The number of occurrences of a word (5.7) and motif (5.9) in a DNA sequence, allowing overlaps

Covariance (2.4) and indicators (2.9)

Prof. Tesler

Math 283
February 25, 2013

Covariance

- Let X and Y be random variables, possibly dependent.
- $\mathrm{Var}(X + Y) = E((X + Y - \mu_X - \mu_Y)^2)$

$$= E\left(\left((X - \mu_X) + (Y - \mu_Y)\right)^2\right)$$

$$= E\left((X - \mu_X)^2\right) + E\left((Y - \mu_Y)^2\right) + 2E\left((X - \mu_X)(Y - \mu_Y)\right)$$

$$= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$$

where the *covariance* of X and Y is defined as

$$\mathrm{Cov}(X, Y) = E\left((X - \mu_X)(Y - \mu_Y)\right)$$

- Expanding gives an alternate formula
  $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$:

$$\mathrm{Cov}(X, Y) = E\left((X - \mu_X)(Y - \mu_Y)\right)$$

$$= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y = E(XY) - E(X)E(Y)$$

Covariance properties

- $\text{Cov}(X, X) = \text{Var}(X)$
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- If X, Y are independent then $\text{Cov}(X, Y) = 0$ and
  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.
  Beware, this is not reversible; $\text{Cov}(X, Y)$ could be 0 for dependent variables.
- $\text{Cov}(aX + b, cY + d) = ac\,\text{Cov}(X, Y)$
- $\text{Var}(X_1 + X_2 + \cdots + X_n) = \text{Var}(X_1) + \cdots + \text{Var}(X_n) + 2\!\!\sum_{1 \leqslant i < j \leqslant n}\!\!\text{Cov}(X_i, X_j)$

Sign of covariance

- ***When $\text{Cov}(X, Y)$ is positive:***
  there is a tendency to have $X > \mu_X$ when $Y > \mu_Y$ and vice-versa, and $X < \mu_X$ when $Y < \mu_Y$ and vice-versa.
- ***When $\text{Cov}(X, Y)$ is negative:***
  there is a tendency to have $X > \mu_X$ when $Y < \mu_Y$ and vice-versa, and $X < \mu_X$ when $Y > \mu_Y$ and vice-versa.

Occurrences of a word in a sequence — notation

- Consider a (long) single-stranded nucleotide sequence $\tau = \tau_1 \ldots \tau_N$ and a (short) word $w = w_1 \ldots w_k$:

$$\tau = \tau_1 \ldots \tau_{19} = \texttt{CTATAGATAGATAGACAGT}$$
$$w = w_1 \ldots w_9 = \texttt{ATAGATAGA}$$

- Say w *occurs* in τ at position j when w is in τ ending at position j:

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
τ_j	C	T	A	T	A	G	A	T	A	G	A	T	A	G	A	C	A	G	T

so w occurs in τ at 11 and 15 (underlined).

- Let $I_j = \begin{cases} 1 & \text{if } w \text{ occurs in } \tau \text{ at } j; \\ 0 & \text{otherwise.} \end{cases}$ $\quad\quad I_{11} = I_{15} = 1$ $\quad\quad$ other $I_j = 0$

I_j is an *indicator* variable (1 when a condition is true, 0 when false).

- $Y = I_k + I_{k+1} + \cdots + I_N$ is the number of times w occurs in τ. Here, $Y = 2$.

Computing mean number of occurrences $\mu = E(Y)$

- Suppose τ is generated by N independent rolls of a 4-sided die, whose sides have probabilities p_A, p_C, p_G, p_T adding up to 1.

- The probability of a word being generated by rolling such a die is the product of the probabilities of its nucleotides:
$$\pi(w) = p_{w_1} \cdots p_{w_k} \qquad \pi(\texttt{ATAGATAGA}) = p_A{}^5 p_T{}^2 p_G{}^2$$

- The probability of w occurring at $j = k, k+1, \ldots, N$ is $\pi(w)$.

- I_j's are indicator variables, so
$$E(I_j) \;=\; 0P(I_j = 0) + 1P(I_j = 1) \;=\; P(I_j = 1) \;=\; \pi(w)$$
for $j = k, k+1, \ldots, N$.

- $Y = I_k + I_{k+1} + \cdots + I_N$ so the mean number of occurrences is
$$\mu \;=\; E(Y) \;=\; E(I_k) + \cdots + E(I_N) \;=\; (N - k + 1)\,\pi(w).$$

Dependencies between positions

- Occurrences at different positions have dependencies;
  if `AGAG` occurs in τ at 20, it cannot occur at 17, 19, 21, or 23,
  while the probability of overlapping occurrences at 18 or 22 is
  increased since 2 letters are already known to be correct in τ.
  So we have to account for overlapping prefixes/suffixes of w.

- Chapter 5.8 does the related problem of counting occurrences
  without overlaps.
  Chapters 4 and 11 do the much more general problem of Markov
  chains.

Self-overlaps of a word

- Define

$$
\varepsilon_r = \begin{cases} 1 & \text{if the first } r \text{ letters of } w \text{ equal the last } r \text{ letters} \\ & \text{of } w \text{ in the exact same order (string equality);} \\ \\ 0 & \text{otherwise.} \end{cases}
$$

- This lets us account for dependencies between I_j and I_{j+k-r}.

		$w:$	A	T	A	G	A	T	A	G	A								
$r = 9$	$\varepsilon_9 = 1$		*A*	*T*	*A*	*G*	*A*	*T*	*A*	*G*	*A*							
$r = 8$	$\varepsilon_8 = 0$			A	T	A	G	A	T	A	G	A						
$r = 7$	$\varepsilon_7 = 0$				A	T	A	G	A	T	A	G	A					
$r = 6$	$\varepsilon_6 = 0$					A	T	A	G	A	T	A	G	A				
$r = 5$	$\varepsilon_5 = 1$						*A*	*T*	*A*	*G*	*A*	T	A	G	A			
$r = 4$	$\varepsilon_4 = 0$							A	T	A	G	A	T	A	G	A		
$r = 3$	$\varepsilon_3 = 0$								A	T	A	G	A	T	A	G	A	
$r = 2$	$\varepsilon_2 = 0$									A	T	A	G	A	T	A	G	A
$r = 1$	$\varepsilon_1 = 1$										*A*	T	A	G	A	T	A	G	A

Computing $\sigma^2 = \mathrm{Var}(Y)$

- Since the I_j's have dependencies, the variance of their sum $Y = I_k + \cdots + I_N$ is NOT necessarily the sum of their variances. We must consider covariance terms as well:

$$\mathrm{Var}(Y) = \sum_{j=k}^{N} \mathrm{Var}(I_j) + 2 \sum_{j,\ell:\ k \leqslant j < \ell \leqslant N} \mathrm{Cov}(I_j, I_\ell)$$

- **First sum:** Note that $I_j^2 = I_j$ since $I_j = 0$ or 1, so

$$\mathrm{Var}(I_j) = E(I_j^2) - (E(I_j))^2 = \pi(w) - \pi(w)^2$$

and the first sum in $\mathrm{Var}(Y)$ is

$$\sum_{j=k}^{N} \mathrm{Var}(I_j) = (N - k + 1)(\pi(w) - \pi(w)^2)$$

- **Second sum:** next few slides.

The covariances sum is complicated:

- If $\ell - j \geqslant k$ then I_j, I_ℓ are independent and $\mathrm{Cov}(I_j, I_\ell) = 0$.

- If $0 < \ell - j < k$, the words ending at ℓ and j overlap by $r = k - (\ell - j)$ letters. Rewrite ℓ as $\ell = j + k - r$:
$$\mathrm{Cov}(I_j, I_\ell) = \mathrm{Cov}(I_j, I_{j+k-r}) = E(I_j I_{j+k-r}) - E(I_j)E(I_{j+k-r})$$

- $I_j I_{j+k-r} = 1$ iff $w_1 \ldots w_k w_{r+1} \ldots w_k$ occurs at position $j + k - r$ in τ.

- $E(I_j I_{j+k-r}) = \varepsilon_r \cdot \pi(w_1 \ldots w_k w_{r+1} \ldots w_k)$.

- $$\mathrm{Cov}(I_j, I_{j+k-r}) = E(I_j I_{j+k-r}) - E(I_j)E(I_{j+k-r})$$
$$= \varepsilon_r \cdot \pi(w_1 \ldots w_k w_{r+1} \ldots w_k) - (\pi(w))^2.$$

Note that this depends on r but not j.

The covariance sum becomes

$$\sum_{j,\ell:\ k \leqslant j < \ell \leqslant N} \mathrm{Cov}(I_j, I_\ell) = \sum_{r=1}^{k-1} \sum_{j=k}^{N-k+r} \left(\varepsilon_r \cdot \pi(w_1 \ldots w_k w_{r+1} \ldots w_k) - (\pi(w))^2 \right)$$

$$= \sum_{r=1}^{k-1} (N - 2k + r + 1) \left(\varepsilon_r \cdot \pi(w_1 \ldots w_k w_{r+1} \ldots w_k) - (\pi(w))^2 \right)$$

$$= \left(\sum_{r=1}^{k-1} \varepsilon_r \cdot (N - 2k + r + 1) \pi(w_1 \ldots w_k w_{r+1} \ldots w_k) \right)$$

$$- \left(\frac{((N - 2k + 2) + (N - k))(k - 1)}{2} (\pi(w))^2 \right)$$

Mean and variance of number of occurrences

Combining all the parts together and simplifiying gives

Mean number of occurrences

$$E(Y) \;=\; (N - k + 1)\, E(I_k) \;=\; (N - k + 1)\, \pi(w)$$

Variance of number of occurrences

$$\mathrm{Var}(Y) \;=\; (N - k + 1)\pi(w) - \left((2k - 1)N - 3k^2 + 4k - 1\right) (\pi(w))^2$$
$$+ 2 \sum_{r=1}^{k-1} \varepsilon_r \cdot (N - 2k + r + 1)\pi(w_1 \ldots w_k w_{r+1} \ldots w_k)$$

Computation for $w = w_1 \ldots w_9 = \texttt{ATAGATAGA}$ ($k = 9$) over all τ of length N

$$\pi(w) \;=\; p_A{}^5 p_T{}^2 p_G{}^2 \qquad\qquad \text{and } w \text{ self-overlaps at } r = 1, 5$$

$$E(Y) \;=\; (N - k + 1)\pi(w) = (N - 8)\pi(w) = (N - 8)p_A{}^5 p_T{}^2 p_G{}^2$$

$$
\begin{aligned}
\mathrm{Var}(Y) \;=\;& (N - k + 1)\pi(w) - \big((2k - 1)N - 3k^2 + 4k - 1\big)(\pi(w))^2 \\
& + 2\sum_{r=1}^{k-1} \varepsilon_r \cdot (N - 2k + r + 1)\pi(w_1 \ldots w_k w_{r+1} \ldots w_k) \\
=\;& (N - 8)\pi(w) - (17N - 208)(\pi(w))^2 \\
& + 2(N - 16)\pi(\texttt{ATAGATAG}\textcolor{red}{\textit{A}}\texttt{TAGATAGA}) \\
& + 2(N - 12)\pi(\texttt{ATAG}\textcolor{red}{\textit{ATAGA}}\texttt{TAGA}) \\
=\;& (N - 8)p_A{}^5 p_T{}^2 p_G{}^2 - (17N - 208)p_A{}^{10} p_T{}^4 p_G{}^4 \\
& + 2(N - 2k + 2)p_A{}^9 p_G{}^4 p_T{}^4 + 2(N - 2k + 6)p_A{}^7 p_G{}^3 p_T{}^3
\end{aligned}
$$

Frequencies of words and motifs in SARS

- The genome of SARS described previously has $N = 29751$ bases:

Nucleotide	Frequency	Proportion
A	8481	$p_A \approx 0.2851$
C	5940	$p_C \approx 0.1997$
G	6187	$p_G \approx 0.2080$
T	9143	$p_T \approx 0.3073$
Total	$N = 29751$	1

- These were used below to compute "Estimated" μ and σ.
- "Observed frequency" y was determined from the DNA sequence.

Word	Estimated		Observed		
	μ	σ	$y =$ Freq.	$z = (y - \mu)/\sigma$	$\Phi(z)$
GAGA	104.5456	10.6943	106	0.1360	0.5541
GCGA	73.2226	8.4830	37	-4.2700	10^{-5}
TGCG	78.9381	8.8018	59	-2.2652	0.0118
motif M	256.7064	17.6583	202	-3.0980	10^{-3}

(M consists of all three words; details on computing μ, σ are later.)

Hypothesis tests on frequencies in SARS

- We have not determined the complete distribution of Y. We will assume it is approximately normal with mean and standard deviation as computed above.

- That lets us compute Z and use it as a test statistic to see if the observed frequencies are consistent with a "random" sequence.

Three possible hypothesis tests

Null Hypothesis H_0: The genome sequence is generated by independent rolls of a 4-sided die with probabilities for each letter $p_A, \ldots, p_T$ as given previously.

vs. one of three alternative hypotheses:

H_1: The word w (or motif M) is over-represented.
H_2: The word w (or motif M) is under-represented.
H_3: The word w (or motif M) is over- or under-represented.

Hypothesis tests (at significance level $\alpha = 5\%$)

Word	Estimated		Observed		
	μ	σ	$y = $ Freq.	$z = (y - \mu)/\sigma$	$\Phi(z)$
GAGA	104.5456	10.6943	106	0.1360	0.5541
GCGA	73.2226	8.4830	37	-4.2700	10^{-5}
TGCG	78.9381	8.8018	59	-2.2652	0.0118
motif M	256.7064	17.6583	202	-3.0980	10^{-3}

- H_0 **vs.** H_1 **(over-represented).** Reject H_0 if Z is too big: $\Phi(Z) \geqslant 0.95$, so $Z \geqslant 1.6449$. In all the cases shown, we accept H_0 (a.k.a. "insufficient evidence to reject H_0").

- H_0 **vs.** H_2 **(under-represented).** Reject H_0 if Z is too small: $\Phi(Z) \leqslant 0.05$, so $Z \leqslant -1.6449$. By this test, GAGA is not under-represented, but each of GCGA, TGCG, and motif M, are considered to be under-represented.

- H_0 **vs.** H_3 **(under or over).** Reject H_0 if Z is too far away from 0: $\Phi(Z) \leqslant 0.025$ (so $Z \leqslant -1.96$) or $\Phi(Z) \geqslant 0.975$ (so $Z \geqslant 1.96$). We accept H_3 for GCGA, for TGCG, and for M, and accept H_0 for GAGA.

Critical regions (at significance level $\alpha = 5\%$)

- For `TGCG` & $N = 29751$, the null hypothesis gives $\mu = 78.9381$ and $\sigma = 8.8018$.
- The *critical region* (where we reject H_0) is blue. The *acceptance region* is white.
- The one-sided critical regions have area $\alpha = 0.05$.
  The two-sided critical regions have area $\alpha/2 = 0.025$ in each part.
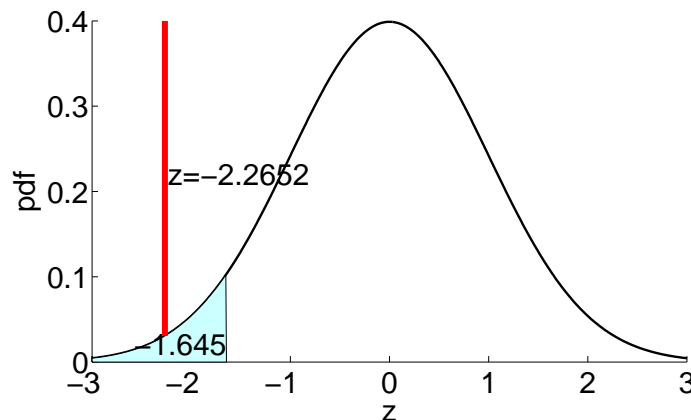- Our test statistic $y = 59$ or $z = -2.2652$ is shown as a red line.

H_1: Over-represented? H_2: Under-represented? H_3: Either over or under?
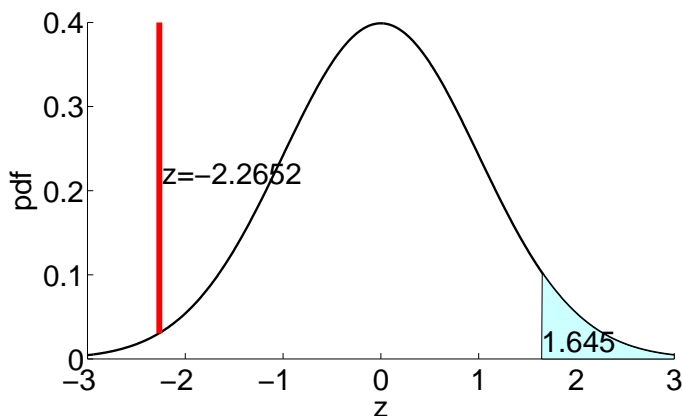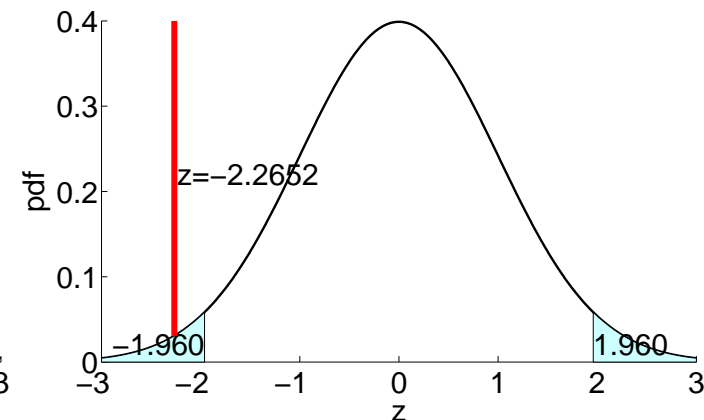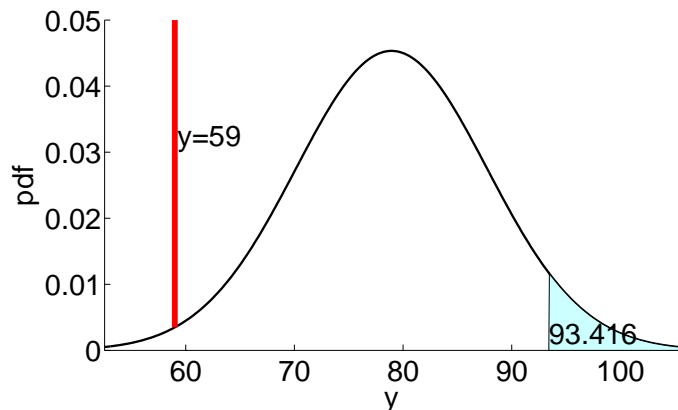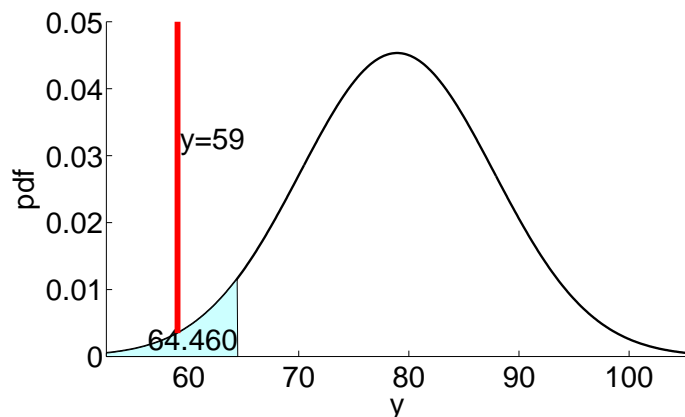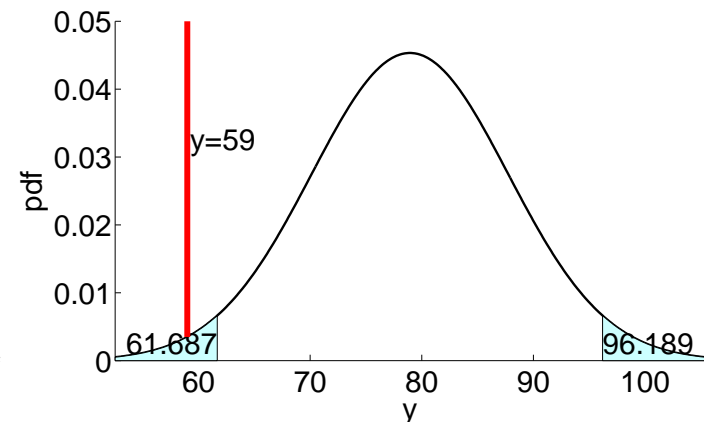
Same tests using P-values (at sig. level $\alpha = 5\%$)

- TGCG has $P(Z \leqslant -2.2652) = \Phi(-2.2652) = 0.0118$.

> - **H_0 vs. H_1 (over-represented?):**
> $$P = P(Z \geqslant -2.2652) = 1 - 0.0118 = 0.9881$$
> Since $P > \alpha$, we accept H_0 (TGCG is not over-represented).
>
> - **H_0 vs. H_2 (under-represented?):**
> $$P = P(Z \leqslant -2.2652) = 0.0118.$$
> Since $P \leqslant \alpha$, we accept H_2 (TGCG is under-represented).
>
> - **H_0 vs. H_3 (either of over or under?):**
> $$P = P(|Z| \geqslant 2.2652) = 2(0.0118) = 0.0236.$$
> Since $P \leqslant \alpha$, we accept H_3 (TGCG is over- or under-represented).

- P-values let us check any α easily.
  At $\alpha = 1\%$, all three tests accept H_0.
  At $\alpha = 2\%$, H_2 says it's under-represented but H_3 does not.

Motifs

- A *motif* is a set M of words that don't contain each other. Usually the words are very similar and have similar lengths.

- Suppose M has m words, all with length k:

$$M = \left\{ w^{(1)}, \dots, w^{(m)} \right\}.$$

- We'll work with an example of $m = 3$ words, each with $k = 4$ letters:

$$M = \left\{ \mathtt{GAGA}, \mathtt{TGCG}, \mathtt{GCGA} \right\}.$$

- When words of length k are generated at random by a 4-sided die, the total probability of the words in M is

$$\pi(M) = \pi(w^{(1)}) + \cdots + \pi(w^{(m)})$$

which is $p_A{}^2 p_G{}^2 + p_C p_G{}^2 p_t + p_A p_C p_G{}^2$ in this example.

Number of occurrences of a motif

- M occurs at position j in a nucleotide sequence τ if any of its words occurs (i.e., ends) there.

- Let $I_j = \begin{cases} 1 & \text{if } M \text{ occurs in } \tau \text{ at } j; \\ 0 & \text{otherwise.} \end{cases}$

- The number of occurrences of M in τ is $Y = I_k + \cdots + I_N$.

- Note that $E(I_j) = \pi(M)$ and
$$E(Y) = (N - k + 1)\,\pi(M)$$
by the same argument as for one word before.
For motifs of length $k = 4$, this becomes $E(Y) = (N - 3)\pi(M)$.

- In the variance formula, $\pi(w)$ is replaced by $\pi(M)$ as well, and we must recompute $\mathrm{Cov}(I_j, I_{j+k-r})$ to take into account overlaps between any two words of M.

Overlaps between words in a motif

- If the first r letters of $w^{(u)}$ equal the last r letters of $w^{(v)}$ ($r = 1, \ldots, k-1$):
  - Set $\varepsilon_r(u, v) = 1$;
  - let $w_r(u, v)$ be $w^{(v)}$ followed by $w^{(u)}$ but overlapped on the r letters;
  - let $\pi_r(u, v) = \pi(w_r(u, v))$.

  Otherwise, set $\varepsilon_r(u, v) = \pi_r(u, v) = 0$.

- For words $w^{(3)} = \texttt{GCGA}$ and $w^{(2)} = \texttt{TGCG}$, the overlaps are

$w^{(2)}$:	T G C G			
$r = 4$	G C G A	$\varepsilon_4(3,2) = 0$		
$r = 3$	*G C G* A	$\varepsilon_3(3,2) = 1$	$w_3(3,2) = \texttt{T}\textit{GCG}\texttt{A}$	$\pi_3(3,2) = \pi(\texttt{TGCGA})$
$r = 2$	G C G A	$\varepsilon_2(3,2) = 0$		
$r = 1$	*G* C G A	$\varepsilon_1(3,2) = 1$	$w_1(3,2) = \texttt{TGC}\textit{G}\texttt{CGA}$	$\pi_1(3,2) = \pi(\texttt{TGCGCGA})$

($r = 4$ was shown, although we're only supposed to go up to $r = k - 1 = 3$.)

Overlap between words in a motif

$\varepsilon_r(u,v)$ $w_r(u,v)$	$v = 1$ $w^{(1)} = \texttt{GAGA}$	$v = 2$ $w^{(2)} = \texttt{TGCG}$	$v = 3$ $w^{(3)} = \texttt{GCGA}$
$u = 1$ $w^{(1)} = \texttt{GAGA}$	$\varepsilon_1(1,1) = 0$ $\varepsilon_2(1,1) = 1$ GA*GA*GA $\varepsilon_3(1,1) = 0$	$\varepsilon_1(1,2) = 1$ TGC*G*AGA $\varepsilon_2(1,2) = 0$ $\varepsilon_3(1,2) = 0$	$\varepsilon_1(1,3) = 0$ $\varepsilon_2(1,3) = 1$ GC*GA*GA $\varepsilon_3(1,3) = 0$
$u = 2$ $w^{(2)} = \texttt{TGCG}$	$\varepsilon_1(2,1) = 0$ $\varepsilon_2(2,1) = 0$ $\varepsilon_3(2,1) = 0$	$\varepsilon_1(2,2) = 0$ $\varepsilon_2(2,2) = 0$ $\varepsilon_3(2,2) = 0$	$\varepsilon_1(2,3) = 0$ $\varepsilon_2(2,3) = 0$ $\varepsilon_3(2,3) = 0$
$u = 3$ $w^{(3)} = \texttt{GCGA}$	$\varepsilon_1(3,1) = 0$ $\varepsilon_2(3,1) = 0$ $\varepsilon_3(3,1) = 0$	$\varepsilon_1(3,2) = 1$ TGC*G*CGA $\varepsilon_2(3,2) = 0$ $\varepsilon_3(3,2) = 1$ T*GCG*A	$\varepsilon_1(3,3) = 0$ $\varepsilon_2(3,3) = 0$ $\varepsilon_3(3,3) = 0$

Dependence between positions

- $I_j I_{j+k-r} = 1$ if there are some overlapping words ($\varepsilon_r(u,v) = 1$) whose combination word $w_r(u,v)$ actually occurs in τ at $j+k-r$.
- $I_j I_{j+k-r} = 0$ if nothing of that form occurs at $j+k-r$.
- So

$$E(I_j I_{j+k-r}) = \sum_{u=1}^{m} \sum_{v=1}^{m} \varepsilon_r(u,v) \pi_r(u,v)$$

replaces the analogous term for the one word case, leading to

Variance of number of occurrences of a motif

$$
\begin{aligned}
\mathrm{Var}(Y) \;=\; & (N-k+1)\pi(M) \\
& -((2k-1)N - 3k^2 + 4k - 1)(\pi(M))^2 \\
& +2\sum_{r=1}^{k-1}(N - 2k + r + 1)\sum_{u=1}^{m}\sum_{v=1}^{m} \varepsilon_r(u,v) \cdot \pi_r(u,v)
\end{aligned}
$$

Example

$M = \{\texttt{GAGA}, \texttt{TGCG}, \texttt{GCGA}\}$ has $m = 3$ words of length $k = 4$, and 5 overlaps

$$
\begin{aligned}
\pi(M) &= \pi(\texttt{GAGA}) + \pi(\texttt{TGCG}) + \pi(\texttt{GCGA}) \\
E(Y) &= (N-3)\,\pi(M) \\
\mathrm{Var}(Y) &= (N-3)\pi(M) - (7N-33)(\pi(M))^2 \\
&\quad + 2(N-5)\pi(\texttt{GA}\textcolor{red}{\texttt{GA}}\texttt{GA}) + 2(N-6)\pi(\texttt{TGC}\textcolor{red}{\texttt{G}}\texttt{AGA}) \\
&\quad + 2(N-5)\pi(\texttt{GC}\textcolor{red}{\texttt{GA}}\texttt{GA}) + 2(N-6)\pi(\texttt{TGC}\textcolor{red}{\texttt{G}}\texttt{CGA}) \\
&\quad + 2(N-4)\pi(\texttt{TGCGA})
\end{aligned}
$$

If all nucleotides have equal probability $1/4$, this becomes

$$
\begin{aligned}
\pi(M) &= 3/4^4 = 3/256 \\
E(Y) &= (N-3)\,(3/256) = 3(N-3)/256 \\
\mathrm{Var}(Y) &= (N-3)(3/256) - (7N-33)(9/65536) \\
&\quad + 2(N-5)4^{-6} + 2(N-6)4^{-7} \\
&\quad + 2(N-5)4^{-6} + 2(N-6)4^{-7} + 2(N-4)4^{-5} \\
&= (913N - 2935)/65536
\end{aligned}
$$

Repeats in *C. elegans* that facilitate homologous pairing in meoisis

Sanford and Perry, *Nucleic Acids Research,* 2001, 29(14):2920-2926.

- **1998:** *C. elegans* is the first multicellular organism completely sequenced. 6 chromosomes, 13–21 Mb each, 100 Mb total.

- **NAR 2001:** Christopher Sanford and Marc Perry (U. Toronto) count all k-mers in *C. elegans* for $2 \leqslant k \leqslant 20$, looking for those over-represented on just one chromosome, plus other constraints.

- They found one unique candidate per chromosome, and speculate these facilitate homologous pairing during meiosis:

Chr.	DNA Seq.	# on that chr. (# per Mb)	# on other (# per Mb)
I	TTGGTTGAGGCT	611 (44.1)	201 (2.5)
II	TTTGTAGTCTAGCA	152 (10.3)	54 (0.7)
III	TGCTAAATATTTAGCA	197 (15.4)	1 (0.0)
IV	GTATAATCATG	347 (21.5)	251 (3.2)
V	TGGGCGCTGCT	713 (34.2)	13 (0.2)
X	TGGTCAGTGCA	335 (19.4)	74 (0.9)

- **RECOMB 2007:** Abby Dernburg (UC Berkeley) announces her lab proved it experimentally (but some k-mers were slightly adjusted).