Sequence Analysis:

Probablility and Statistics

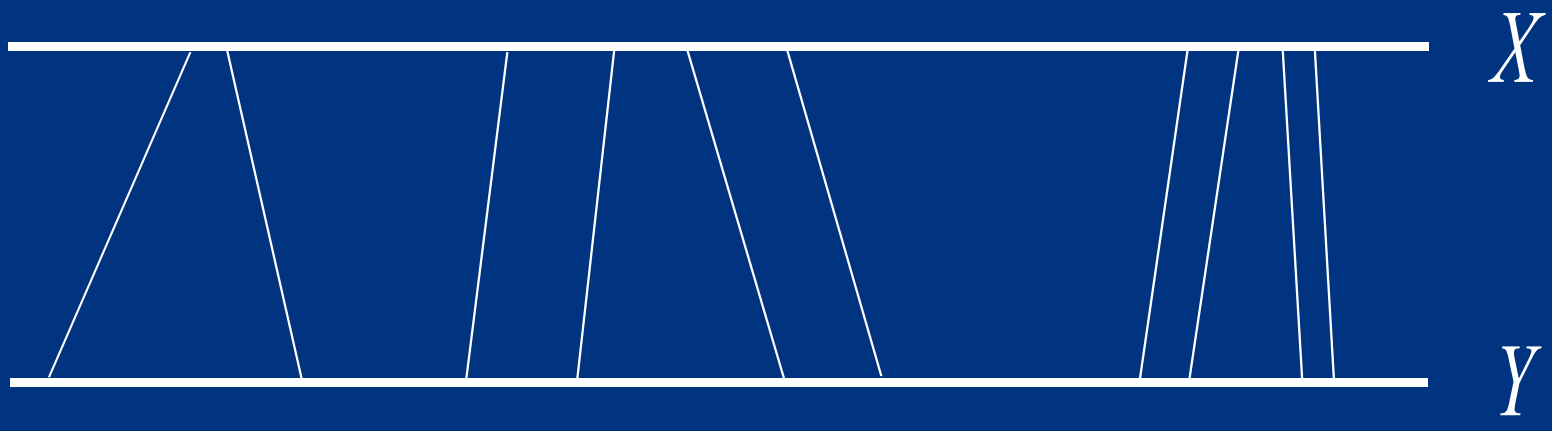
Michael S. Waterman

University of Southern California, Los Angeles
Fudan University, Shanghai

Outline

- Reminder: Algorithms for Sequence Alignment
  - Global Alignment
  - Local Alignment
- Statistical Distribution of Alignment Scores
  - Global Alignment
  - Local Alignment
- Word Counting
  - In One Sequence
  - Sequence Comparison

Global Alignment



$$S(X,Y) = max \left\{ \sum_{\substack{aligned \\ i,j}} s(x_i, y_j) - \delta \#indels \right\}$$

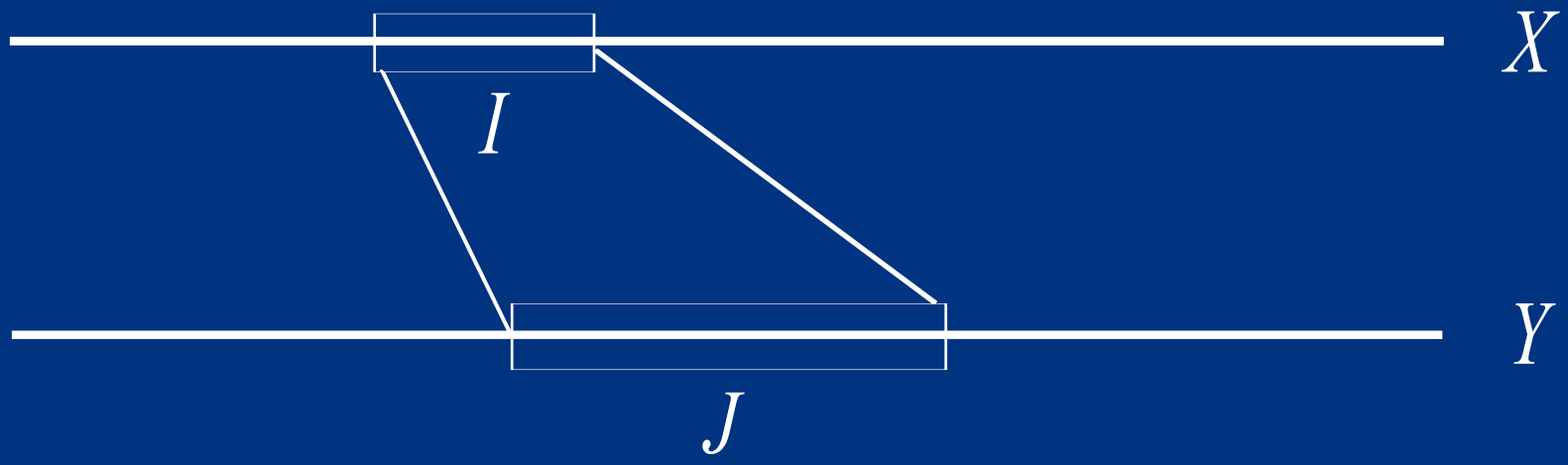
$$S_{i,0} = -i\delta \quad S_{0,j} = -j\delta$$

$$S_{i,j} = max \left\{ \begin{array}{l} S_{i-1,j-1} + s(x_i, y_j) \\ S_{i-1,j} - \delta \\ S_{i,j-1} - \delta \end{array} \right\}$$

$$s(x_i, y_j) = \left\{ \begin{array}{ll} +1 & x_i = y_j \\ -\mu & x_i \neq y_j \end{array} \right.$$

$$S(X, Y) = S_{n,m}$$

Local Alignment



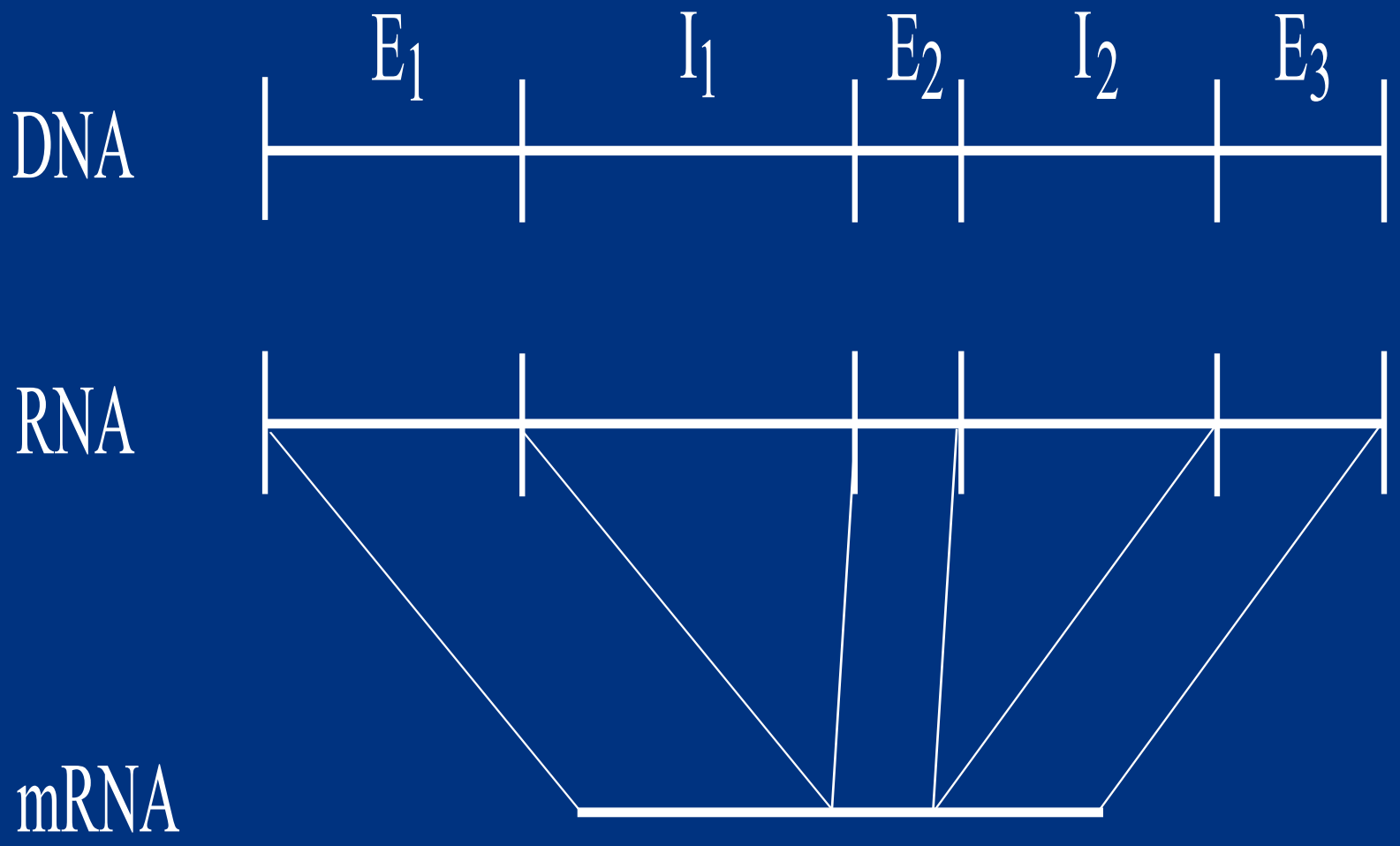
$$H(X,Y) = max\{S(I,J) : I \subset X, J \subset Y\}$$

$$H_{i,0} = 0 \qquad H_{0,j} = 0$$

$$H_{i,j} = max \left\{ \begin{array}{l} H_{i-1.j-1} + s(x_i, y_j) \\ H_{i-1,j} - \delta \\ H_{i,j-1} - \delta \\ 0 \end{array} \right\}$$

$$H(I, J) = max_{i,j} H_{i,j}$$

Smith and Waterman J.Mol.Biol.(1981)


E1
I1
E2
I2
E3
DNA
RNA
mRNA

STATISTICS OF SEQUENCE MATCHING

- Global matching of random sequences has results from subadditive ergodic theory
- Local matching has strong laws and many distributional results
- Local matching statistics are as important as computational efficiency in biological database searching (BLAST)

Statistics of long matches

- Such matches are rare in random sequences and occur in an approximately Poisson number of clumps
- HTTTHHHHTHHHTTHTHHT
- number of runs of 3Hs $= 3$
- number of clumps of 3Hs $= 2$
- Cannot directly apply $Bin(n, p) \approx Poisson(\lambda = np)$

Statistics of the number of short words

- w occurs frequently, in overlapping clumps
- Cannot apply $Bin(n,p) \approx Normal(np, np(1-p))$
- $N_w =$ number of w occurences in sequence of length n
- σ_w^2 is a function of self-overlap of w
- $(N_w - nP(w))/\sqrt{n}\sigma_w \approx N(0,1)$

Multivariate word counts

Theorem. Let $\{A_i\}_{i\geq 1}$ be a stationary, irreducible, aperiodic first-order Markov chain. Let $W = \{w_1, \ldots, w_m\}$ be a set of words and $\mathbf{N} = (N_1(n), \ldots, N_m(n))$ be the count vector. Then $n^{-1}\mathbf{N}$ is asymptotically normal with mean μ and covariance matrix $n^{-1}\Sigma$. If $\det(\Sigma) \neq 0$, then

$$n^{1/2}\Sigma^{-1/2}\,(\ N/n - \mu) \Longrightarrow \Phi(\mathbf{0}, \mathbf{1}).$$

The covariance matrix is calculated using the overlap polynomial between the words. Results from *Lindstrom (1990)*, thesis Using Stein's Method, rates of convergence
Haiyan Huang (2001/2), thesis.

GLOBAL ALIGNMENT

What do we know about the statistical distribution of global alignment scores? We assume the sequence letters are iid.

Chavatal-Sankoff in 1975 had a proof that $\mathbb{E}(S_n) \sim \alpha n$, but to this day no one knows α for any non trivial example.

Their case study was for length of longest common subsequence (LCS), which is the global alignment score for $s(a, b)$ equal 1 if $a = b$ and 0 for all other scoring weights.

We are guaranteed that α is at least $\mathbb{P}(A = B)$. For binary uniformly distributed sequences, $\alpha \geq 0.5$. In fact the value is approximately 0.82, showing the power of alignment.

Assume $\mathbf{A} = A_1 A_2 \cdots A_n$ and $\mathbf{B} = B_1 B_2 \cdots B_n$ with A_i, B_j iid. We apply Kingman's subaddative ergodic theorem.

Define $S_n = S(\mathbf{A}, \mathbf{B})$. Then, there is a constant $\rho \geq \mathbb{E}(s(A, B))$ such that

$$\lim_{n \to \infty} \frac{S_n}{n} = \rho$$

probability 1 and in the mean.

Azuma-Hoeffding

Assume $\mathbf{A} = A_1 A_2 \cdots A_n$ and $\mathbf{B} = B_1 B_2 \cdots B_n$ with A_i, B_j iid.

Then there is a computable constant c such that,

$$\mathbb{P}(S - \mathbb{E}S \geq \gamma n) \leq e^{-\gamma^2 n/2c^2}.$$

This large-deviations result is useless for practical p-values, but is very useful in proving theorems.

Variance

There is a conjecture credited to me that $Var(S_n) = \Theta(n)$. Numerical evidence from simulations are the basis of the conjecture. It is well known (due to Michael Steele) that

$$Var(S_n) \leq \kappa n \text{ for a known constant } \kappa.$$

Chvatal and Sankoff conjectured that $Var(S_n)$ is of order $o(n^{2/3})$, which as it turned out is the same order (when properly rescaled), as obtained by Baik, Deift and Johansson in their much celebrated result on the Longest Increasing Subsequence (LIS) of a random permutation.

Matzinger and Lember established the $\Theta(n)$ variance order for binary sequences with one symbol having extremely small probability.

Houdre and Matzinger recently show the conjecture holds for binary sequences where one symbol has "somewhat larger score" than the other.

The problem that I would like to see solved is the full probability distribution for LCS and general global alignment scores, just as has been established for LIS of a random permutation.

LOCAL ALIGNMENT

Consider a coin tossing sequence of length n with $\mathbb{P}(H) = p$. What is the length of the longest run of heads?

The expected number of clumps of k heads beginning with a tail (forgetting end effects) is

$$n(1-p)p^k.$$

If we set this expectation equal to 1 and solve for k, we get

$$k = log_{1/p}(n) + log_{1/p}(1-p).$$

Exact matching

Let $A_1, A_2, \ldots, B_1, B_2, \ldots$ be independent and identically distributed with $0 < p \equiv \mathbb{P}(X_1 = Y_1) < 1$.

Define $H_n = \max\{m : A_{i+k} = B_{j+k}$ for $k = 1$ to $m,\ 0 \leq i, j \leq n - m\}$.

Then

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{H_n}{\log_{1/p}(n)} = 2\right) = 1.$$

Note that $2\log_{1/p}(n)$ is the solution for k in

$1 = n^2 p^k$

Matching sequences with different distributions

Let $A_1, A_2, \ldots$ be distributed as ξ, $B_1, B_2, \ldots$ be distributed as ν with all letters independent and $p = \mathbb{P}(X_1 = Y_1) \in (0, 1)$.

Then there is a constant $C(\xi, \nu) \in [1, 2]$ such that

$$\mathbb{P}\left(\lim_{n\to\infty} H_n / \log_{1/p}(n) = C(\xi, \nu)\right) = 1.$$

In addition

$$C(\xi,\nu) = \sup_{\gamma} \min \left\{ \frac{\log(1/p)}{\mathcal{H}(\gamma,\xi)}, \frac{\log(1/p)}{\mathcal{H}(\gamma,\nu)}, \frac{2\log(1/p)}{\log(1/p)+\mathcal{H}(\gamma,\beta)} \right\}$$

where $\beta_a \equiv \xi_a \nu_a / p$, $\mathcal{H}(\beta,\nu) = \sum_a \beta_a \log(\beta_a/\nu_a)$, and γ ranges over the probability distributions on the alphabet. Here log can be to any base.

And

$C(\xi, \gamma) = 2$ if and only if

$$\max\{\mathcal{H}(\beta, \nu), \mathcal{H}(\beta, \xi)\} \leq \left(\frac{1}{2}\right) \log(1/p).$$

Let's take an example.

Let $A_1, A_2, \ldots$ have $\mathbb{P}(A_i = H) = \mathbb{P}(A_i = T) = 1/2$ and $B_1, B_2, \ldots$ have $\mathbb{P}(B_i = H) = \theta = 1 - \mathbb{P}(B_i = T)$ where $\theta = [0, 1]$.

$$p = \mathbb{P}(A_i = B_i) = 1/2\theta + 1/2(1 - \theta) = 1/2$$

For $\theta = 1$ (or 0), the value $H_n =$ length of the longest run of H's (or Ts) in $\mathbf{A}$.. Therefore in both cases, $H_n \sim \log_2(n)$. However, if $\theta = 1/2$, A_i and B_i have the same distribution and $H_n \sim 2\log_2(n)$. The theorem tells us

$$H_n \sim C\log_2(n), C \in [1, 2],$$

and $C = 2$ if and only if

$$\max\{\mathcal{H}(\beta, \nu), \mathcal{H}(\beta, \xi)\} \leq 1/2\log(2).$$

if and only if θ is in $[0.11002786\ldots, 0.88997214\ldots]$.

Poisson approximation

The $log(n)$ results hold for scoring too. Here we set deletion weights ($g(k) = \infty$ for all k), $\mathbb{E}(s(A, B)) < 0$ and $s^* = \max s(a, b) > 0$.

For p the largest root of $1 = \mathbb{E}(\lambda^{-s(A,B)})$

$$\mathbb{P}(\lim_{n\to\infty} H_n / \log_{1/p}(n^2) = 1).$$

Therefore for the case of no indels we know the center of the score is $2\log_{1/p}(n)$.

A Chen-Stein style theorem can be proved.

Under the above assumptions, there is a constant $\gamma > 0$ such that

$$\mathbb{P}(H(\mathbf{A}, \mathbf{B}) > \log_{1/p}(nm) + c) \approx 1 - e^{-\gamma nmp^c}.$$

While this result has not been proved for the usual alignment problem with indels, it is indication of the range of validity of Poisson approximation.

The e-values in BLAST are from a version of this result.

A Phase Transition

The result below is true for much more general scoring but for simplicity we study

$$s(a,b) = \{+1, a = b; \quad -\mu, a \neq b\}$$

and $g(k) = -\delta k$ where $(\mu, \delta) \in [0, \infty]^2$.

The following limit exists by Kingman:

$$\rho = \rho(\mu, \delta) = \lim_{k \to \infty} \frac{\mathbb{E}(S_k)}{k} = \sup_{k \geq 1} \frac{\mathbb{E}(S_k)}{k}.$$

The solution of $\rho = \rho(\mu, \delta) = 0$ is a line in parameter space.

(I) Obviously when $\rho = \rho(\mu, \delta) > 0$ the score H_n grows linearly.

(II) Much less obviously when $\rho = \rho(\mu, \delta) < 0$ the score H_n grows logarithmically.

LAST CONJECTURE

Poisson Approximation is valid in the logarithmic region.

No one is close to proving this. It is probably further away than the full distribution of LCS.

STATISTICS OF NO. OF SHORT WORDS IN A SEQUENCE

- w occurs frequently, in overlapping clumps
- Cannot apply $Bin(n,p) \approx Normal(np, np(1-p))$
- $N_w =$ number of w occurences in sequence of length n
- σ_w^2 is a function of self-overlap of w
- $(N_w - nP(w))/\sqrt{n}\sigma_w \approx N(0,1)$

MULTIVARIATE WORD COUNTS IN A SINGLE SEQUENCE

Theorem. Let $\{A_i\}_{i\geq 1}$ be a stationary, irreducible, aperiodic first-order Markov chain. Let $W = \{w_1, \ldots, w_m\}$ be a set of words and $\mathbf{N} = (N_1(n), \ldots, N_m(n))$ be the count vector. Then $n^{-1}\mathbf{N}$ is asymptotically normal with mean μ and covariance matrix $n^{-1}\Sigma$. If $\det(\Sigma) \neq 0$, then

$$n^{1/2}\Sigma^{-1/2}\,(\ N/n - \mu) \Longrightarrow \Phi(\mathbf{0}, \mathbf{1}).$$

The covariance matrix is calculated using the overlap polynomial between the words. Results from *Lindstrom (1990)*, thesis
Using Stein's Method, rates of convergence
Haiyan Huang (2001/2), thesis.

We start with two sequences of iid letters

$\mathbf{A} = A_1 A_2 \cdots A_n$
$\mathbf{B} = B_1 B_2 \cdots B_m$

$$f_a = P(A_i = a) = P(B_j = a), \quad a \in \mathcal{A}$$

$$p_k = \sum_{a \in \mathcal{A}} f_a^k.$$

D_2 has been defined as the dot product of the k-word count vectors. It is computed in linear time.

$$D_2 = \sum_{w \in k-word} n_A(w) n_B(w)$$

Define the match indicator $C_{i,j} = 1\{A_i = B_j\}$, and the k-word match indicator at position (i, j)

$$Y_{i,j} = C_{i,j} C_{i+1,j+1} \cdots C_{i+k-1,j+k-1}.$$

Note: $\mathbf{E}C_{i,j} = p_2$ and $\mathbf{E}Y_{i,j} = p_2^k$

$$D_2 = \sum_{v \in I} Y_v.$$

MOTIVATION: To find useful distributions for D_2 and p-values

For LARGER k

We should have approximately a Poisson number of clumps of matching k-words
Each clump has a geometric number of matching k-words since a clump implies k matches and additional matches occur with probability p_2

Therefore using Chen-Stein we expect to obtain a compound Poisson approximation

For SMALLER k

We have $(n-k+1)(n-k+1)$ rv's $C_{i,j}$ which are 1 with probability p_k and 0 otherwise.
If Cs are independent, D_2 is $Bin(n, p_2)$, n large.
That is, approximately a normal.

Therefore using Stein's method we expect to obtain a normal approximation.

$$W = \frac{D_2 - \mathbf{E}D_2}{\sqrt{\mathsf{Var}(D_2)}} = \sum_v \frac{Y_v - \mathbf{E}Y_v}{\sqrt{\mathsf{Var}(D_2)}}.$$

Stein-Rinot-Rotar. Let $X_j \in \mathcal{R}^d$, and let $W = \sum_{j=1}^n X_j$.

$$|X_j| \leq \mathrm{B}.$$

Let $|\mathcal{S}_i|$ and $|\mathcal{N}_i|$ be subsets of $\{1, \ldots, n\}$, $i \in \mathcal{S}_i \subset \mathcal{N}_i, i = 1, \ldots, n$. Constants $\mathrm{C}_1 \leq \mathrm{C}_2$:

$$\max\{|\mathcal{S}_i|, i = 1, \ldots, n\} \leq \mathrm{C}_1; \max\{|\mathcal{N}_i|, i = 1, \ldots, n\} \leq \mathrm{C}_2,$$

where for sets $|\cdot|$ denotes cardinality.

Then, there exists a universal constant c such that

$$\sup_{h\in\mathcal{C}} |\mathbf{E}h(W) - \Phi h| \leq c\{2\mathrm{C}_2\mathrm{B} + n(2+\sqrt{\mathbf{E}W^2})\mathrm{C}_1\mathrm{C}_2\mathrm{B}^3 + \chi_1 + \chi_2 + \chi_3\}.$$

where

$$\chi_1 = \sum_{j=1}^n \mathbf{E}|\mathbf{E}(X_j|\sum_{k\notin\mathcal{S}_j} X_k)|,$$

$$\chi_2 = \sum_{j=1}^n \mathbf{E}|\mathbf{E}(X_j(\sum_{k\in\mathcal{S}_j} X_k)^T) - \mathbf{E}(X_j(\sum_{k\in\mathcal{S}_j} X_k)^T|\sum_{l\notin\mathcal{N}_j} X_l)|$$

$$\chi_3 = |I - \sum_{j=1}^n \mathbf{E}(X_j(\sum_{k\in\mathcal{S}_j} X_k)^T)|.$$

$n = m$, $k = -\frac{\alpha}{\log(p_2)} \log(n)$ and $n >> k$

$\sup_{h \in cxH} |\mathbf{E}h(W) - \Phi h| \leq c\{2C_2B + n^2(2+1)C_1C_2B^3\}$
$\leq c\{8\frac{k^{\frac{1}{2}}}{(p_3/p_2^2-1)^{\frac{1}{2}} n^{\frac{1}{2}-\alpha}} + 12\frac{k^{\frac{1}{2}}}{(p_3/p_2^2-1)^{\frac{3}{2}} n^{\frac{1}{2}-3\alpha}}\}$ which has a rate

$$O(\frac{\sqrt{\log(n)}}{n^{\frac{1}{2}-3\alpha}})$$

as n goes to ∞.

When $\alpha < \frac{1}{6}$, the error bound will go to zero.
For $k = \alpha \log_{1/p_2}(n)$ with $0 < \alpha < 1/6$, D_2 is approximately normal.

THE GLITCH:

When uniformly distributed $p_3 = p_2^2$ and $p_3/p_2^2 - 1 = 0$
For the uniform we have NO bound, it is the exceptional case!

The Non-Normal Case

Alphabet is $\{0, 1\}$,
$P(0 \text{ appears }) = p, \quad P(1 \text{ appears }) = q.$

Denote $\#$ of $0, 1$ in the two sequences by X and Y respectively, then

$$D_2 = XY + (n - X)(n - Y).$$

$$\mathbf{E}(D_2) = n^2(1 - 2pq),$$

and

$$\text{Var}(D_2) = 2n^2pq(1 - 2pq) + 2n^2(n - 1)pq(p - q)^2$$
$$= \text{O}(n^2) \text{ if } p = q = \tfrac{1}{2}; = \text{O}(n^3) if p \neq q$$

Next:

$$\begin{aligned}
\frac{D_2 - \mathbf{E}(D_2)}{\sigma} &= \frac{2npq}{\sigma}\left(\frac{X-np}{\sqrt{npq}}\right)\left(\frac{Y-np}{\sqrt{npq}}\right) \\
&\quad + n(2p-1)\frac{\sqrt{npq}}{\sigma}\left(\frac{Y-np}{\sqrt{npq}}\right) \\
&\quad + n(2p-1)\frac{\sqrt{npq}}{\sigma}\left(\frac{X-np}{\sqrt{npq}}\right) \\
&= \frac{2npq}{\sigma}\frac{(X-np)}{\sqrt{npq}}\frac{(Y-np)}{\sqrt{npq}} \quad : \quad p = q = \frac{1}{2}
\end{aligned}$$

So the limit is normal if $p \neq q$ and the product of independent normals if $p = q$

$$\tilde{X}_{\mathbf{w}} = X_{\mathbf{w}} - \bar{n}p_{\mathbf{w}} \text{ and } \tilde{Y}_{\mathbf{w}} = Y_{\mathbf{w}} - \bar{m}p_{\mathbf{w}};$$

$$D_2^* = \sum_{\mathbf{w}\in\mathcal{A}^k} \tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}.$$

$$D_2 = D_2^* + \bar{n}\sum_{\mathbf{w}\in\mathcal{A}^k} p_{\mathbf{w}}Y_{\mathbf{w}} + \bar{m}\sum_{\mathbf{w}\in\mathcal{A}^k} p_{\mathbf{w}}X_{\mathbf{w}} - \bar{m}\bar{n}\sum_{\mathbf{w}\in\mathcal{A}^k} p_{\mathbf{w}}^2.$$

Shepp

In 1964 Larry Shepp observed that, if X and Y are independent mean zero normals, X with variance σ_X^2, Y with variance σ_Y^2, then $\frac{XY}{\sqrt{X^2+Y^2}}$ is again normal, with variance τ such that $\frac{1}{\tau} = \frac{1}{\sigma_X} + \frac{1}{\sigma_Y}$.

We introduce

$$D_2^S = \sum_{\mathbf{w}\in\mathcal{A}^k} \frac{\tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}}{\sqrt{\tilde{X}_{\mathbf{w}}^2 + \tilde{Y}_{\mathbf{w}}^2}}.$$

We also use

$$D_2^{*vc} = \sum_{\mathbf{w}} \frac{\widetilde{X}_{\mathbf{w}} \widetilde{Y}_{\mathbf{w}}}{\sqrt{\bar{n}\bar{m}\widehat{p_{\mathbf{w}}}}},$$

This statistic comes from the normalized $\sum_{\mathbf{w}} \frac{\widetilde{X}_{\mathbf{w}} \widetilde{Y}_{\mathbf{w}}}{\sqrt{\widehat{\mathrm{Var}X_{\mathbf{w}}}\widehat{\mathrm{Var}Y_{\mathbf{w}}}}}$, but as the variance is costly to compute, replacing it by the estimated mean of the word occurrence across the two sequences when the size of the word pattern is large by Poisson approximation.

Thanks for listening!

Thanks for listening!