Comunicação
III Simpósio Doutoral do Departamento de Informática
da Universidade do Minho

Sérgio Deusdado

Compressão de Informação em Sequências Biológicas

Orientação Científica:
Prof. Doutor Paulo Carvalho

Fevereiro de 2006

1 Resumo

1.1 Introdução

O ADN constitui o meio físico onde todas as propriedades dos organismos vivos são codificadas. A codificação ocorre numa sequência de nucleótidos interligados entre si e entre os seus complementares numa estrutura de dupla hélice, os genes. Assim, a sub-cadeia … AACTGTTGTTGTTAGAA … poderá ser o exemplo de uma fracção de um "texto" que, dependendo da complexidade do organismo, poderá atingir os biliões de caracteres.
A tarefa de sequenciar o código genético de todas as espécies, incluindo aqueles que já se extinguiram, implica armazenar uma quantidade de informação colossal. Em jeito de futurismo seria como começar a escrever o inventário genómico da futura arca de Noé que será necessário levar para colonizar planetas distantes, e note-se que a natureza levou muitos milhões de anos a produzir este património. Actualmente existem em bases de dados distribuídas por todo o mundo, *gigabytes* de informação correspondente a sequências de nucleótidos, que perfazem o ADN e ARN, bem como os aminoácidos, os elementos das proteínas. Esta proliferação em massa de informação genómica implica a necessidade de algoritmos de compressão que optimizem e racionalizem o seu armazenamento e comunicação.

1.2 Resumo

Se o código ADN fosse uma *string* puramente aleatória então a entropia seria máxima e a melhor maneira de a representar seria usando dois bits por cada uma dos quatro diferentes símbolos, porém existem regularidades, propriedades específicas da sequência estatisticamente comprováveis, que provam a existência de entropia redutível, num grau ainda por descobrir, que abrem caminho a uma investigação que conduzirá a uma melhor conhecimento das razões pelas quais a Natureza usa tão peculiar código, conduzindo a descobertas filogenéticas e necessariamente à compressão da informação. A compressibilidade das sequências de ADN naturais não é linear, algumas denotam grande entropia e pouco melhor se consegue que o s 2 bits/base, noutras, os melhores algoritmos chegam a ganhar 40%, obtendo a marca de 1,6 bits/base. Os resultados são variáveis consoante a complexidade do organismo, de facto os seres eucariotas possuem um maior número de regularidades, logo possuem códigos com maior grau de compressibilidade.
É neste equipamento que surge este trabalho, que primordialmente visa fazer emergir uma nova e mais avançada metodologia para a compressão de sequências biológicas.

1.3 Objectivos Estratégicos

Optimizar a compressão em sequências biológicas, mormente em ADN, estudando e aproveitando as especificidades estatisticamente comprovadas.

2 Contribuições

Esperam-se contribuições nas áreas seguintes:
- Teoria da informação biológica;
- Filogenia;
- Biomedicina, sendo o exemplo mais sonante a luta contra o cancro.

3 Publicações

Dado o plano de doutoramento ter registado uma reestruturação, não existem ainda publicações neste novo enquadramento.

4 Enquadramento

Segundo a analogia de Ridley, o genoma (dê-se como exemplo o humano) é uma espécie de livro de tamanho gigantesco: tem «23 capítulos, chamados cromossomas; cada capítulo contém vários milhares de histórias, chamadas genes; cada história é feita de parágrafos, chamados exões (*exons*) que são interrompidos por anúncios chamados intrões (*introns*) ; cada parágrafo é feito de palavras, chamadas codões; e cada palavra é escrita com letras, chamadas bases.» Todos os seres vivos, quer se tratem de plantas, animais ou insectos, partilham código genético, o que revela um passado comum entre todas as espécies de seres vivos.
A informação genética armazenada no genoma apresenta uma organização espacial que deriva de alguns processos conhecidos e estudados, necessariamente susceptíveis de análise estatística.

4.1 Enquadramento Científico

Em 1953, Jame Watson e Francis Crick determinaram a estrutura do ácido desoxirribonucleico (ADN), e com isso germinaram uma nova extensão da ciência, a biologia molecular, que não parou de evoluir porque encontrou aliados poderosos, os computadores, e as suas capacidades de processamento e interligação, onde se estribou para a obtenção do seu marco maior, a sequenciação do genoma humano. Actualmente, somos detentores dos códigos da vida e de meios de computação poderosos, assim, torna-se possível projectar novas proteínas e novas formas de vida que interferirão necessariamente com a de todos nós. Vamos subindo degraus na pirâmide evolucional das ciências e a bioinformática [1] assume-se cada vez como uma ciência de ponta, que assenta sobretudo nos espaços de intersecção entre a biologia molecular, a informática, as bases de dados, a matemática biológica e a estatística.
O trabalho insere-se no domínio da bioinformática, onde se conjugam revelações bioestatísticas do ADN e a engenharia de software para perceber melhor os códigos da vida.

4.2 Motivação

A investigação neste início de século XXI regista um empenho e um pendor muito considerável na área das biociências, torna-se assim necessário aportar maior número de

recursos a esta área, uma área eminentemente das ciências da nova era do conhecimento. Desta vaga de investigação surgirão certamente resultados significativos para a saúde e vida de todos.
Na vertente puramente informática torna-se necessário, cada vez mais implementar medidas "ecológicas" na informação que armazenamos ou veiculamos pela Internet, assim a compressão de informação é uma necessidade óbvia, por outro lado, quase invariavelmente, as tarefas de análise, interpretação e compressão de dados estão subtilmente interligadas. As sequências biológicas, mormente o ADN, devem ser analisadas nesta perspectiva, e os resultados serão certamente utilizáveis com proficuidades por biólogos moleculares nas suas áreas de investigação.

4.3 Objectivos Detalhados

São objectivos deste trabalho:

- Desenvolver um algoritmo de compressão de sequências biológicas que supere o estado da arte nestas matérias nos casos de:
  - ADN e ARN;
  - Proteínas.

- Sabendo-se que os códigos genéticos naturais, na perspectiva da informação, são de natureza variável conforme a complexidade dos organismos, implementar um processo de compressão adaptativo;

- Priorizar o desempenho nos processos de compressão e descompressão;

- Privilegiar a eficiência no tocante às pesquisas na versão comprimida necessárias à investigação;

- Dotar o formato de condições de escalabilidade para comunicação dos dados;

- Incluir as preocupações com inferências filogenéticas resultantes da análise da sequência de AND.

4.4 Trabalhos alternativos

Em 1993 surge a primeira ferramenta de compressão de sequências de ADN, trata-se do *Biocompress* [10], baseado nos princípios de Ziv e Lempel associando detecções de factores e palindromas.

Em [11] descrevem-nos o algoritmo GenCompress que reclamou, no início deste século, o estatuto de mais avançado.

Em [12] descreve-se a investigação de compressão de ADN baseada em dicionário *off-line* usando a transformada de Burrows-Wheeler (BWT).

Em [13] descreve-se a implementação e os resultados da aplicação de um modelo NML (*normalized maximum likelihood*) na compressão sem perdas de ADN.

Em [14], o desafio da compressão de ADN foi revisitado e nele nos dão a conhecer o algoritmo *DNAPack*, que a seguir se compara aos restantes de referência fazendo uso da tabela seguinte.

sequence	length	BioCompress-2	GenCompress	CTW-LZ	DNACompress	DNAPack
CHMPXX	121024	1.6848	1.6730	1.6690	1.6716	**1.6602**
CHNTXX	155844	1.6172	1.6146	1.6120	1.6127	**1.6103**
HEHCMVCG	229354	1.8480	1.8470	1.8414	1.8492	**1.8346**
HUMDYSTROP	33770	1.9262	1.9231	1.9175	1.9116	**1.9088**
HUMGHCSA	66495	1.3074	1.0969	1.0972	**1.0272**	1.039
HUMHBB	73308	1.8800	1.8204	1.8082	1.7897	**1.7771**
HUMHDABCD	58864	1.8770	1.8192	1.8218	1.7951	**1.7394**
HUMHPRTB	56737	1.9066	1.8466	1.8433	1.8165	**1.7886**
MPOMTCG	186609	1.9378	1.9058	1.9000	**1.8920**	1.8932
PANMTPACGA	100314	1.8752	1.8624	1.8555	1.8556	**1.8535**
VACCG	191737	1.7614	1.7614	1.7616	**1.7580**	1.7583
Average	—	1.7837	1.7428	1.7389	1.7254	**1.7148**

Figura 1 - Os algoritmos de compressão de ADN de referência e a sua comparação de desempenho em bits/base.

4.5 Bibliografia Principal

[1] J. Cohen, "Computer Science and Bioinformatics," *Communications of the ACM*, vol. 48, 2005.

[2] D. R. Powell, D. L. Dowe, L. Allison, and T. I. Dix, "Discovering Simple DNA Sequences by Compression," *Department of Computer Science, Monash University, Clayton, Australia*, 1998.

[3] J. R. Lobry, "The Black Hole Of Symmetric Molecular Evolution." Lyon, France: University of Lyon, 2000.

[4] J. K. Lanctot, M. Li, E. Yang, and, "Estimating DNA sequence entropy," presented at Symposium on Discrete Algorithms, 2000.

[5] D. Loewenstern and P. N. Yanilos, "Significantly Lower Entropy Estimates for Natural DNA Sequences," *Computational Biology*, vol. 6, nº1, 1997.

[6] H. Herzel, "Complexity of Symbol Sequences," *Systems Analysis Modelling Simulation*, vol. 5, pp. 435-444, 1988.

[7] M. Farach, M. Noordewier, S. Savari, L. Shepp, A. Wyner, and J. Ziv, "On the entropy of DNA: algorithms and measurements based on memory and rapid convergence," presented at Sixth annual ACM-SIAM symposium on Discrete algorithms, San Francisco, California, 1995.

[8] L. Gatlin, *Information Theory and the Living Systems*: Columbia University Press, 1972.

[9] K. Sayood, *Lossless Compression Handbook*: Academic Press, Elsevier Science USA, 2003.

[10] S. Grumbach and F. Tahi, "Compression of DNA Sequences," *IEEE*, pp. 340-350, 1993.

[11] X. Chen, S. Kwong, and M. Li, "A compression algorithm for DNA sequences and its applications in genome comparison," presented at Annual Conference on Research in Computational Molecular Biology, Tokyo, Japan, 2000.

[12] D. Adjeroh, Y. Zhang, A. Mukherjee, M. Powell, and T. Bell, "DNA Sequence Compression Using the Burrows-Wheeler Transform," presented at IEEE Computer Society Bioinformatics Conference (CSB'02), 2002.
[13] I. Tabus, G. Korodi, and J. Rissanen, presented at Data Compression Conference (DCC'03), 2003.
[14] B. Behzadi and F. L. Fessant, "DNA Compression Challenge Revisited," presented at Symposium on Combinatorial Pattern Matching (CPM'2005), Korea, 2005.
[15] M. Farach and S. Kannan, "Efficient algorithms for inverting evolution," *Journal of the ACM (JACM)*, vol. 46, pp. 437-449, 1999.