TAMS 23: Statistiska metoder i bioinformatik Föreläsning 6: Modellering av DNA och Ord i DNA

Timo Koski

Matematisk statistik /Linköpings tekniska högskola

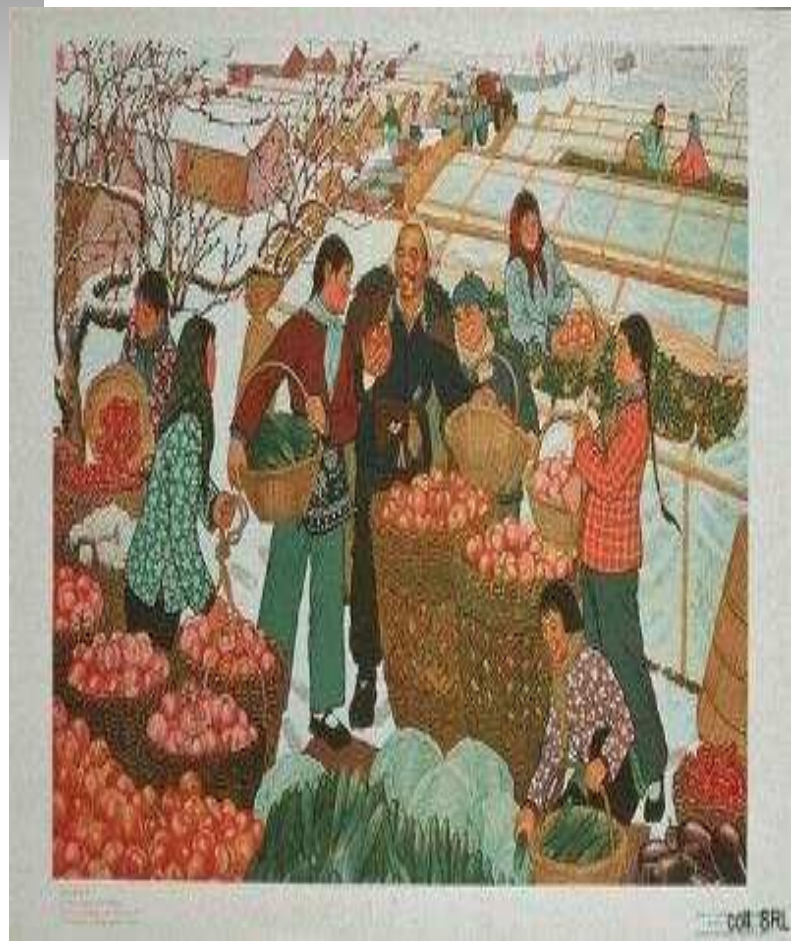
Lecture 6

The lecture is base on portions of chapter 5 and on section 10.4 of Ewens and Grant. The lecture is, however, as far as the occurrences (or appearances) of DNA words is concerned, a mix of the original research contributions and the material in the textbook.
The sections 5.1 ’shotgun sequencing‘, section 5.4 ’long repeats‘, section 5.5 r-scans are offered as possible examination items, ’presentationer’.

Lecture 6: Contents

1) Weight Matrix Model
2) Markov Modelling
3) Words in a DNA sequence
   (A) The number of occurrences
   (B) The length between one occurrence of a word and the next
   (C) The waiting time till appearance of a word including the Markovian case.
   (D) p.g.f. for the waiting time till appearance of a word.

An Image of Practical Life Science



Weight Matrix Model

A weight matrix $\mathcal{M}_0$ is a simple model often used by molecular biologicists as a representation for a *family of signals*. The sequences containing the signals are supposed to have equal length (=n) and to have no *gaps* (no positions are blank).

Weight Matrix Model

A weight matrix $\mathcal{M}_0$ has as entries the probabilities $p_i(x_j)$ (e.g. *observed relative frequency*) for that a string should have one of the bases

$$\{x_1, x_2, x_3, x_4\} = \{A, T, C, G\}$$

at position i:

$$\mathcal{M}_0 : \begin{matrix} p_1(x_1) & \dots & p_n(x_1) \\ p_1(x_2) & \dots & p_n(x_2) \\ p_1(x_3) & \dots & p_n(x_3) \\ p_1(x_4) & \dots & p_n(x_4) \end{matrix} .$$

The weight matrix model is often called a *profile*.

Prob(Data | Model)

The probability of a finite sequence $\mathbf{x} = x_{l_1} x_{l_2} \ldots x_{l_n}$ given the model $\mathcal{M}_0$ is given by

$$P\left(\mathbf{x}|\mathcal{M}_0\right) = \prod_{j=1}^{4} \prod_{i=1}^{n} p_i\left(x_j\right)^{I_{i,x_j}(\mathbf{x})},$$

where the indicator $I_{i,x_j}\left(\mathbf{x}\right)$, a function of $\mathbf{x}$, is 0 if $x_j \neq x_{l_i}$, i.e., if the symbol x_j does not appear in position i in the string $\mathbf{x}$ and is 1 otherwise. Thus the bases in the different positions are *independent* given $\mathcal{M}_0$.

Prob(Data | Model)

A sequence of strings $\mathbf{x}^1, \ldots, \mathbf{x}^t$ is training data, i.e., of known cases of members of a signal family. We take them to be generated *independently* given $\mathcal{M}_0$, is by multiplication of the preceding expressions assigned the probability

$$P\left(\mathbf{x}^1, \ldots, \mathbf{x}^t | \mathcal{M}_0\right) = \prod_{s=1}^{t} P\left(\mathbf{x}^s | \mathcal{M}_0\right)$$

$$= \prod_{j=1}^{4} \prod_{i=1}^{n} p_i\left(x_j\right)^{n_i(x_j)},$$

where $n_i\left(x_j\right)$ is the number of times the symbol x_j ap-

Prob(Data | Model$_{\mathrm{MLE}}$)

$$P\left(\mathbf{x}^1,\ldots,\mathbf{x}^t|\mathcal{M}_0\right)=\prod_{s=1}^{t}P\left(\mathbf{x}^s|\mathcal{M}_0\right)$$

$$=\prod_{j=1}^{4}\prod_{i=1}^{n}p_i\left(x_j\right)^{n_i\left(x_j\right)},$$

where $n_i\left(x_j\right)$ is the number of times the symbol x_j appears on position i in $\mathbf{x}^1,\ldots,\mathbf{x}^t$. The maximum likelihood estimate is

$$\widehat{p_i}\left(x_j\right)=\frac{n_i\left(x_j\right)}{n}$$

This will be shown during a later 'lektion'.

Example: Promoter Regions

RNA polymerase molecules start transcription by recognizing and binding to promoter regions upstream of the desired transcription start sites. Unfortunately promoter regions do not follow a strict pattern. It is possible to find a DNA sequence (called the consensus sequence) to which all of them are very similar.

Example: Promoter Regions

For example, the consensus sequence in the bacterium *E. Coli*, based on the study of 263 promoters, is TTGACA followed by 17 random base pairs followed by TATAAT, with the latter located about 10 bases upstream of the transcription start site. None of the 263 promoter sites exactly match the above consensus sequence.

Weight Matrix Model

By constructing the weight matrix of the TATAAT region (n=6) we can compute the probability of a DNA sequence $\mathbf{x} = x_{l_1} x_{l_2} \ldots x_{l_6}$ and compute the probability

$$\mathrm{Prob}(\mathbf{x} \mid 'promoter')$$

This means that successive bases are thought as being generated independently from the distributions in the weight matrix table. Similarly, we can compute using, e.g., a weight matrix model of the signal family

$$\mathrm{Prob}(\mathbf{x} \mid 'non-promoter')$$

with weight matrix from a non-promoter region and compare

$$\frac{\mathrm{Prob}(\mathbf{x} \mid 'promoter')}{\mathrm{Prob}(\mathbf{x} \mid 'non-promoter')}.$$

to decide in $\mathbf{x}$ is a member of the family.

Markov Model $\mathcal{M}_1$

A Markov model is a statistical 'counter hypothesis' to weight models. The

	A	C	G	T
A	$p_{\mathrm{A}\|\mathrm{A}}$	$p_{\mathrm{A}\|\mathrm{C}}$	$p_{\mathrm{A}\|\mathrm{G}}$	$p_{\mathrm{A}\|\mathrm{T}}$
C	$p_{\mathrm{C}\|\mathrm{A}}$	$p_{\mathrm{C}\|\mathrm{C}}$	$p_{\mathrm{C}\|\mathrm{G}}$	$p_{\mathrm{A}\|\mathrm{T}}$
G	$p_{\mathrm{G}\|\mathrm{A}}$	$p_{\mathrm{G}\|\mathrm{C}}$	$p_{\mathrm{A}\|\mathrm{G}}$	$p_{\mathrm{G}\|\mathrm{T}}$
T	$p_{\mathrm{T}\|\mathrm{A}}$	$p_{\mathrm{T}\|\mathrm{C}}$	$p_{\mathrm{T}\|\mathrm{G}}$	$p_{\mathrm{T}\|\mathrm{T}}$

matrix $\mathbf{P}$ contains $16 - 4$ unknown probabilities that will have to be learned from training data, i.e., of known cases of member of a signal family.

Prob $(\text{Data} \mid \mathcal{M}_1)$

If

$$\{X_n\}_{n=0}^{\infty} \in \text{Markov}\left(P, p_{X_0}\right),$$

where

$$p_{X_0} = \left(P\left(X_0 = 1\right), \ldots, P\left(X_0 = J\right)\right),$$

then

$$P\left(X_0 = j_0, X_1 = j_1, \ldots, X_n = j_n\right) = p_{X_0}\left(j_0\right) \prod_{l=1}^{n} p_{j_{l-1}|j_l}.$$

An Image of Practical Life Science

Learning with Markov chains

$$\mathbf{P} = \begin{pmatrix} p_{1|1} & p_{1|2} & p_{1|2} & p_{1|4} \\ p_{2|1} & p_{2|2} & p_{2|3} & p_{2|4} \\ p_{3|1} & p_{3|2} & p_{3|3} & p_{3|4} \\ p_{4|1} & p_{|2} & p_{4|3} & p_{4|4} \end{pmatrix}$$

We change the interpretation: the function $p(\mathbf{x}|\mathcal{M}_1)$ is regarded as a function of $\mathbf{P}$ (or the probabilities in $\mathbf{P}$) and called a *likelihood function* and denoted by $L_{\mathbf{x}}(\mathbf{P})$

$$L(\mathbf{P}) = p_{j_0}(0) \prod_{l=1}^{n} p_{j_{l-1}|j_l}.$$

The maximum likelihood estimate of $\mathbf{P}$ is obtained by maximizing $L(\mathbf{P})$ as a function of $\mathbf{P}$.

Learning with Markov chains

The maximum likelihood estimate $\widehat{p}_{i|j}$ of $p_{i|j}$ is

$$\widehat{p}_{i|j} = \frac{n_{i|j}}{n_i}, \text{ for all } i \text{ and } j.$$

Here $n_{i|j}$ is the number of times the sequence contains the pair of bases (i, j) (in this order), i.e., the number of transitions from i to j and n_i is the number of times the base i occurs in the sequence.

This will be shown during a later 'lektion'.

Modelling with Markov chains

$$\widehat{p}_{i|j} = \frac{n_{i|j}}{n_i}, \text{ for all } i \text{ and } j.$$

Here $n_{i|j}$ is the number of times the sequence contains the pair of bases (i, j) (in this order), i.e., the number of transitions from i to j and n_i is the number of times the base i occurs in the sequence.
Hence Markov models assume that there is biological information contained in the frequency of pairs of bases following each other.

An Image of Practical Life Science

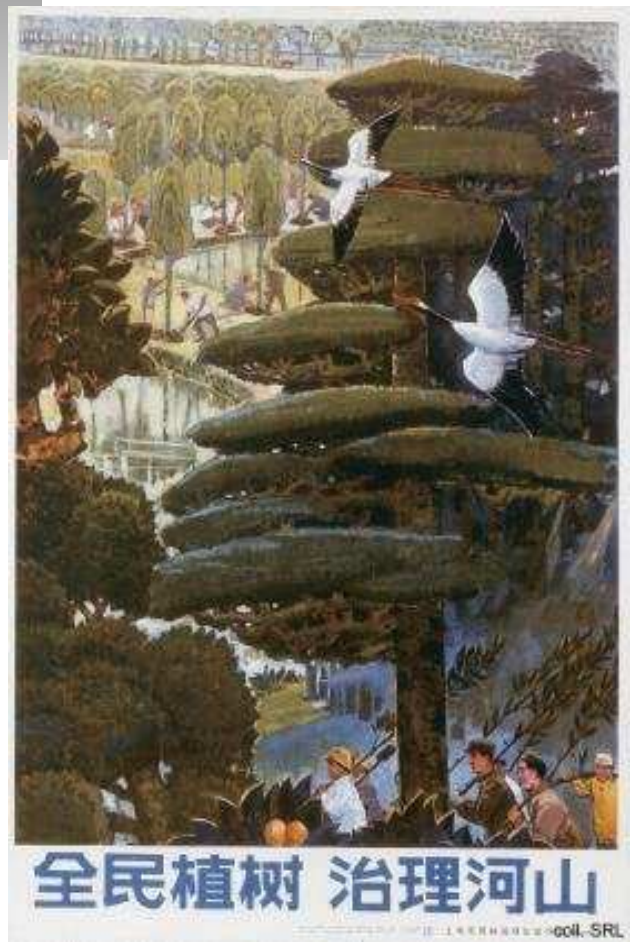
Learning & Parametric Inference

In maximum likelihood estimation we have regarded the transition probabilities as *parameters* and used training data to *infer* their values.

Inference: the process of deriving a conclusion from from fact and/or premise.

In probabilistic modelling of sequences the facts are the observed sequences, the premise is represented by the model and the conclusions concern unobserved quantities.

Another Image of Practical Life Science



Words in a DNA sequence

Word means here a subsequence of a larger sequence. This is a unique subsequence, not a family of signals.

Words in A DNA sequence

Frequency of occurrence of specific short nucleotide words (like GAGA) within part or all of a nucleotide sequence is of interest for several biological reasons.

The occurrence of a word in unexpectedly large numbers in a segment may provide information about structure or function.

Words in a DNA sequence

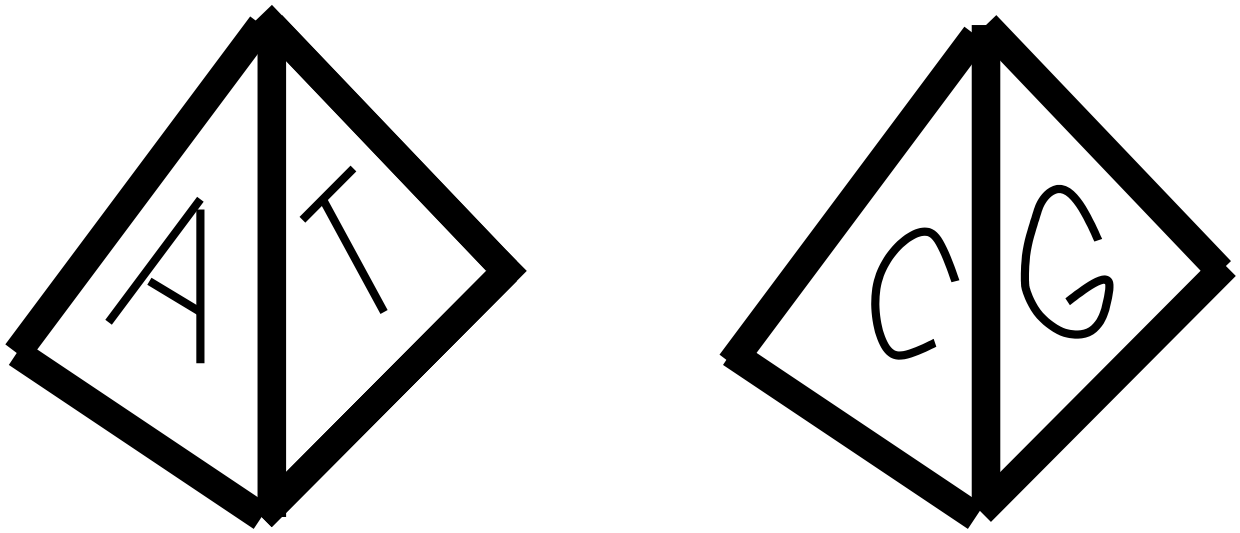
Information about structure or function:

- a segment of DNA may have originally been formed by replication of smaller segments. Such replication might be exact initially but might be altered over time. But in regions where retention of function is necessary, exact repeats would be expected to occur.
- the occurrence of of repeated subsequences may identify locations where an insert has been introduced into the DNA sequence.

DNA Words : statistical assumptions

Let Z assume values in $\{A, T, C, G\}$ and let $Z \in U(1, 4)$. Z is a nucleotide chosen at random. Hence the probability of the occurrence of a word of length L is

$$p_L = \left(\frac{1}{4}\right)^L .$$

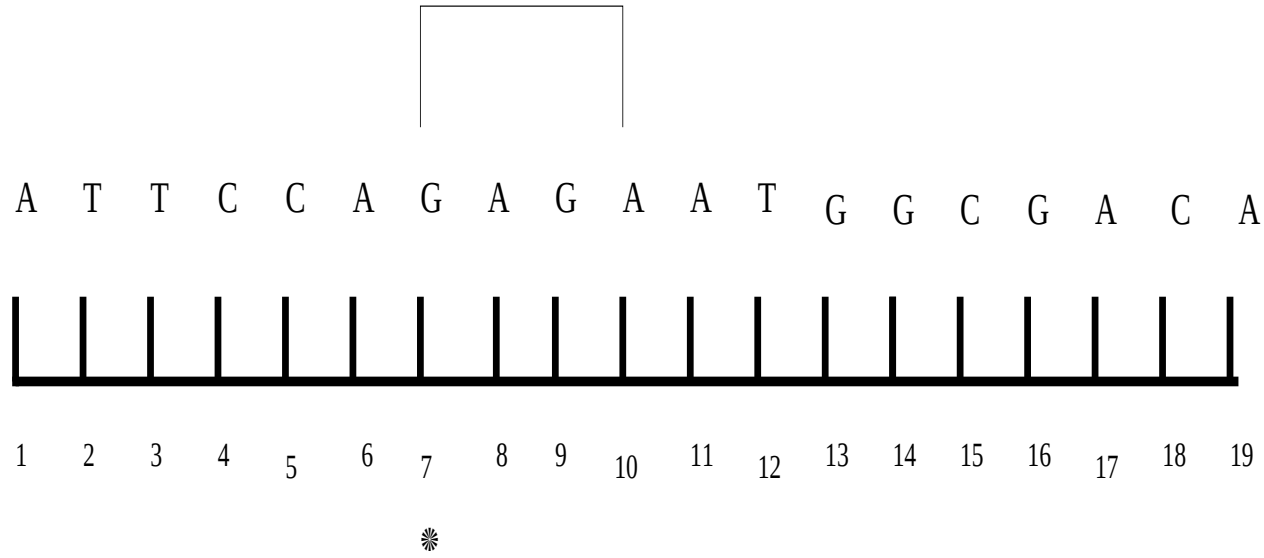


DNA Words : statistical assumptions

The DNA dice is biologically unreasonable as a model for a whole genome, but can serve well as a statistical null hypothesis to detect important deviations from random noise.

(A) Occurrence of a word

We say that ’a word occurs at position i‘ (Like GAGA in position 7 in the figure) if it is found to begin at position i in a longer sequence.

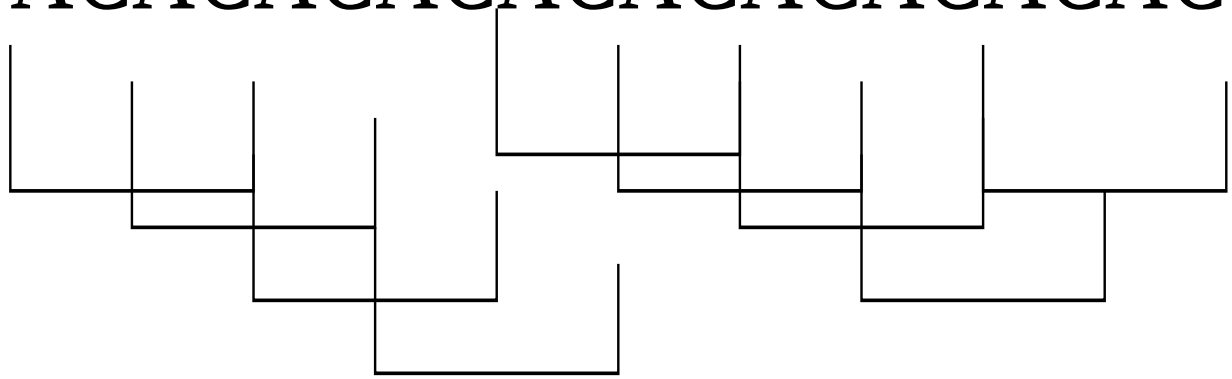


(A) Occurrence of a word

Let the random variable $Y_1(M)$ be the number of occurrences of a nucleotide word (like GAGA) of length L (like $L = 4$) within a nucleotide sequence of length M, $L \leq M$. Then $n = M - L + 1$ is the maximum value achievable by $Y_1(M)$.

Occurrence of a word: overlap capability (examples)

The word ACAC cannot occur more than 9 times in a sequence of length 20, because of



its ’overlap capability‘.

Occurrence of a word: overlap capability (examples)

The word AAAA can occur between 0 and 17 times in a sequence of length 20, because of its greater ’overlap capability‘. The word ACGT has no overlap capability except in the trivial case.

Occurrence of a word

Let the random variable $Y_1(M)$ be the number of occurrence of a nucleotide word of length L within a nucleotide sequence of length M. M is thought to be so large that end effects can be neglected. Let its probability function be denoted by

$$p_{Y_1(M)}(y; L, M, \mathbf{w}).$$

This does depend on the overlap capability ('$\mathbf{w}$' (to be defined)) and is therefore not a binomial probability function.

Overlap capability

We define the overlap capability of a word as follows.
Let $\mathbf{s}$ be a word, $\mathbf{s} = s_1 \ldots s_L$, so that s_i are its nucleotides read from left to right.
The overlap capability $\mathbf{w}$ is a binary sequence $\mathbf{w} = w_1 \ldots w_L$, where w_i is defined as follows:
$w_i = 1$ if it is possible for the word's first i letters to overlap its last i letters (in the same order) and $w_i = 0$ otherwise.
Overlap capability is also known as the autocorrelation of a word.

Overlap capability

The overlap capability $\mathbf{w}$ of $\mathbf{s} = s_1 \dots s_L$ is a binary sequence $\mathbf{w} = w_1 \dots w_L$, where w_i is defined as follows: $w_i = 1$, if it is possible for the word's first i letters to be equal to its last i letters (in the same order) and $w_i = 0$ otherwise.
More formally

$$w_i = \begin{cases} 1 & \text{if } s_k = s_{L+k-i}, k = 1, \dots, i \\ 0 & \text{elsewhere.} \end{cases}$$

Overlap capability: examples (Ewens & Grant p. 169)

$L = 4$

$$w_i = \begin{cases} 1 & \text{if } s_k = s_{4+k-i}, k = 1, \ldots, i \\ 0 & \text{elsewhere.} \end{cases}$$

	w_1	w_2	w_3	w_4
GAGA	0	1	0	1
GGGG	1	1	1	1
GAAG	1	0	0	1
GAGC	0	0	0	1

Number of Occurrences of a Word

We are not going to derive the probability function

$$p_{Y_1(M)}(y; L, M, \mathbf{w})$$

for $Y_1(M)$, the number of occurrence of a nucleotide word of length L within a nucleotide sequence of length M. Instead we are going find

- $E\left[Y_1(M)\right]$
- $\mathrm{Var}\left[Y_1(M)\right]$

$E[Y_1(M)]$, Expected Number of Occurrences of a Word

Let $n = M - L + 1$ and $i = 1, \ldots, n$

$$I_i = \begin{cases} 1 & \text{if the word begins in position } i \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$Y_1(M) = I_1 + I_2 + \ldots I_n,$$

and

$$E[Y_1(M)] = E[I_1] + E[I_2] + \ldots E[I_n] =$$

$$= \sum_{i=1}^{n} \text{Prob}\,(\text{ the the word begins in position } i\,) = n \cdot p_L = n\left(\frac{1}{4}\right)^L.$$

$E[Y_1(M)]$, *Expected Number of Occurrences of a Word*

The expected number of occurrences of a word

$$E[Y_1(M)] = n\left(\frac{1}{4}\right)^L.$$

does not depend on the overlap capability !

$\mathrm{Var}\left[Y_1(M)\right]$, *variance of the Number of Occurrences of a Word*

$$\mathrm{Var}\left[Y_1(M)\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}\left(I_i I_j\right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} E\left(I_i I_j\right) - \sum_{i=1}^{n} E\left(I_i\right) \sum_{j=1}^{n} E\left(I_j\right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} E\left(I_i I_j\right) - n^2 \left(\frac{1}{4}\right)^{2L}$$

$E\left(I_i I_j\right)$

The correlations between the indicators I_i and I_j that are near neighbors depend on $\mathbf{w}$.

$$E\left(I_i I_{i+k}\right)$$

is just the probability that the word occcurs at both position i and position $i+k$.

- If $0 \leq k \leq \min(L-1, n-1)$, then

$$E(I_i I_{i+k}) = w_{L-k} \left(\frac{1}{4}\right)^{L+k}$$

- If $\min(L-1, n-1) \leq k \leq n$, then $\mathbf{w}$ is irrelevant, and

$$E(I_i I_{i+k}) = \left(\frac{1}{4}\right)^{2L}$$

Counting the cases in $\sum_{i=1}^{n}\sum_{j=1}^{n} E\left(I_i I_j\right)$

- There are n terms among the n^2 in $\sum_{i=1}^{n}\sum_{j=1}^{n} E\left(I_i I_j\right)$ such that $i = j$.
- If $n > 1$, there are $2(n-k)$ terms such that $j = i + k$ or $i = j + k$ for $k = 1, \ldots, \min\left(L-1, n-1\right)$.
- If $n > L$, there are $2\sum_{k=1}^{n-L} k = (n-L)(n-L+1)$ remaining terms that do not depend on $\mathbf{w}$.

$$\sum_{i=1}^{n}\sum_{j=1}^{n} E\left(I_i I_j\right)$$

$$= np_L + p_L^2(n-L)(n-L+1) + 2p_L^2 \sum_{k=1}^{\min(L-1,n-1)} (n-k) w_{L-k} \left(\frac{1}{4}\right)^k$$

$\mathrm{Var}\,[Y_1(M)]$, *variance of the Number of Occurrences of a Word*

$$\mathrm{Var}\,[Y_1(M)]$$

$$= np_L(1-np_L)$$

$$+p_L^2(n-L)(n-L+1)+2p_L^2\sum_{k=1}^{\min(L-1,n-1)}(n-k)w_{L-k}\left(\frac{1}{4}\right)^k.$$

In the right hand side the second term equals 0 if $n \leq L$ and the third term equals 0 if $n = 1$. If $L = 1$, the third term in the equation equals zero, and the variance is that of a binomial R.V..

$\mathrm{Var}\,[Y_1(M)]$, ***generalization***

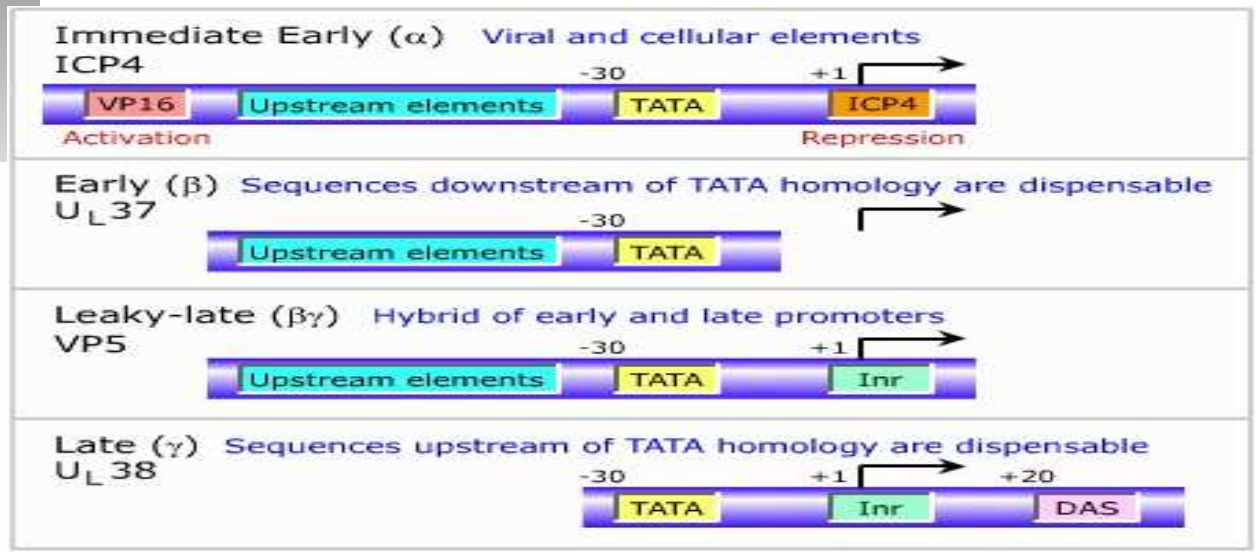
Let the the alphabet is generic (J symbols) and the probabilities be $p_1, \ldots p_J$, and $\mathbf{s}$ be a word with L letters with probabilities p_{j_m}, $m = 1, \ldots, L$. Let $P_k = \prod_{m=1}^{k} p_{j_m}$ be the product of probabilities of the first k letters in $\mathbf{s}$. Then we insert in the preceding formula P_L for $\left(\frac{1}{4}\right)^L$ and P_k for $\left(\frac{1}{4}\right)^k$.

Reference

The results about $E[Y_1(M)]$ and $\mathrm{Var}[Y_1(M)]$ presented above are due to
J. F. Gentleman and R. C. Mullin: The Distribution of the Frequency of Occurrence of Nucleotide Subsequences, Based on their Overlap Capability. *Biometrics*, 45, pp. 35–52, 1989.
Gentleman and Mullin find also the full probability distribution

$$p_{Y_1(M)}(y; L, M, \mathbf{w}).$$

(B) The length between one occurrence of a word and the next



(B) The length between one occurrence of a word and the next

Let $\mathbf{s}$ be word, $\mathbf{s} = s_1 \dots s_L$ with overlap capability $\mathbf{w}$.
$\mathbf{w} = w_1 \dots w_L$.
Let Y_2 be the distance to the next occurrence of $\mathbf{s}$ after $\mathbf{s}$ has occurred at i.
Let

$$p_{Y_2}(y) \stackrel{def}{=} P\left(Y_2 = y\right),$$

We are going to find a recursive formula for $p_{Y_2}(y)$.

Some pertinent events

Let Y_2 be the distance to the next occurrence of $\mathbf{s}$ after $\mathbf{s}$ has occurred at i.

$$p_{Y_2}(y) = P\left(Y_2 = y\right),$$

Let

$F =$ $\mathbf{s}$ occurs at $i + y$ but nowhere between i and $i + y$

Then $P\left(F\right) = p_{Y_2}(y)$.

More pertinent events

$F =$ $\mathbf{s}$ occurs at $i + y$ but nowhere between i and $i + y$

and let

$E =$ $\mathbf{s}$ occurs at $i + y$

$A_j =$ $\mathbf{s}$ occurs at $i + y$ and $i + j$ but nowhere between i and $i + j$

Then

$$E = F \cup A_1 \cup A_2 \cup \ldots \cup A_{y-1}$$

Decomposition of events

Decomposition

$$E = F \cup A_1 \cup A_2 \cup \ldots \cup A_{y-1}$$

yields, since the events are in the right hand side are disjoint,

$$p_{Y_2}(y) = P(F) = P(E) - \sum_{j=1}^{y-1} P(A_j).$$

We shall now compute the probabilities in the right hand side of this expression.

Decomposition of events: the cases $y \in \{1, \ldots, L-1\}$

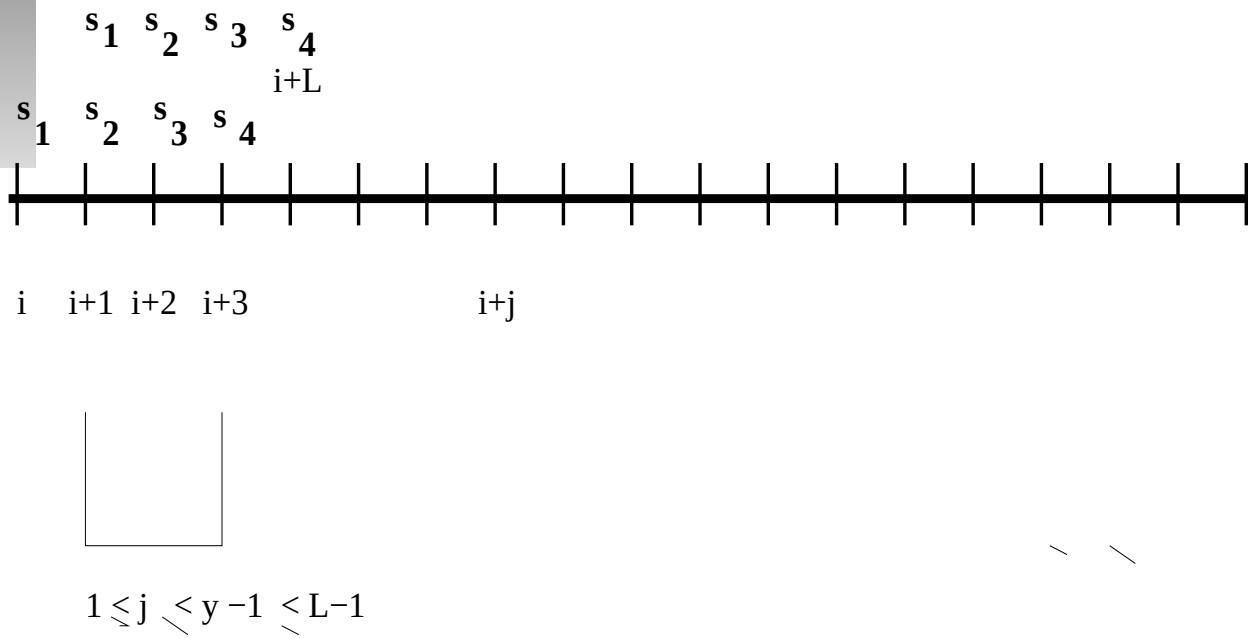
Let $1 \leq y \leq L-1$.

$$E = F \cup A_1 \cup A_2 \cup \ldots \cup A_{y-1}$$

$$p_{Y_2}(y) = P(F) = P(E) - \sum_{j=1}^{y-1} P(A_j).$$

Decomposition of events: the cases

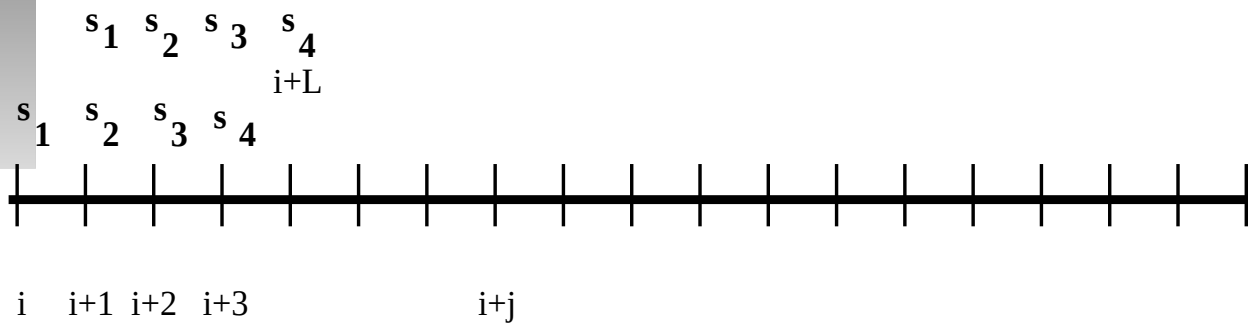
$y \in \{1, \ldots, L-1\}$, $P(E)$

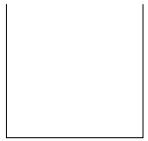


$$P(E) = w_{L-y} \left(\frac{1}{4}\right)^{y}$$

Decomposition of events: the cases

$y \in \{1, \ldots, L-1\}$, $P(A_j)$





$1 \leq j \leq y-1 \leq L-1$

For $1 \leq j \leq y-1 \leq L-1$ we have that

$$P(A_j) = p_{Y_2}(j) w_{L+j-y} \left(\frac{1}{4}\right)^{y-j},$$

since the overlaps are to be taken into account.

$p_{Y_2}(y)$ *for* $y \in \{1, \ldots, L-1\}$

$$p_{Y_2}(y) = P(F) = P(E) - \sum_{j=1}^{y-1} P(A_j);$$

given for $1 \leq y \leq L-1$ we have

$$p_{Y_2}(y) = w_{L-y} \left(\frac{1}{4}\right)^y - \sum_{j=1}^{y-1} p_{Y_2}(j) w_{L+j-y} \left(\frac{1}{4}\right)^{y-j}.$$

Decomposition of events: the cases

$y \geq L$, $P(E)$

Let $L \leq y$.

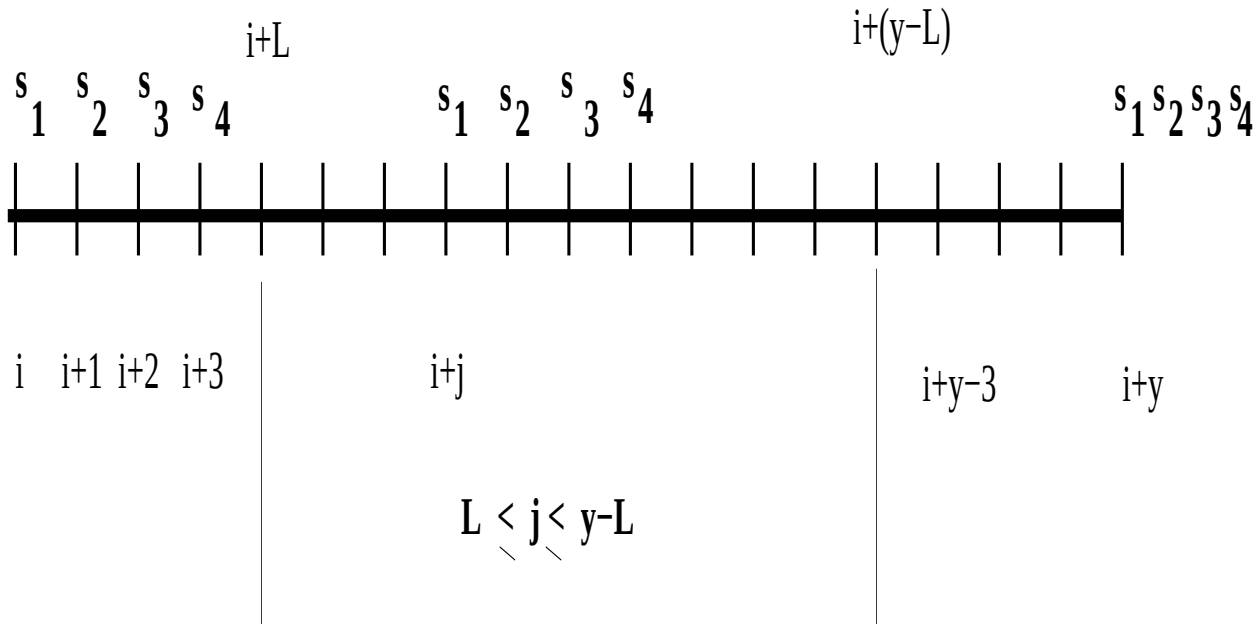
$$P(E) = p_L = \left(\frac{1}{4}\right)^L$$

Probability $P(A_j)$; $L \leq j \leq y - L$

For $L \leq j \leq y - L$ we have that

$$P(A_j) = p_{Y_2}(j) p_L,$$

since the events that $\mathbf{s}$ occurs at $i + y$ and $i + j$ but nowhere between i and $i + j$ are in this case independent.

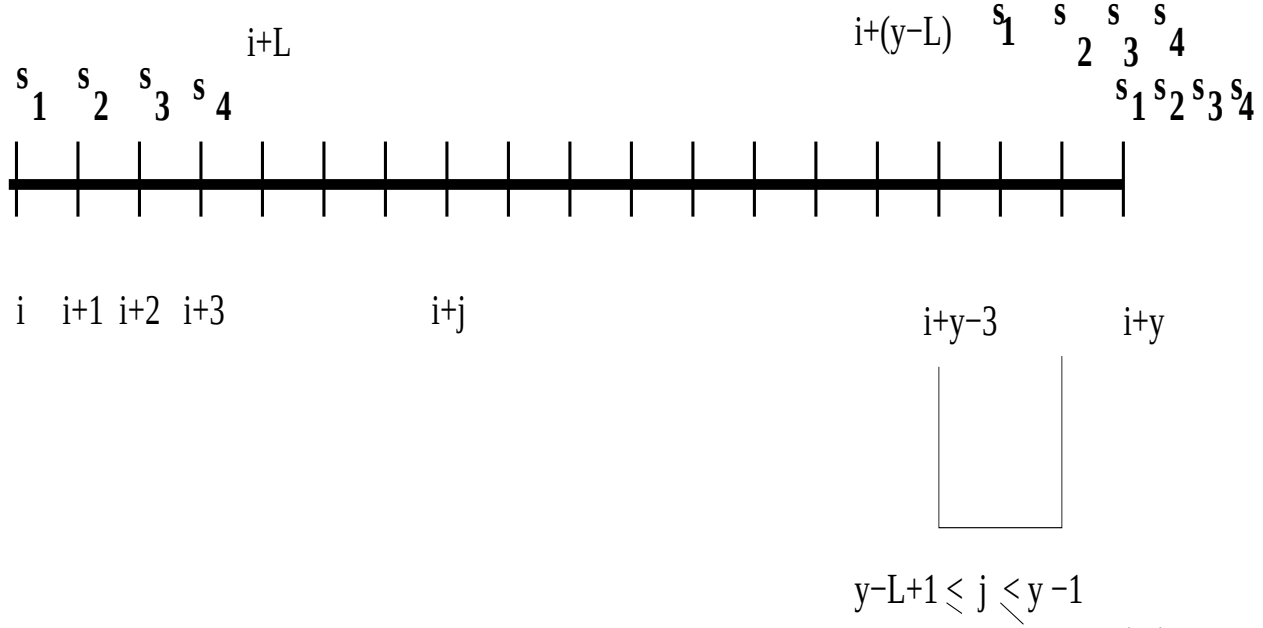


Probability $P(A_j)$ **;** $y - L + 1 \leq j \leq y - 1$

For $y - L + 1 \leq j \leq y - 1$ we have that

$$P(A_j) = p_{Y_2}(j) w_{L+j-y} \left(\frac{1}{4}\right)^{y-j},$$

since for $y - L + 1 \leq j \leq y - 1$ overlaps are to be taken



into account.

$p_{Y_2}(y)$ *for* $y \geq L$

Hence for $L \leq y$.

$$p_{Y_2}(y) = \left(\frac{1}{4}\right)^L - \left(\frac{1}{4}\right)^L \sum_{j=1}^{y-L} p_{Y_2}(j) - \sum_{j=y-L+1}^{y-1} p_{Y_2}(j) w_{L+j-y} \left(\frac{1}{4}\right)^{y-}$$

and the recursion of Ewens and Grant is complete.

occurrence of a word: beginning at the origin

$$p_{Y_3}(y) = \left(\frac{1}{4}\right)^L - \left(\frac{1}{4}\right)^L \sum_{j=L}^{y-L} p_Y(j) - \sum_{j=y-L+1}^{y-1} p_Y(j) w_{L+j-y} \left(\frac{1}{4}\right)^{y-j}$$

This formula is due to Gunnar Blom and Daniel Thorburn (1982), as the probability function of the the waiting time Y_3 until the word has appeared, in the sense that all the symbols have been seen, hence $p_{Y_3}(y) = 0$ for $y \leq L$.

The waiting time of a word: The Markov Case

We consider a stationary Markov chain $(X_t)_{t=0}^{\infty}$ $\in \mathrm{Markov}\,(P, \phi)$ with the state space S and let

$$\tau(u, v) = \prod_{i=u}^{v} p_{s_{i-1}|s_i}$$

designate the probability that the chain should generate $s_u \ldots s_v$ after s_{u-1}.

The Markov Case

We denote by $P_\phi(\mathbf{s})$ is the stationary probability of observing $\mathbf{s}$ so that

$$P_\phi(\mathbf{s}) = \phi(s_1) \cdot \prod_{i=2}^{h} p_{s_{i-1}|s_i}.$$

The Markov Case: The recursion

The probability function of the waiting time of the word $\mathbf{s} = s_1 \dots s_h$ $p(l)$ at time $t = l$ in a sample path of $(X_t)_{t=0}^{\infty}$, is for $l > h$

$$p(l) = P_\phi(\mathbf{s}) - L_2 - L_3,$$

where

$$L_2 = \sum_{z=h}^{l-h} p(z) \cdot p_{s_k|s_1}(l - z - h + 1) \cdot \tau(2, h),$$

where $p_{s_k|s_1}(n)$ is the probability of transition from s_k to s_1 in n steps, and

$$L_3 = \sum_{z=l-h+1}^{l-1} p(z) \cdot w_{h-l+z} \cdot \tau(h - l + z + 1, h).$$

Furthermore, $p(l) = 0$ for $l < h$ and $p(h) = P_\phi(\mathbf{s})$.

The Markov Case: The recursion

The result can clearly be derived in the same way as the result by Blom and Thorburn was derived above and is due to

S. Robin and J.J. Daudin (1999): Exact distribution of word occurrences in a random sequence of letters. *Journal of Applied Probability*, 36, pp. 179−193.

The p.g.f. of the waiting time of a word

Gunnar Blom and Daniel Thorburn (1982) derived also the p.g.f. of the waiting time of a word.

The p.g.f. of the waiting time of a word

$$p_{Y_3}(y) = \left(\frac{1}{4}\right)^L - \left(\frac{1}{4}\right)^L \sum_{j=L}^{y-L} p_Y(j) - \sum_{j=y-L+1}^{y-1} p_Y(j) w_{L+j-y} \left(\frac{1}{4}\right)^{y-j}$$

$$\Leftrightarrow$$

$$4^L p_{Y_3}(y) = 1 - \sum_{j=L}^{y-L} p_{Y_3}(j) - 4^L \sum_{j=y-L+1}^{y-1} p_{Y_3}(j) w_{L+j-y} \left(\frac{1}{4}\right)^{y-j}$$

Here

$$1 - \sum_{j=L}^{y-L} p_{Y_3}(j) = \sum_{j=y-L+1}^{\infty} p_{Y_3}(j)$$

and we use this in the right hand side

The p.g.f. of the waiting time of a word

$$\Leftrightarrow$$

$$4^L p_{Y_3}(y) = \sum_{j=y-L+1}^{\infty} p_{Y_3}(j) - 4^L \sum_{j=y-L+1}^{y-1} p_{Y_3}(j) w_{L+j-y} \left(\frac{1}{4}\right)^{y-j}$$

Now we multiply this equality by t^y and sum over y:

$$4^L \mathbf{p}_{Y_3}(t) = \sum_{j=1}^{\infty} \sum_{y=L}^{L+j-1} p_{Y_3}(j) t^y$$

$$-4^L \sum_{y=L}^{\infty} \sum_{j=y-L+1}^{y-1} p_{Y_3}(j) w_{L+j-y} \left(\frac{1}{4^{y-j}}\right) t^y$$

The p.g.f. of the time to the waiting time of a word

$$4^L \mathbf{p}_{Y_3}(t) = \sum_{j=1}^{\infty} \sum_{y=L}^{L+j-1} p_{Y_3}(j) t^y - 4^L \sum_{y=L}^{\infty} \sum_{j=y-L+1}^{y-1} p_{Y_3}(j) w_{L+j-y} \left(\frac{1}{4^{y-j}} \right) t^y.$$

can be rewritten (lots of details omitted) in the form

$$4^L \mathbf{p}_{Y_3}(t) = \frac{t^L}{1-t} \sum_{j=1}^{\infty} p_{Y_3}(j) \left(1 - t^j\right) - 4^L \mathbf{p}_{Y_3}(t) \sum_{s=1}^{L-1} t^s w_{L-s} \left(\frac{1}{4} \right)^s.$$

$$\Leftrightarrow$$

$$4^L \mathbf{p}_{Y_3}(t) = \frac{t^L}{1-t} (1 - \mathbf{p}_{Y_3}(t)) - 4^L \mathbf{p}_{Y_3}(t) \sum_{s=1}^{L-1} t^s w_{L-s} \left(\frac{1}{4} \right)^s.$$

The p.g.f. of the time to the waiting time of a word

$$4^L \mathbf{p}_{Y_3}(t) = \frac{t^L}{1-t}(1 - \mathbf{p}_{Y_3}(t)) - 4^L \mathbf{p}_{Y_3}(t) \sum_{s=1}^{L-1} t^s w_{L-s} \left(\frac{1}{4}\right)^s .$$

If we solve this equation w.r.t. $\mathbf{p}_{Y_3}(t)$ we get $\Rightarrow$

$$\mathbf{p}_{Y_3}(t) = \frac{1}{1 + (1-t)d(4/t)},$$

where $d(x) = \sum_{r=1}^{L} w_r x^r$ (= the generating function of the overlap capabilities).

End of Lecture 6