

A Word Count Statistic from Computational Biology

Michael S. Waterman
University of Southern California

Collaborators

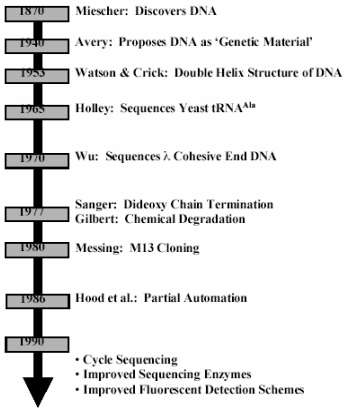
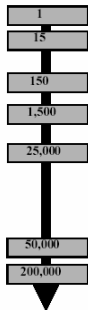
Ross A. Lippert, Celera Genomics, MIT
Haiyan Huang, Harvard Medical School, UC Berkeley
Gesine Reinert, Oxford University

Outline

- Biological Sequence History
- Sequence Comparison History
- Comparison by Composition, D_2
- D_2 Approximately Compound Poisson
- D_2 Approximately Normal
- D_2 Approximately ???
- Numerical Simulations

History of DNA Sequencing

Efficiency
(bp/person/year)



DNA Sequencing History

Shotgun sequencing

1995

TIGR (The Institute of Genomic Research)

The first complete DNA sequence of the genome of a free living organism --- the bacterium *Haemophilus influenzae* (1.8Mbp).

1996

International Consortium

The first complete DNA sequence of the genome of a eukaryote --- the yeast *Saccharomyces cerevisiae* (12Mbp).

1998

International Consortium

The first complete DNA sequence of the genome of a multicellular organism --- the roundworm *Caenorhabditis elegans* (97Mbp).

1999

Celera Genomics

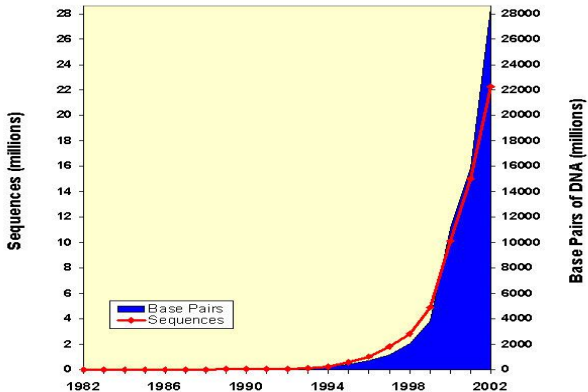
The entire genomic sequence of the fruitfly *Drosophila melanogaster* (137Mbp).

2000

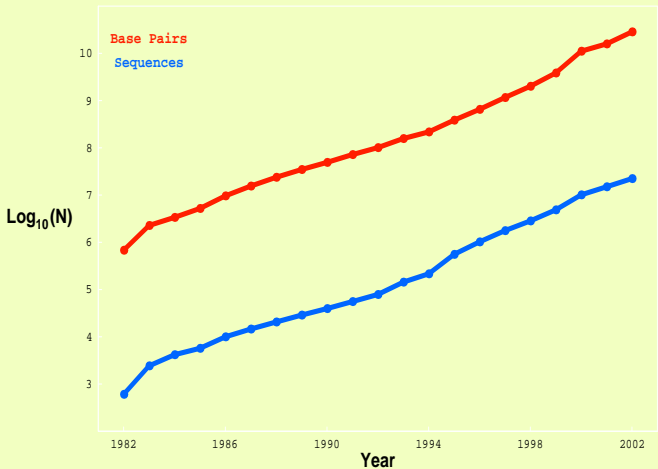
International Consortium & Celera

The first draft of the sequence of the entire human genome (3000 Mbp).

Growth of GenBank



Growth of GenBank (Log Scale)



SEQUENCE COMPARISON

Two sequences can be related in an alignment

what
wh-y

Scoring:

<i>Alignment</i>	<i>Count</i>	<i>Neg. Cost</i>
Identities	2	+1
Mismatches	1	$-\mu$
Indels	1	$-\delta$

$$S = \left\{ \begin{array}{l} \text{what} \\ \text{wh-y} \end{array} \right\} = 2 - \mu - \delta$$

GLOBAL ALIGNMENT



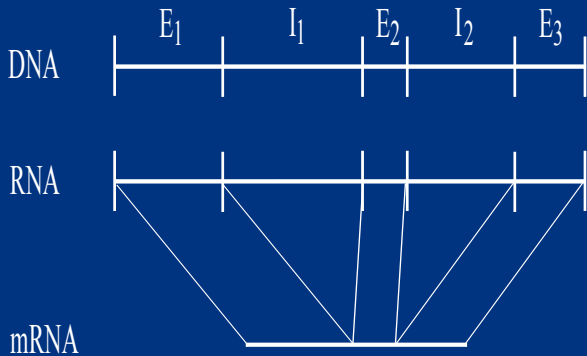
$$S(X, Y) = \max \left\{ \sum_{\substack{\text{aligned} \\ i, j}} s(x_i, y_j) - \delta \# \text{indels} \right\}$$

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

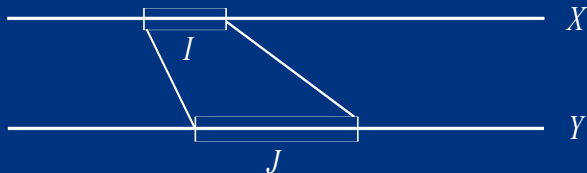
$$S_{i,j} = \max \left\{ \begin{array}{l} S_{i-1,j-1} + s(x_i, y_j) \\ S_{i-1,j} - \delta \\ S_{i,j-1} - \delta \end{array} \right\}$$

$$s(x_i, y_j) = \begin{cases} +1 & x_i = y_j \\ -\mu & x_i \neq y_j \end{cases}$$

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



LOCAL ALIGNMENT



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

$$M_{i,0} = 0 \quad M_{0,j} = 0$$

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

$$M(I, J) = \max_{i,j} M_{i,j}$$

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

STATISTICS OF SEQUENCE MATCHING 1983-2000

- Global matching of random sequences has results from subadditive ergodic theory and large deviations
- Local matching has strong laws and many distributional results, but not yet including the full biological models
- Local matching statistics are as important as computational efficiency in biological database searching (BLAST)

BLAST

- Most used software in molecular biology
- Searches DNA and protein sequence databases
- Heuristic for local alignment algorithm given before
- Algorithm for local matching is word based
- Statistical significance estimation essential

STATISTICS OF LONG LOCAL 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 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 occurrences 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

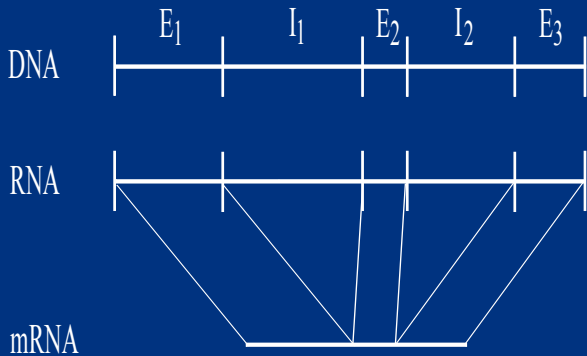
Theorem. Let $\{A_i\}_{i \geq 1}$ be a stationary, irreducible, aperiodic first-order Markov chain. Let $W = \{w_1, \dots, w_m\}$ be a set of words and $\mathbf{N} = (N_1(n), \dots, N_m(n))$ be the count vector. Then $n^{-1}\mathbf{N}$ is asymptotically normal with mean μ and covariance matrix $n^{-1}\Sigma$.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

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

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.

D_2 Motivation

- Many partial RNA sequences (ESTs) corresponding to genes are produced of approx 500 letters.
- Each sequence is compared to EST and DNA databases. The comparison must be very rapid to be useful and not expensive.
- A simple comparison statistic (and relatives) was used that takes linear time. It simply counts the number of k -word matches there are between two sequences, regardless of order of matching.
- Implicitly this statistic was assumed to be normally distributed.



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.

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

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

Let X_v be the declumped matching indicators associated with the Y_v by,

$$\begin{aligned} X_{i,j} &= Y_{i,j} & : & \quad i = 0 \text{ or } j = 0 \\ X_{i,j} &= (1 - C_{i,j})Y_{i,j} & : & \quad \text{else} \end{aligned}$$

D_{2*} , the “declumped D_2 ”, is $\sum_{v \in I} X_v$.

Chen-Stein Let X_i for $i \in I$ be indicator random variables such that X_i is independent of $\{X_j\}$, $j \notin J_i$. Let $W = \sum_{i \in I} X_i$ and $\lambda = \mathbf{E}W$ and let Z be a Poisson random variable with $\mathbf{E}Z = \lambda$. Then

$$\|W - Z\| \leq 2(b_1 + b_2) \frac{1 - e^{-\lambda}}{\lambda} \leq 2(b_1 + b_2),$$

and in particular

$$|P(W = 0) - e^{-\lambda}| \leq (b_1 + b_2) \frac{1 - e^{-\lambda}}{\lambda},$$

where

$$b_1 = \sum_{v \in I} \sum_{u \in J_v} \mathbf{E}X_v \mathbf{E}X_u,$$

and

$$b_2 = \sum_{v \in I} \sum_{v \neq u \in J_v} \mathbf{E}(X_u X_v).$$

$$n = m, k = -\frac{\alpha}{\log(p_2)} \log(n), n \gg k \gg 1.$$

$$\lambda = (1 - p_2)n^{2-\alpha} = O(n^{2-\alpha}).$$

Bound $b_1 + b_2$ by

$$\begin{aligned} b_1 + b_2 \leq & 4kn^{3-2\alpha} + 4kn^{3-(\frac{3}{2}+3\delta)\alpha} \\ & + (2k)^2 p_2^{\frac{1}{2}+\gamma} n^{2-(2\gamma+1)\alpha}, \end{aligned}$$

Here $\delta \in (0, 1/6], \gamma \in (0, .5 - 1/(2k + 1)]$ Thus, $b_1 + b_2 \leq$

$$O\left(\frac{\log(n)}{n^{2\alpha-3}}\right) + O\left(\frac{\log(n)}{n^{(\frac{3}{2}+3\delta)\alpha-3}}\right) + O\left(\frac{(\log(n))^2}{n^{(2\gamma+1)\alpha-2}}\right).$$

Clearly, for $\alpha \geq 2$ we can ensure that $b_1 + b_2$ goes to 0.

$$P(\text{Clump Size} = k + T = k + t) = (1 - p_2)p_2^t,$$

The resulting model of D_2 is a compound Poisson process of independent geometrically distributed variables T_i ,

$$D_2 \sim \sum_{i=1}^{Z(\lambda)} (1 + T_i),$$

where $\lambda = (1 - p_2)p_2^k n^2$

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 C s are independent, D_2 is $Bin(n, p_2^k)$, n large. That is, approximately a normal.
- Therefore using Stein's method we expect to obtain a normal approximation.

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

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

$$|X_j| \leq B.$$

Let $|\mathcal{S}_i|$ and $|\mathcal{N}_i|$ be subsets of $\{1, \dots, n\}$,
 $i \in \mathcal{S}_i \subset \mathcal{N}_i, i = 1, \dots, n$. Constants $C_1 \leq C_2$:

$$\max\{|\mathcal{S}_i|, i = 1, \dots, n\} \leq C_1; \max\{|\mathcal{N}_i|, i = 1, \dots, n\} \leq C_2,$$

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

Then, there exists a universal constant c such that

$$\sup_{h \in \mathcal{C}} |\mathbf{E}h(W) - \Phi h| \leq c\{2C_2B + n(2 + \sqrt{\mathbf{E}W^2})C_1C_2B^3 + \chi_1 + \chi_2 + \chi_3\}.$$

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 $n \gg k$

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

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

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 $p_3/p_2^2 - 1 = 0$

$$p_2 = \sum_{a \in \mathcal{A}} 1/A^2 = 1/A$$

$$p_3 = \sum_{a \in \mathcal{A}} 1/A^3 = 1/A^2$$

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 0s 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 $\text{Var}(D_2) =$

$$2n^2pq(1 - 2pq) + 2n^2(n - 1)pq(p - q)^2$$

$$= O(n^2) \text{ if } p = q = \frac{1}{2};$$

$$= O(n^3) \text{ if } p \neq q$$

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 $p = q$

NORMAL CASES

- $D_2 = X_1Y_1 + X_2Y_2 + \dots + X_dY_d$, // where d is alphabet size.
- Normal limit when d is large, as when 4^k is large.
- These quantities should be small with respect to n .
Difficult cases to prove, see G. Reinert.

SIMULATIONS

- Simulate 2500 comparisons (2×2500 sequences)
- Compare score distributions to both compound Poisson and Normal distributions
- p -values from Kolmogorov-Smirnov test
- When distributions match, the p -value will be uniformly distributed on $(0,1)$;
when the fit is poor the p -values will be near 0.

SIMULATIONS, COMPOUND POISSON

- For non-uniform letters, $p_G = p_C = 1/3$ and $p_A = p_T = 1/6$.
- Good approximation if $k > 2\log_{1/p}(2^x \times 10^2) \approx 1.2x + 7.2$
- For uniform letters, $p_A = p_G = p_C = p_T = 1/4$.
- Good approximation if $k > 1.5\log_{1/p}(2^x \times 10^2) \approx 0.75x + 5$

k/n	100	200	400	800	1600	3200	6400	12800
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
3	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
4	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
5	0.15181	0.00010	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
6	0.01670	0.06138	0.00067	0.00000	0.00000	0.00000	0.00000	0.00000
7	0.08230	0.24738	0.12469	0.00075	0.00000	0.00000	0.00000	0.00000
8	0.78766	0.67140	0.04178	0.14229	0.00667	0.00001	0.00000	0.00000
9	0.24738	0.36250	0.13325	0.67140	0.20728	0.01066	0.00179	0.00000
10	0.50706	0.57613	0.06613	0.52972	0.02357	0.95655	0.14229	0.03027

K-S probabilities for non-uniform D_2 compared to compound Poisson

k/n	100	200	400	800	1600	3200	6400	12800
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
2	0.00059	0.00029	0.00006	0.00409	0.00003	0.00023	0.00549	0.00075
3	0.07657	0.05693	0.02787	0.00605	0.29296	0.02787	0.11658	0.15181
4	0.30940	0.69525	0.03027	0.42112	0.59975	0.10167	0.46307	0.88737
5	0.18346	0.78766	0.11658	0.67140	0.74228	0.29296	0.96602	0.18346
6	0.00104	0.64747	0.13325	0.19509	0.04885	0.69525	0.90400	0.26195
7	0.46307	0.23342	0.46307	0.04885	0.08230	0.48483	0.71892	0.22006
8	0.48483	0.27715	0.03562	0.76524	0.30940	0.91930	0.96602	0.24738
9	0.83039	0.48483	0.19509	0.00199	0.11658	0.74228	0.34418	0.08230
10	0.14229	0.23342	0.78766	0.62356	0.08230	0.07657	0.22006	0.16183

K-S probabilities for uniform D_2 compared to compound Poisson

SIMULATIONS, NORMAL

- For non-uniform letters, $p_G = p_C = 1/3$ and $p_A = p_T = 1/6$.
- Good approximation if $k < 1/6 \log_{1/p}(2^x \times 10^2) \approx x/10 + 0.6$
- For uniform letters, $p_A = p_G = p_C = p_T = 1/4$.
- Good approximation if k is large enough to have 4^k large but still have $2^x \times 10^2 = n > 4^k$ or $k < 1/2x + 3.32$

k/n	100	200	400	800	1600	3200	6400	12800
1	0.05862	0.00419	0.13668	0.11486	0.31036	0.09010	0.91967	0.00506
2	0.03365	0.00006	0.00061	0.66297	0.29724	0.16957	0.66064	0.68674
3	0.00000	0.01002	0.05023	0.39328	0.05444	0.05082	0.77163	0.38298
4	0.00000	0.00004	0.00039	0.14959	0.03058	0.26901	0.59183	0.93879
5	0.00000	0.00004	0.00048	0.14381	0.03832	0.04490	0.55703	0.62759
6	0.00000	0.00000	0.00022	0.00403	0.00601	0.08003	0.59902	0.32061
7	0.00000	0.00000	0.00000	0.00009	0.00475	0.56324	0.29819	0.46705
8	0.00000	0.00000	0.00000	0.00000	0.00058	0.11751	0.32351	0.17059
9	0.00000	0.00000	0.00000	0.00000	0.00000	0.00005	0.15591	0.18042
10	0.00000	0.00000	0.00000	0.00000	0.00000	0.00002	0.02962	0.11055

K-S probabilities for non-uniform D_2 compared to normal

k/n	100	200	400	800	1600	3200	6400	12800
1	0.00000	0.00000	0.00000	0.00000	0.00010	0.00002	0.00001	0.00002
2	0.03811	0.15773	0.28724	0.47452	0.07759	0.19055	0.25803	0.00939
3	0.04802	0.15361	0.12058	0.55153	0.70760	0.22644	0.81058	0.31066
4	0.00000	0.05730	0.04796	0.81343	0.68940	0.65794	0.98177	0.69245
5	0.00000	0.00001	0.23410	0.18908	0.77291	0.10750	0.08259	0.08706
6	0.00000	0.00000	0.00144	0.07070	0.08660	0.72020	0.06702	0.45234
7	0.00000	0.00000	0.00000	0.00000	0.02782	0.69609	0.26900	0.06839
8	0.00000	0.00000	0.00000	0.00000	0.00000	0.06281	0.65713	0.05397
9	0.00000	0.00000	0.00000	0.00000	0.00000	0.00001	0.00000	0.32139
10	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00001	0.00011

K-S probabilities for uniform D_2 compared to normal

CONCLUSIONS

- Small k – Normal and Compound Poisson
- Large k – Compound Poisson
- For $k = 6$ which is often used in practice, Compound Poisson is best unless distribution is uniform (where equivalent to normal)
- Other versions of the count statistics (several have been proposed) need to be studied