Compressão de Sequências de DNA

Lauro Didier Lins

16 de Novembro de 2004

A referência básica desta aula é o capítulo 7, *Compressing DNA Sequences*, do livro *Current Topics in Computational Biology* (2002). O capítulo foi escrito pelo professor Ming Li que atualmente trabalha na Universidade of Waterloo no Canadá.

O que é a compressão de uma Seqüência de DNA?

Exemplo:

Seqüência do DNA humano de referência L78833. Tamanho 126826 bp. (fonte: GenBank)

```
>gi|1698398|gb|L78833.1|HUMBRCA1 Human BRCA1, Rho7
 and vatI genes, complete cds, and ipf35 gene,
 partial cds
ACGGGGTCTCGAAAAAAGGAGAATGGGATGAGAAG
GATATATGGGTAGTGTCATTTTTTAACTTGCAGAT
TTCATCCTAGTCTTCCAGTTATCGTTTCCTAGCAC
TCCATGTTCCCAAGATAGTGTCACCACCCCAAGGA...
```

Compressão utilizando o programa *GenCompress*

```
C:\users\lauro\ufpe\phd\cursos\compbio>gencompress
L78833-HUMBRCA1.acgt
 Unconditionally compress L78833-HUMBRCA1.acgt.
 Searching for approximate repeats!
 The compressed filename is L78833-HUMBRCA1.GEN!
.....................................
```

```
.....................................
.....................................

 The size of original file is 126826 bytes.
 The size of compressed file is 29137 bytes.
 The compression ration is 77.026004%.
    (defined by 1-|compressed_file|/|original_file|)

 Note: To verify the correctness of compression,
you need follow the next two steps and then see what happens.
       1>  gendecompress  original_file.gen  [-c reference_file]
       2>  comparetwofile  original_file  original_file.out
```

O número de bits por base de nucleotídeo após a compressão é dado por:

$$\frac{8 \times 29137}{126826} \approx 1.83791$$

Aplicações da Compressão de Seqüências de DNA

- armazenar eficientemente as seqüências de DNA (aplicação óbvia).

- ferramenta para medir a "quantidade de informação" de uma seqüencia de DNA. A partir dessa medida é possível, como veremos adiante, definir uma "distância" entre seqüências. Ou seja, a partir da compressão podemos comparar pequenas seqüências de DNA ou até genomas inteiros.

Programas para a Compressão de DNA:

1. Ziv e Lempel.
2. Biocompress & Biocompress-2.
3. *Cfact*.
4. GenCompress.

Ziv e Lempel

- Ziv, J. e Lempel, A. propuseram este método em 1977.
- algoritmo de propósito geral para compressão de dados.
- base para os bem conhecidos arquivos .ZIP
- A idéia básica é encontrar repetições exatas anteriores na sequência e codificá-las através de um *link* para a primeira ocorrência.
- Apenas uma passagem pela entrada.

Biocompress & Biocompress-2

- Grumbach, S. e Tahi, F. no mesmo espírito de Ziv e Lempel propuseram um método de compressão específico para seqüências de DNA.

- a idéia essencial é encontrar repetições exatas anteriores e palíndromos complementares ocorridos anteriormente na seqüência e codificá-los através de um *link* para a primeira ocorrência.

- Apenas uma passagem pela entrada.

- O Biocompress-2 utiliza, nas regiões sem casamentos, *order-2 arithmetic coding*.

Cfact

- Rivals, É., Delahaye, J-P. e Delgrange, O., em 1995, definiram um novo método para compressão de DNA.
- Procura pelo casamento exato mais longo na seqüência inteira. Para isso utiliza uma árvore de sufixo.
- Duas passagens pela entrada. A primeira passagem é para criar a árvore de sufixo.
- As regiões sem repetição são codificadas com 2 bits por base.

GenCompress

- Chen, X., Kwong, S. e Li, M., em 2000, propuseram outro algoritmo específico para compressão de DNA.
- A idéia essencial é a de casamento aproximado.
- utiliza *order-2 arithmetic coding* quando é vantajoso.
- É feito em uma só passagem.
- obtém resultados significativamente melhores do que os anteriores.
- utilizado como ferramenta para comparação de genomas.

Comparação: Biocompress × GenCompress

Sequência	Tam.(bp)	Com-ress	Arith-2	Bio-comp.2	Gen-Comp.1	Gen-Comp.2
MTPACGA	100314	2.116	1.873	1.875	1.861	1.861
MPOMTCG	186608	2.202	1.966	1.938	1.898	1.898
CHNTXX	155844	2.187	1.934	1.617	1.614	1.614
CHMPXX	121024	2.075	1.837	1.685	1.669	1.670
HUMGHCSA	66495	2.194	1.938	1.307	1.092	1.097
HUMHBB	73323	2.195	1.918	1.877	1.813	1.814
HUMHDABCD	58864	2.230	1.943	1.877	1.800	1.809
HUMHDYSTROP	38770	2.233	1.924	1.926	1.924	1.923
HUMHPRTB	56737	2.202	1.929	1.907	1.826	1.830
VACCG	191737	2.167	1.898	1.761	1.761	1.761
HEHCMVCG	229354	2.213	1.965	1.848	1.847	1.847

Valores em bits por base de nucleotídeo.

Comparação: Cfact × GenCompress

Sequência	Tam.(bp)	LZW 15	Arith-2	*Cfact*	GenComp.1	GenComp.2
atatsgs	9647	2.237	1.951	1.785	1.664	1.673
atefla23	6022	2.297	1.994	1.585	1.541	1.540
atrdnaf	10014	2.300	2.009	1.814	1.789	1.786
atrdnai	5287	2.239	1.994	1.468	1.419	1.410
hsg6pdgen	52173	2.168	1.937	1.928	1.785	1.800
xlxfg512	19338	2.084	1.923	1.490	1.376	1.385
mmzp3g	10833	2.244	1.953	1.911	1.854	1.857
celk07e12	58949	2.108	1.912	1.713	1.597	1.605

Valores em bits por base de nucleotídeo.

Teoria da informação

> *Information theory is a branch of the mathematical theory of probability and mathematical statistics, that quantifies the concept of information. It is concerned with information entropy, communication systems, data transmission and rate distortion theory, cryptography, data compression, error correction, and related topics.* (fonte: http://en.wikipedia.org/wiki/Information_theory)

- Informação e Incerteza (Entropia).

$$H = -\sum_{s \in S} p_s \log_2 p_s$$

Estimadores da Entropia de uma Sequencia de DNA

- Farach, M., Noordweider, M., Savari, S. Shepp, L., Wyner, A. e Ziv, A. (1994) propuseram uma novo método de estimar a entropia do DNA chamado de *match length entropy estimator*. Este método foi usado para comparar as diferenças entre exons e introns e, ao contrário do que era esperado, eles acharam que a entropia dos exons era 73% das vezes maior do que a dos introns e que a variabilidade dos introns era 80% das vezes maior do que a dos exons.

- Lowenstern, D. e Yianilos, P. (1999) criaram o programa CDNA para estimar a entropia de uma seqüência de DNA. A observação básica utilizada por eles foi a de que seqüências de DNA contém muito mais repetições aproximadas (*near repeats*) do que seria esperado normalmente. O CDNA utiliza dois parâmetros para capturar as repetições aproximadas: w o comprimento da substrig e h a distância de hamming.

- Lanclot, K., Li, M., e Yang, E.H. (2000) baseados na idéia de Kieffer e Yang (1999) sobre códigos baseados em gramáticas que reconhecem repetições e reversões complementares de seqüências de DNA desenvolveram o programa para estimação da entropia de seqüências de

DNA chamado GTAC (*Grammar Transform Analysis and Compression*)

Algoritmo	Universal	Linear Runtime	Entropy estimate
UNIX compress	sim	sim	pior
Match length	limitado	sim	-
Biocompress-2	sim	sim	terceiro melhor
CDNA	não	não	segundo melhor
GTAC	sim	sim	melhor

Comparação: CDNA $\times$ GTAC

Seq	Tam.(bp)	Compress	Biocomp-2	CDNA	GTAC
PANMTPACGA	100314	2.12	1.88	1.85	1.74
MPOMTCG	186608	2.20	1.94	1.87	1.78
CHNTXX	155844	2.19	1.62	1.65	1.53
CHMPXX	121124	2.08	1.68	-	1.58
SCCHRIII	315339	2.18	1.92	1.94	1.82
HUMGHCSA	66495	2.19	1.31	0.95	1.10
HUMHBB	73323	2.20	1.88	1.77	1.73
HUMHDABCD	58864	2.21	1.88	1.67	1.70
HUMHDYSTROP	38770	2.23	1.93	1.93	1.81
HUMHPRTB	56737	2.20	1.91	1.72	1.72
VACCG	191737	2.14	1.76	1.81	1.67
HEHCMVCG	229354	2.20	1.85	-	1.74

Compress e Biocompress-2 são algoritmos de compressão sem perda. CDNA e GTAC são estimadores de entropia.

Experimento com GTAC

- Aproximadamente 15% do genoma da *E. Coli* (procarionte) é não-codificante.

- Se regiões não-codificantes têm um papel definido, o fato de elas serem mais regulares do que as regiões codificantes fortaleceria a conjectura de que as regiões não-codificantes nos procariontes não são lixo. A aplicação do GTAC na *E.Coli* confirmou esta hipótese:
  - 1.85 bits/base para as regiões codificantes (4.090.525 pares de bases)
  - 1.80 bits/base para as regiões não-codificantes (640.039 pares de bases)

Uma Medida de Distância Entre Seqüências de DNA

Dadas duas seqüências x e y, de DNA ou não, Chen et al.(2000) e Li et al.(2000) definiram a seguinte medida de distância entre x e y:

$$d(x,y) = 1 - \frac{K(x) - K(x\,|\,y)}{K(xy)},$$

onde $K(x)$ é a complexidade de Kolmogorov da seqüência x, ou seja, o tamanho em bits do menor programa que faz um computador universal produzir x como sua única saída. $K(x\,|\,y)$ é a complexidade condicional de Kolmogorv assumindo que o programa tem a seqüência y como informação livre. Um teorema de Kolmogorov prova que $K(x) - K(x\,|\,y) = K(y) - K(y\,|\,x)$ a menos de um erro logaritmico.

Comparando Genomas completos

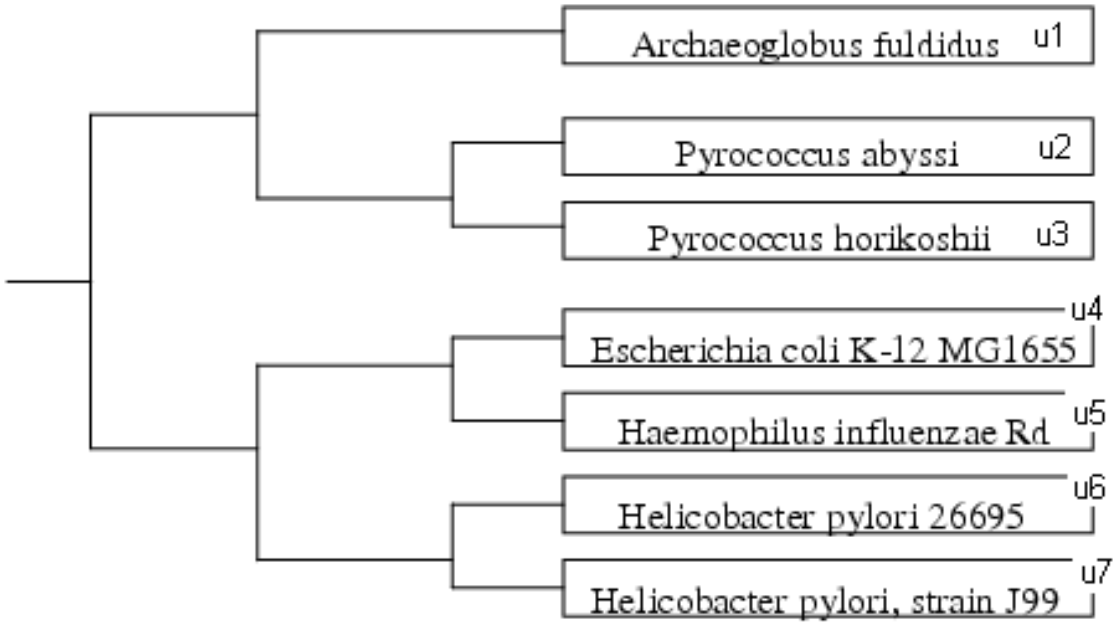
- Neste experimento realizado em Li et al.(2000) calculou-se a distância dois a dois dos seguintes genomas completos:
  - Archaea Bacteria: *Archaeoglobus fulgifus* ($u1$), *Pyrococcus abyssi* ($u2$), *Pyrococcus horikoshii* OT3 ($u3$).
  - Batcetria: *Escherichia coli* ($u4$), *Haemophilus influenze Rd* ($u5$), *Helicobacter pylori* 26695 ($u6$), *Helicobacter pylori*, strain J99 ($u7$).

- A distância foi calculada usando a função $d(x, y)$ trocando a complexidade de Kolmogorov pelo algoritmo de compressão de DNA GenCompress. A complexidade condicional de Kolmogorov também foi trocada por uma versão condicionada do GenCompress.

- A tabela a seguir apresenta os valores $100(1 - d(x, y))$ para cada par x e y.

Seq.	$u1$	$u2$	$u3$	$u4$	$u5$	$u6$	$u7$
$u1$		0.018326	0.019550	-0.000548	-0.002399	-0.001765	-0.002259
$u2$	0.023072		0.797546	0.000089	-0.000988	0.000812	0.000705
$u3$	0.023055	0.794383		-0.000391	0.000617	0.000109	-0.000109
$u4$	0.000373	-0.001084	-0.000537		0.048760	0.008160	0.008371
$u5$	0.000274	0.000145	-0.000044	0.049059		0.018303	0.017776
$u6$	-0.002217	0.000307	-0.000140	0.009068	0.016523		43.069863
$u7$	-0.001314	-0.000782	-0.000062	0.009796	0.019680	43.171044	

- A partir da matriz de distâncias anterior foi possível reconstruir a seguinte filogenia exatamente como informada no GenBank:


Archaeoglobus fulgidus u1
Pyrococcus abyssi u2
Pyrococcus horikoshii u3
u4
Escherichia coli K-12 MG1655
u5
Haemophilus influenzae Rd
u6
Helicobacter pylori 26695
u7
Helicobacter pylori, strain J99