Análise de uma seqüência de DNA

Statistics for Biology and Health
(Ewens, W.J. & Grant, G.R.)
Capítulo 5

Carla Rêgo Monteiro

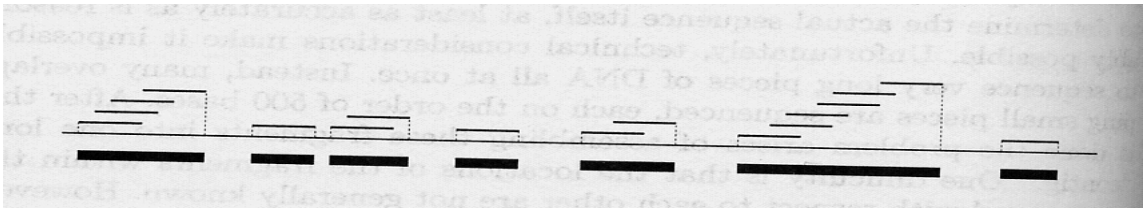
- Construção de uma seqüência
- Modelar DNA
- Verificar Sinais de um DNA
- Long Repeats
- r-Scans
- Análise de Padrões

■ Shotgun Sequencing

→Analisar uma seqüência de DNA :
Construir a seqüência

→ Difícil de construir seqüências longas.

→Construir várias seqüências pequenas.
(algumas sobrepostas)



→ Problema :
Juntar as seqüências que se sobrepõem em “contigs” .
(Shotgun sequencing)

→Cobertura:

Uma cobertura de nX é obtida se para uma seqüência de DNA de tamanho G, o tamanho total de fragmentos sequenciados é nG.

→Genoma humano

Duas estratégias são usadas para seqüênciar o genoma humano.

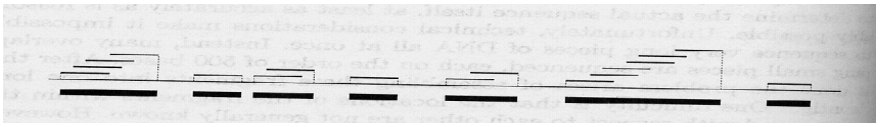
1-Dividir o genoma em peças de tamanho 100,000 com localizações conhecidas.

(shotgun sequencing com 8Xcobertura)

2- Partir o genoma em peças bem pequenas.

(whole shotgun sequencing com alta cobertura)

→Construindo um DNA



→ Questões:

1- Qual é a proporção média do genoma coberto pelos contigs?

2- Qual é o número médio de contigs?

3-Qual o tamanho médio de um contig?

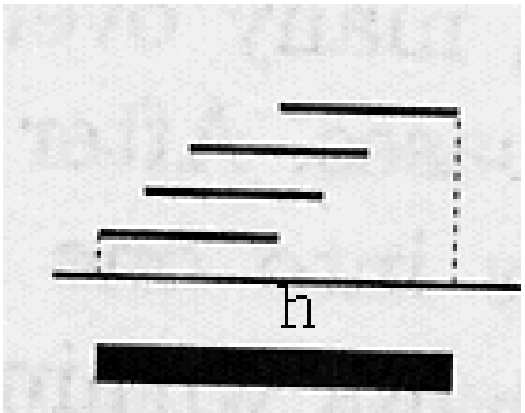
Suponha que temos Nfragmentos de mesmo tamanho L de um genoma de tamanho G. G>>L

Assim a cobertura **a** = NL/G

Os fragmentos são escolhidos aleatoriamente do genoma. Assim a posição do lado esquerdo de uma seqüência tem a mesma probabilidade de ser escolhida em cada intervalo do genoma. P_e ~Unif(0,G)

P(x<P_e<x+h) = $\int_x^{x+h} \frac{1}{G} dp_e = \frac{h}{G}$

Observe que podemos escolher 0, 1, 2,....N fragmentos cuja a posição esquerda esteja dentro do intervalo h.



Defina X como o número de fragmentos com a posição esquerda no intervalo h.

$X \sim Bin(N, h/G)$ $\quad E(X) = \frac{Nh}{G}$

$\rightarrow$ $B(N, h/G)$ $\rightsquigarrow$ $Poi(Nh/G)$

(quando n é grande e h/G pequeno)

Seja L o tamanho do intervalo.

$X \sim Poi(a)$, $a = NL/G$

$P(X>1) = 1 - P(X=0) = 1 - e^{-a}$

1- A probabilidade que um ponto aleatoriamente escolhido seja coberto por no mínimo um fragmento será

$$P(X>1) = 1- e^{-a}$$

Então se queremos uma probabilidade de cobertura de 99% o número médio de fragmentos cobrindo um ponto será

$0.99= 1- e^{-a} \Rightarrow a=4.6 \Rightarrow$ cobertura de $4.6X$

2- Número médio de contigs

Cada contig tem um único fragmento mais a direita. Então o número médio de contigs será igual o número médio de fragmentos mais a direita.

$$N \times P(X=0) = Ne^{-a} = Ne^{-NL/G}$$

P(X=0)= probabilidade que nenhum outro fragmento ter seu ponto final do lado esquerdo neste fragmento.

Quanto maior "**a**" menor o número de contigs.

O número de contigs é maximizado quando a=1, cobertura 1X.

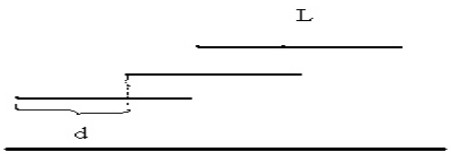
Seja G= 100,000, L=500

a	0.5	0.75	1	1.5	2	3	4	5	6	7
Número médio de contigs	60.7	70.8	73.6	66.9	54.1	29.9	14.7	6.7	3.0	1.3

→ Problema: Alta cobertura é difícil . (junk DNA)

3- Tamanho médio do contig

Seja **d** a distância entre o ponto final do lado esquerdo de um fragmento e o ponto final do lado esquerdo do fragmento seguinte.



D~Geo(N/G) $\rightarrow$ Exp(N/G)

Então o fragmento 2 irá sobrepor o fragmento 1 se D < L

$$P(D<L) = \int_0^L \frac{N}{G} e^{-\frac{N}{G}d} dd = 1 - e^{-NL/G} = 1 - e^{-a}$$

→Um contig é construído até que um fragmento pare de sobrepor o anterior.

Seja Y o número de sucessivos fragmentos sobrepostos até a sobreposição falhar.

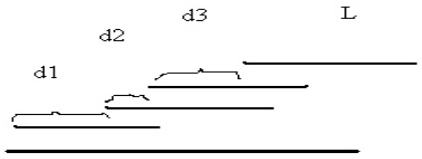
$Y \sim Geo(1- e^{-a})$

O número médio de sucessivos fragmentos sobrepostos num contig.é

$$E(Y) = \frac{1 - e^{-a}}{1 - (1 - e^{-a})} = e^{a} - 1$$

Suponha que temos n fragmentos num contig. Então o tamanho total deste contig será

L(o tamanho do último fragmento) +
(n–1) distâncias



Nós vimos que $D_i \sim Exp(\lambda)$. $\lambda = N/G$ $0<d_i<\infty$

Queremos $0<d_i<L$

$$f_{D_i'}(d_i) = P(x<D_i<x+h \mid 0<d_i<L) = \frac{\lambda e^{-\lambda d_i}}{1-e^{-\lambda L}}$$

Assim o tamanho médio das distâncias é

$$E(D_i^*)=\int_0^L \frac{d_i \lambda e^{-\lambda d_i}}{1-e^{-\lambda L}} \partial d_i = \frac{1}{\lambda} - \frac{L}{e^{\lambda L}-1}$$

Observe que os D_i's são independentes e que nós não sabemos quantos D_i's teremos.

$S= \sum D_i$ é uma variável aleatória

$$E(S)= E(Y)E(D_i^*) = (e^a -1)\left(\frac{1}{\lambda} - \frac{L}{e^{\lambda L} -1}\right) = \frac{e^a -1}{\lambda} - L$$

Então o tamanho médio de um contig será

$$L + \frac{e^a - 1}{\lambda} - L = \frac{e^a - 1}{\lambda} = L\frac{e^a - 1}{a}$$

Exemplo: L=500, a=2 e G=100,000 $\Rightarrow$ N=aG/L=400

1- A proporção média do genoma coberto pelos contigs é

$$1- e^{-2} = 86.5\%$$

2- O número médio de contigs é

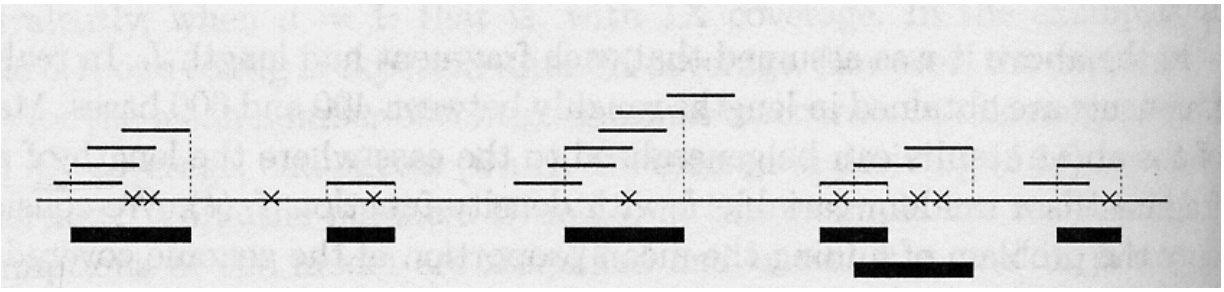
$$400e^{-2} = 55$$

3- O tamanho médio de um contig é

$$500(e^{2}-1)/a \approx 1598 \text{ ncleotídeos}$$

$\rightarrow L \sim f_L(l)$ $\quad E(L) \quad \Rightarrow a = \frac{NE(L)}{G}$

■Contigs Ancorados



Existem pequenas seqüências no DNA que são únicas e cuja localização são conhecidas e providem marcadores.

Uma âncora é um ponto no genoma que localiza qualquer fragmento cobrindo ele.

Um fragmento ancorado é qualquer fragmento que contém pelo menos uma âncora.

Um contig ancorado é uma coleção de fragmentos colocados juntos por âncoras.

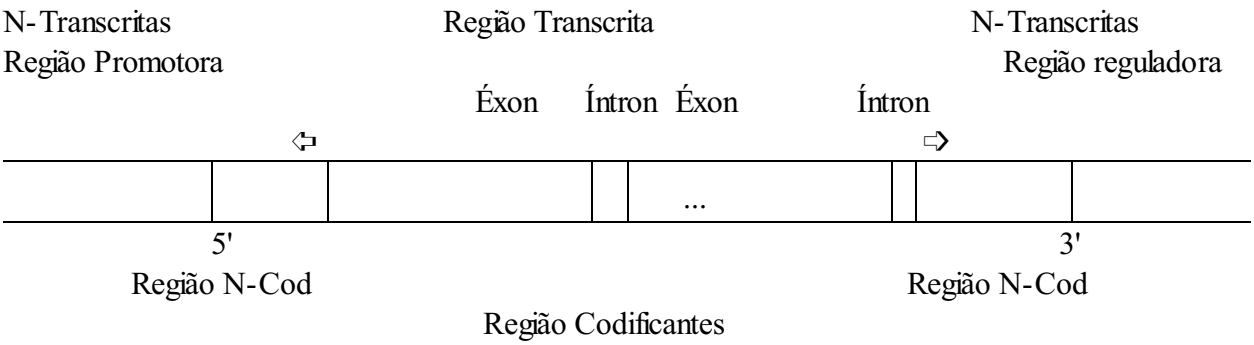
→ Questões:

1- Qual é a proporção média do genoma coberto pelos contigs ancorados?

2- Qual é o número médio de contigs ancorados?

3-Qual o tamanho médio de um contig ancorado?

▪ Modelando o DNA



Íntrons - Regiões não codificantes

→Íntrons e Éxons tem diferentes propriedades estatísticas.

→Modelos são ajustados com base nestas propiedades.

→Testes são construídos para verificar se um fraguimento de DNA é parte de uma região codificante de um gene.

→ Training data são usados para ajustar estes modelos.

→Suposições: Nucleotídeos em várias posições são considerados independentes e identicamente distribuídos. (modelo mais simples)

→ Para observar diferenças entre regiões codificantes e não codificantes, observa-se a diferença entre freqüências dos 4 nucleotídeos A,G,C,T em duas regiões (íntron vs éxon)

→ Suposição de independência é satisfeita?

→ Teste - Markov Chains

H_0: nucleotídeos de um site independe dos nucléotídeos do site precedente.

H_1: um determinado nucleotídeo de um site depende dos nucleotídeos do site precedente.

Seja Yij o número de vezes, em uma seqüência de DNA de interesse, que um nucleotídeo do tipo i é seguido pelo nucleotídeo j. i,j = 1,...,4;
1=a, 2=g, 3=c, 4=t

		Nucleotídeo j no site k+1				
		a	g	c	t	Total
Nucleotídeo i no site k	a	Y_{11}	Y_{12}	Y_{13}	Y_{14}	$Y_{1.}$
	g	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{2}
	c	Y_{31}	Y_{32}	Y_{33}	Y_{34}	Y_{3}
	t	Y_{41}	Y_{42}	Y_{43}	Y_{44}	Y_{4}
Total		$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	$Y_{.4}$	Y

$$X^2 = \sum_{i=1}^{4}\sum_{j=1}^{4} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_9 \qquad \text{sob } H_0$$

$$E_{ij} = \frac{Y_{i.}Y_{.j}}{Y}$$

→Modelos de Markov de primeira ordem ajustam-se melhor aos dados do que modelos assumindo independência.

→Modelos de Markov de ordem maiores são considerados ainda melhor.

■ Modelando Sináis no DNA

Um Sinal é uma pequena seqüência de DNA.

Genes contém sinais no DNA para indicar os começo e o fim de regiões codificadoras.
(limite entre éxon e íntron)

Um mesmo sinal pode aparecer em várias seqüências de DNA.

Membros de um sinal: Seqüências que contém o mesmo sinal.

→Questão:

São todos os membros de um sinal conhecidos?
Resp: Não.

→Solução:

Usar membros conhecidos de um sinal para acessar a probabilidade de que uma determinada seqüência de DNA seja membro daquele sinal.

Exemplo:

Para investigar sinais de um DNA humano pode se usar um conjunto de membros de sinais obtidos de um banco de dados de genes humanos.

→Alguns sinais são mais complexos do que outros
(Modelos mais complexos)

■ Matrizes Ponderadas: Independência

Testes para verificar se uma determinada seqüência de DNA é membro de um sinal são mais simples quando temos garantida a seguinte suposição:

Um nucleotídeo de qualquer posição no sinal é independente do nucleotídeo de uma outra posição no sinal.

Verificando independência:

H_0: O nucleotídeo de qualquer posição no sinal é independente do nucleotídeo de uma outra posição no sinal.

H_1: Pelo menos um nucleotídeo de uma posição no sinal não é independente do nucleotídeo de uma outra posição no sinal.

Seja Yij o número de vezes, em um sinal que um nucleotídeo do tipo i esta na posição a e um nucleotídeo j esta na posição b. i,j = 1,...,4 ; 1=a, 2=g, 3=c, 4=t

a,b = 1,...,n n é o tamanho do sinal.

Teremos $\binom{n}{2}$ tabelas comparando pares de posições a, b.

Tabela 2

		Nucleotídeo j na posição b				
		a	g	c	t	Total
Nucleotídeo i na posição b	a	Y_{11}	Y_{12}	Y_{13}	Y_{14}	$Y_{1.}$
	g	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{2}
	c	Y_{31}	Y_{32}	Y_{33}	Y_{34}	Y_{3}
	t	Y_{41}	Y_{42}	Y_{43}	Y_{44}	Y_{4}
Total		$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	$Y_{.4}$	Y

Exemplo:

Se o sinal é de tamanho 5, teremos 10 tabelas.

→Problema: Múltiplos Testes.

Seja 95% o nível de confiança escolhido.

$(.95)^{10} = 0.599 \quad \rightarrow \quad \alpha = 1 - .599 = 0.40$

→ por chance estaremos rejeitando mais

→Alternativa:

Testes de Bonferroni : α/g ;
g = número de testes.

Uma vez que não rejeitamos a hipótese de **independência**

Matriz Ponderada M para um sinal de tamanho n

		Posição				
		1	2	3	...	n
Nucleotídeo	a	p_{11}	p_{12}	p_{13}	...	p_{1n}
	g	p_{21}	p_{22}	p_{23}	...	p_{2n}
	c	p_{31}	p_{32}	p_{33}	...	p_{3n}
	t	p_{41}	p_{42}	p_{43}	...	p_{4n}
Total		1	1	1		1

pij = proporção de casos, no “training data” , que o nucleotídeo i aparece na posição j do sinal

Matriz Ponderada M para um sinal de tamanho 5

		Posição				
		1	2	3	4	5
Nucleotídeo	a	0.33	0.34	0.19	0.20	0.21
	g	0.22	0.27	0.23	0.24	0.21
	c	0.31	0.18	0.34	0.30	0.25
	t	0.14	0.21	0.24	0.26	0.33
Total		1	1	1	1	1

Exemplo : Sinal atata

$$P(s|M) = .33x.21x.19x.26x.21 = 0.00072$$

❑ Se rejeitarmos a hipótse de independência?

→Dependências de primeira ordem - Markov

Distribuição de probabilidade das posições.

→ primeira posição

a	g	c	t
p_a	p_g	p_c	p_t

p_i =proporção de casos, no “training data” , que o nucleotídeo i aparece na posição 1 do sinal

→ segunda a n-ésima posição :

A distribuição de probabilidade do nucleotídeo i na posição k depende do nucleotídeo na posição k-1.
(k=2,...n)

Matrizes de probababilidade de transição
Markov Chain

K-ésima posição

$$\begin{vmatrix} p_{aa} & p_{ag} & p_{ac} & p_{at} \\ p_{ga} & p_{gg} & p_{gc} & p_{gt} \\ p_{ca} & p_{cg} & p_{cc} & p_{ct} \\ p_{ta} & p_{tg} & p_{tc} & p_{tt} \end{vmatrix}$$

p_{ij} = probabilidade que o nucleotídeo i ocorra na posição k dado que o nucleotídeo j estava na posição k-1. (Estimadas pelo training data)

- Máxima dependência

→Dependencias de ordes maiores : matrizes de tranzição ainda maiores. (Limite de dados)

→ Dependências mais informativas

→ Decomposição de máxima dependência- MDD- (Burge-1997)

→Achar a posição que tem mais influência nas outras posições

→Construir uma tabela nxn cuja posição (i,j) é o valor observado da estatística X^2 obtida em tabelas como a Tabela 2.

Tabela de X^2

	1	2	3	4	5	total
1		34.2*	7.1	37.2*	2.8	81.3
2	34.2*		.4	72.4*	4.5	111.5
3	7.1	.4		15.3	98.3*	121.1
4	37.2*	72.4*	15.3		14.2	139.1 ←
5	2.8	4.5	98.3*	14.2		119.8

→A posição 4 é a que apresenta maior influência.

Contruir um Modelo que determina as distribuições das posicões 1,2,3 e 5 condicional a posição 4.

→Dividir os membros de um sinal no training data em 4 subconjuntos, T_a, T_g, t_c e T_t

T_x = todas as seqüências com o nucleotídeo x na posição 4. x = a, g, c,t

n_x = números de seqüências em $T_{x.}$

$n = n_a + n_{g+} n_{c+} n_t$ = total de membros no training data.

$p_x = n_x/n$ = probabilidade do nucleotídeo x estar na posição 4.

→ Para cada subconjunto T_x calcular a matriz ponderada M_x .

→ Assim, o modelo M consiste das distribuições $\{p_a, p_g, p_c, p_t\}$ com 4 matrizes ponderadas $\{ M_a, M_g, M_c, M_t\}$

→ Probabilidade de uma sequencia **s** ser membro deste sinal é

$$P(s| M)$$

→ Se o nucleotídeo x ocorre na posição 4 de s a matrix ponderada M_x é usada.

Assim $P(s|M) = p_x.p_1.p_2.p_3.p_5,$

p_k = é a proporção de casos, no subconjunto T_x, que o nucleotídeo i aparece na posição k.
(obtido de M_x)

Exemplo : Sinal **atata** - posição 4 = t , p_t = .40

Matriz Ponderada M_t para um sinal de tamanho 5

		Posição				
		1	2	3	4	5
Nucleotídeo	a	0.33	0.34	0.19	0.00	0.21
	g	0.22	0.27	0.23	0.00	0.21
	c	0.31	0.18	0.34	0.00	0.25
	t	0.14	0.21	0.24	1.00	0.33
Total		1	1	1	1	1

$$P(s|M) = p_t.p_1.p_2.p_3.p_5 =$$
$$0.40x0.33x0.21x.19x0.21=$$
$$0.001106$$

Este processo pode ser repitido recursivamente.

Considere agora o subconjunto T_x

Ache a posição mais influente entre as posições restantes , 1,2,3,5(tabela X^2).

Considere a posição 2

Divida T_x em 4 subgrupos, T_{xa}, T_{xg}, T_{xc} e T_{xt}

T_{xy} = todas as seqüências com o nucleotídeo x na posição 4 e o nucleotídeo y na posição 2.

y = a, g, c,t.

n_{xy} = números de seqüências em $T_{xy.}$

$n_x = n_{xa} + n_{xg} + n_{xc} + n_{xt}$ - total de membros no subconjunto T_x.

$p_{xy} = n_{xy}/n_x$ = probabilidade do nucleotídeo y estar na posição 2 e nucleotídeo x esta na posição 4.

→ Para subconjunto conjunto T_{xy} calcular a matriz ponderada M_{xy} .

Assim $P(s|M) = p_x.p_{xy}.p_1.p_3.p_5.$ p_k obtida da matriz M_{xy}

O processo é repetido enquanto se tem pelo menos 100 membros em cada subgrupo. $T_{xyzw...}>100$

■ **Long Repeats**

→ Interesse em um determinado nucleotídeos , a.

→Questão: Há significativa evidência de longas repetições deste nucleotídeo?

→Teste

H_0: As sequências do nucleotídeo **a** ocorre de forma aleatória na seqüência.

H_1: Há tendência a longas repetições do nucleotídeo **a**

→Considere uma seqüência de DNA de tamanho N.

Suponha que a probabilidade que o nucleotídeo **a** ocorra em qualquer site é um valor conhecido p.

→Sucesso - se ocorre a no site i (p)
→Fracasso- se não ocorre a no site i (1-p)

→Seja Y = o número de sucessos, **a**, até aparecer a primeira falha, **c,g,t**.

Y é o tamanho da repetição de **a**.

$Y \sim Geo(p)$

$P(Y = y) = p^y(1-p)$ $\quad y = 0,1,2...$

→ Estatística do Teste

Suponha agora que temos n sequências,(s,s,...,f), independentes observadas.

Seja Y_{max} = o tamanho da maior seqüência, a estatística do teste.

y = o tamanho observado da maior seqüência

Então p-valor = $P(Y_{max} \geq y) = 1- (1-p^y)^n$ sob a hipótese H_0

→ Quem é n?

aaaac aat g g aac aaat

O número de seqüências é igual ao número de falhas ocorridas.

N(1-p) é o número experado de n.

p-valor = $1 - (1-p^y)^{N(1-p)}$

Uma aproximação deste p-valor pode ser encontrada usando exponencial aproximação quando N é grande e $(1-p)Np^y \leq 1$e n=N(1-p). Assim

p-valor* $\approx 1 - e^{-N(1-p)p^y}$

Conecção com BLAST

Exemplo: Seja N=100,000, p = 0.25 e y=10

n = 100,000x0.75 = 75,000

p-valor = 0.0690272 não rejeita H_0 ao nível de 0.05

p-valor* = 0.0690275

As seqüências de **a** aparencem de forma aleatória.

→O número de falhas numa seqüência, n, é aleatório.

n~ Bin(N, 1-p) , E(n) = N(1-p)

p-valor ≈ 1- (1 - (1-p)p^{y})^{N} = 0.0690276

O p-valor* ainda aproxima bem.

Assumindo a aproximação exponencial

→Distribuição de Y_{max} :

$$\mu_{\max} = \frac{\gamma + \log n}{-\log p} - \frac{1}{2}$$

$$\sigma^2_{\max} = \frac{\pi^2}{6(\log p)^2} + \frac{1}{12}$$

γ e π são funções de p.

→Questão:

Há significativa evidência de longas repetições em **qualquer** nucleotídeo?

Distribuição de Y_{max} :

$$\mu_{\max} = \frac{0.6359 + 2\log n + \log(1-p)}{-\log(P)} - 1$$

$$\sigma^2_{\max} = \frac{\pi^2}{6(\log P)^2}$$

P - soma dos quadrados das proporções dos 4 nucleotídeos.

■ **r-Scans**

→Palavra : gaga

→Observar no fragmento L se esta palavra ocorre n vezes. Palava <<<L

→Questão:

Estas palavras ocorrem aleatoriamente no fragmento de tamanho L?

→Testes : Karlin & Macken(1991)

H_0: Os pontos onde as palavras aparecem são iid variáveis aleatórias Unif(0,L)

H_{11}: Os pontos ocorrem em uma overly dispersed fashion ou

H_{12}: Os pontos ocorrem in a clumped fashion

❑ Testes H_0 vs H_{11}

Sejam U_1, ...U_{n+1} subinterlavos em (0,1) definidos por n pontos no qual a palavra de interesse ocorre.

Estatística do teste :

U_{max}

$$E(U_{max}) = \frac{1}{n+1}\left(\frac{1}{n+1} + \frac{1}{n} + \ldots + \frac{1}{1}\right)$$

$$Var(U_{max}) = \frac{\pi^2}{6(n+1)^2}$$

$n \to \infty$

p-valor = $$P(U_{max} \geq u) = 1 - e^{-(n+1)e^{-(n+1)u}}$$

$\Rightarrow$ $$P(U_{max} \geq \frac{\log(n+1) + u}{n+1}) = 1 - e^{-e^{-u}}$$

Estatística do teste - rscans

$$R_1(r)=U_1+ U_2 + \ldots+U_r$$
$$R_2(r)= \quad U_2+ U_3 + \ldots+U_{r+1}$$
$$\vdots$$
$$R_{n-r+1}(r)= \quad U_{n-r+1} + \ldots+ U_n$$

$R_{max}(r)$= a estatística do teste

$$P(R_{\max} \geq \frac{\log(n+1)+(r+1)\log(\log(n+1))+u}{n+1}) \approx$$

$$1- e^{-e^{-\frac{u}{(r-1)!}}}$$

❑ Testes H_0 vs H_{12}

Sejam U_1, ...U_{n+1} subinterlavos em (0,1) definidos por n pontos no qual a palavra de interesse ocorre.

Estatística do teste :

U_{min}

p-valor = $P(U_{\min} \leq u) = 1 - e^{-u(n+1)^2}$, $n \to \infty$

Estatística do teste - rscans

$$R_1(r)=U_1+ U_2 + \ldots+U_r$$
$$R_2(r)= \quad U_2+ U_3 + \ldots+U_{r+1}$$
$$\vdots$$
$$R_{n-r+1}(r)= \quad U_{n-r+1} + \ldots+ U_n$$

$R_{min}(r)$= a estatística do teste

$$P(R_{\min} \leq u) \approx 1 - e^{-u^r (n+1)^{r+1}/r!}$$

■Análise de Padrão

Seja uma sequência de nucleotídeos de tamanho N iid.

→Interesse : gaga

Questões:

1- Qual é o número médio de vezes que a palavra gaga aparece na seqüência?

2-Qual é o tamanho médio entre duas ocorrências desta palavra?

Razões

1-Suspeita que a palavra gaga ocorre na seqüência mais freqüênte que ela deveria se seguisse um padrão iid.

→ As freqüências gggg, ggga,gaag são diferêntes da freqüência de gaga - iid.

2-Descobrir sinais promotores comuns a região upstream do gene através de padrões encontrados no DNA.

→Criar dicionários de palavras, de diferentes tamanhos, com suas respectivas probabilidades.

→Qualquer palavra que ocorrer mais freqüênte que o esperado na região upstream é candidato a ser um sinal promotor.

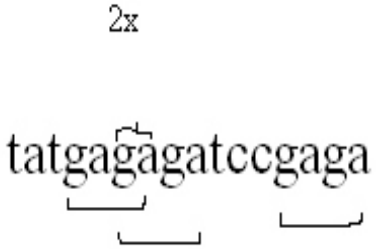
■Overlaps Counted

Questão:

1-Qual o número médio de vezes que a palavra **gaga** aparece na seqüência?

(Sobreposições)

2-Qual a distancia média entre ocorrências da palavra.



❑ Número de Ocorrências(Sobreposições)

H_0: A palavra gaga ocorre aleatoriamente

$Y_1(N)$ = número de vezes que a palavra gaga ocorre na seqüência de DNA de tamanho N.

Suposição: Os nucleotídeos ocorrem independentes em cada site com P(a)=P(g)=P(c)=P(t)=0.25

$$I_j = \begin{cases} 1 & gaga \quad termina \quad na \quad posicao \quad j \\ 0 & outro \quad caso \end{cases}$$

$P(I_j = 1)$ = P(**g** estar na posição j-3, **a** estar na posição j-2, **g** estar na posição j-1, **a** estar na posição j) = $(0.25)^4 = 1/256$

$Y_1(N) = I_4 + I_5 + ... + I_N \quad \rightarrow Y_1(N) \not\approx \text{Bin}(N-3, 1/256)$

A distribuição de Y(N) depende da palavra.

gagag gagaga $N \geq 6$

$E(Y_1(N)) = (N-3) / 256$

$Var(Y_1(N)) = (281N - 895) / 65536$

Estatística do Teste : (N grande)

$$Z = \frac{Y_1(N) - E(Y_1(N))}{\sqrt{Var(Y_1(N))}} \sim N(01)$$

p-valor = $P(|Z| \geq z)$

Exemplo: N=100,000 , y(N)=4023

→$E(Y_1(N)) = (N–3) / 256 = 390624$

$Var(Y_1(N)) = (281N- 895) / 65536 = 4287.71$

P-valor = 0.038 rejeita H_0 ao nível 0.05

→ Variância para qualquer palavra

Seja w_j = 1 se as primeiras j letras da palavra de interesse são iguais e na mesma ordem que as últimas letras.

= 0 caso contrário

$$\text{Var}(Y_1(N)) = \frac{(249 + 128w_3 + 32w_2 + 8w_1)N - (775 + 512w_3 + 160w_2 + 48w_1)}{65536}$$

❑ Distância entre as ocorrências

→ $Y_2(N)$ = a distância até a próxima ocorrência da palavra após o site i. ≠ geométrica

$E(Y_2(N)) = 256$

$Var(Y_2(N)) = 512\sum_{j=1}^{4} w_j 4^j - 67328$ depende da palavra

→ $Y_3(N)$ = número de sites até a primeira ocorrência da palavra de interesse. começando na origem)

$$E(Y_3(N)) = \mu_3 = \sum_{j=1}^{4} w_j 4^j$$

$$Var(Y_2(N)) = \mu_3^2 + \mu_3 - 2\sum_{j=1}^{4} j w_j 4^j$$

Exemplo : N=1,000,000

Palavra	Experânça			Variância		
	$Y_1(N)$	$Y_2(N)$	$Y_3(N)$	$Y_1(N)$	$Y_2(N)$	$Y_3(N)$
gaga	3906.24	256	272	4284	71936	72144
gggg	3906.24	256	340	6363	106752	113436
gaag	3906.24	256	260	3292	65792	65804
gagc *	3906.24	256	256	3799	63744	63744

$Y_2(N)$ = a distância até a próxima ocorrência da palavra após o site i.

$Y_3(N)$ = número de sites até a primeira ocorrência da palavra de interesse,

$Y_1(N)$ = número de vezes que a palavra gaga ocorre na seqüência de DNA de tamanho N.

■ **Repetições sem contar as sobreposições**

Reocorrências das palavras

Seja R_1 a primeira ocorrência da palavra - primeira reocorrência;

R_2 a segunda ocorrência da palavra que não se sobrepõe a 1º;

R_3 a terceira ocorrência da palavra que não se sobrepõe a 2º;

⋮

Razão: Como as seqüências são cortadas.

❑ Distância entre as ocorrências

Seja $Y_4(N)$ o número médio de sites entre consecutivas reocorrências (distância).

$$E(Y_4(N)) = \mu = \sum_{j=1}^{4} w_j 4^j$$

$$Var(Y_4(N)) = \sigma^2 = \mu_3^2 + \mu_3 - 2\sum_{j=1}^{4} j w_j 4^j$$

Mesma distribuição de $Y_3(N)$

❑ Número de reocorrências(sem sobreposições)

Seja $Y_5(N)$ o número de reocorrências em uma seqüência de tamanho N.

$$E(Y_5(N)\ ) \approx N / \mu$$

$$Var(Y_5(N)\ ) \approx \frac{N\sigma^2}{\mu^3}$$

Exemplo : N=1,000,000

Palavra	Experança				Variância			
	Ocorrências		Distâncias		Ocorrências		Distâncias	
	$Y_1(N)$	$Y_5(N)$	$Y_3(N)$	$Y_4(N)$	$Y_1(N)$	$Y_5(N)$	$Y_3(N)$	$Y_4(N)$
gaga	3906.24	3676.47	272	272	4284	3585	72144	72144
gggg	3906.24	2941.17	340	340	6363	2886	113436	113436
gaag	3906.24	3846.15	260	260	3292	3744	65804	65804
gagc *	3906.24	3906.25	256	256	3799	3799	63744	63744

$Y_1(N)$ = número de vezes que a palavra gaga ocorre na seqüência de DNA de tamanho N.

$Y_5(N)$ = número de reocorrências em uma seqüência de tamanho N.

$Y_3(N)$ = número de sites até a primeira ocorrência da palavra de interesse,

$Y_4(N)$ = número médio de sites entre consecutivas reocorrências.