ANÁLISE DE REPETIÇÕES EM DADOS BIOLÓGICOS

JOSÉ P. LOUSADO*
JOSÉ L. OLIVEIRA**

* Docente da Escola Superior de Tecnologia e Gestão de Lamego e investigador do Centro de Estudos em Educação, Tecnologias e Saúde (CI&DETS) do Instituto Politécnico de Viseu.

** Docente da Universidade de Aveiro.

Resumo

A descodificação dos genomas veio criar novos desafios na comunidade científica ligada à área da computação e da informática. Diariamente são alimentadas inúmeras bases de dados com biliões de registos provenientes de equipamentos cada vez mais evoluídos, que auxiliam na descodificação dos genomas. Determinar o quão importante e relevante são esses dados, de forma a retirar valor acrescentado – informação, e obviamente transformá-los em conhecimento, é o grande desafio actual para a comunidade de investigadores de bioinformática. A análise de genomas, bem como dos proteomas dos vários organismos permitem-nos observar o comportamento ao nível da evolução das espécies. Neste estudo focamos a atenção num aspecto particular dessa análise: as repetições de determinados codões e dos respectivos aminoácidos nos vários organismos eucariotas, especificamente em genes ortólogos. Pertencente a várias fases da evolução das espécies, o objectivo principal centra-se na obtenção de resultados quanto à evolução dessas repetições ao longo de milhões de anos. Sabemos hoje que essas repetições no ser humano são a causa de diversas doenças neuro-degenerativas, entre outras, pelo que esta análise permitirá verificar o estado de conservação ou repressão, dessas repetições ao longo do processo de especiação, bem como ao nível do relacionamento que poderá existir entre essas repetições e as doenças nos seres superiormente evoluídos. Para este estudo foi desenvolvido um algoritmo de detecção de padrões de repetição, que possibilita uma análise detalhada da localização de uma determinada sequência, bem como das sequências que melhor se ajustam ao padrão de repetição inicial.

Palavras-chave: Bioinformática, Detecção de padrões, Sistemas de Apoio à Decisão.

Abstract

The decoding of the genomes has created new challenges on the scientific community linked to the area of computation and information technologies. Daily, new data is added to numerous databases with billions of records, coming from more advanced equipment, helping in decoding the genomes. Determine how important and relevant are these data in order to find value-added information and obviously turn them into knowledge,is the main challenge for the bioinformatics research community. The analysis of genomes and proteomes of several organisms allow us to observe the behaviour at the evolution of species. In this study, our focus goes to a particular aspect of this analysis: the repetition of some codons and their amino acids inside several orthologous genes in eukaryotic organisms. Belonging to different stages of evolution, the main objective focuses on achieving results on the evolution of these repetitions over millions of years. We now know that these repetitions in humans are the source of several neurodegenerative diseases among others. This analysis will verify the conservation or repression, of these repetitions throughout the process of speciation as well at the level of relationship that may exist between these repetitions and those diseases. For this study we have developed an algorithm for detecting patterns of repetition, which allows a detailed analysis of the location of each sequence and the sequences that best fit the initial pattern.

Keywords: bioinformatics, pattern detection, Decision Support Systems

1. Introdução

Desde há vários anos que se vêm a estudar as repetições de certos codões e respectivos aminoácidos em determinados genes, sendo várias as áreas de interesse, desde a análise evolutiva das espécies, com conservação, redução ou expansão dessas repetições (Sher Ali & Rajesh Gopal, 1998), até aos organismos mais evoluídos, nomeadamente o homem, no que às doenças genéticas diz respeito. Doenças como a Doença de Huntington's (Ferro, Catalano, Dell'Eva, Fortunati, & Pfeffer, 2002;

Herishanu, et al., 2009; Paul, 2007; Pearson C.E., 2005), entre outras doenças neurodegenerativas, estão comprovadamente relacionadas com essas repetições em determinados genes (Freed, Cooper, Brennecke, & Moses, 2005; Pearson C.E., 2005), bem como na relação com alguns tipos de Cancro (Ferro, et al., 2002; Pearson C.E., 2005).

Perante este cenário real, várias questões surgiram, nomeadamente, saber qual a evolução que essas cadeias de codões e obviamente de aminoácidos tiveram ao longo do tempo. Terão sido reprimidas? Ou pelo contrário, terão sido ampliadas? Terá esse fenómeno influência significativa ao nível da especiação e evolução dos organismos?

Atendendo à evolução genética das espécies, este estudo foca-se na evolução de determinados genes homólogos[1] (ortólogos[2]) (Fu & Jiang, 2008) ao longo da cadeia evolutiva de diversos organismos, seleccionando-se à partida os genes que apresentavam repetições no mínimo de 10 aminoácidos repetidos, sendo essa uma regra assumida como sendo a mínima aceitável, referida na literatura por alguns autores (Gabriela Moura, 2005; Jones & Pevzner, 2006).

Tratando-se tipicamente de uma análise de padrões e existindo um vasto leque de aplicações que poderiam facilitar a obtenção de resultados, verificou-se que a utilização dessas ferramentas disponíveis on-line (Gordon, Nekludova, McCallum, & Fraenkel, 2005; Pearson, Wood, Zhang, & Miller, 1997; Stoye, 1997; Tatusova & Madden, 1999) (BLASTx, FASTA, etc.) não corresponderam ao pretendido, uma vez que a comparação entre duas sequências não se resume a um alinhamento destas, mas antes à obtenção da subsequência dentro de um determinado gene, que seja o mais aproximada possível da sequência de repetição de codões ou de aminoácidos, obtida a partir do genoma do organismo ancestral. Essa subsequência terá normalmente um comprimento entre 10 e 70 codões ou aminoácidos, podendo nalguns casos ser maior. Por esse motivo, os algoritmos de alinhamento típico, como os referidos anteriormente, apresentam resultados bastante distantes, tendo mesmo nalguns casos apresentado o resultado de "não existência de similaridade" (Tatusova & Madden, 1999). Como resposta ao pretendido foi implementado um algoritmo baseado na distância de Levenshtein (Levenshtein, 1966), que permite encontrar a subsequência dentro de um gene de um organismo superior, que esteja à "menor distância" da subsequência do organismo ancestral – a melhor aproximação. Para além desse algoritmo foi adoptada a metodologia referida em (Moura, et al., 2007), para a contagem de repetições de codões, sendo neste caso implementada uma aplicação específica para esse fim, mas para aminoácidos, integrada com a ferramenta indicada anteriormente.

[1] Genes que têm origem num gene comum, que evoluíram de forma diferente podendo ter funções distintas.
[2] Genes homólogos mas que concorrem em espécies diferentes, tendo, no entanto, um ancestral comum.

2. Detecção de subsequências repetidas de aminoácidos

Como já foi referido anteriormente, para suportar o estudo foram seleccionados 8 organismos eucariotas (Tabela 1), seguindo-se a evolução natural na ordem filogenética. Atendendo ao facto de que o organismo Schizosaccharomyces pombe é o mais antigo, foi eleito como o ponto de partida para o trabalho que se seguiu.

Tabela 1 – Lista de organismos do estudo e respectiva fonte de dados original

Organism (Kegg ID)	Geneal. Tree Order	Source Database
Schizosaccharomyces pombe (spo)	1	ftp://ftp.sanger.ac.uk/pub/yeast/pombe/CDS_bases/
Aspergillus fumigatus (afm)	2	ftp://ftp.ncbi.nih.gov/genomes/Fungi/Aspergillus_fumigatus/
Candida albicans (cal)	3	http://www.candidagenome.org/download/sequence/ Assembly21/archive/orf_genomic_assembly_21.20081020.fasta.gz
Sacharomyces cerevisiae (sce)	4	ftp://ftp.ncbi.nih.gov/genomes/Fungi/Saccharomyces_cerevisiae/
Arabidopsis thaliana (ath)	5	ftp://ftp.ncbi.nih.gov/genomes/Arabidopsis_thaliana/
Drosophila melanogaster (dme)	6	ftp://ftp.ensembl.org/pub/release-52/fasta/drosophila_melanogaster /cdna/Drosophila_melanogaster.BDGP5.4.52.cdna.abinitio.fa.gz
Mus musculus (mmu)	7	ftp://ftp.ensembl.org/pub/release-52/fasta/mus_musculus/ cdna/Mus_musculus.NCBIM37.52.cdna.abinitio.fa.gz
Homo sapiens (hsa)	8	ftp://ftp.ensembl.org/pub/release-52/fasta/homo_sapiens/ cdna/Homo_sapiens.NCBI36.52.cdna.abinitio.fa.gz

Sabendo à partida que o nosso interesse são as repetições de codões e principalmente de aminoácidos, o estudo iniciou-se com análise do genoma (apenas a região codificante – orfeoma) do organismo Schizosaccharomyces pombe, tendo-se isolado os genes que possuem as subsequências de aminoácidos iguais, em número superior ou igual a 10. Nessa fase identificamos cada gene, o aminoácido repetido, o número de repetições, bem como a sua localização registando-se a posição inicial da contagem. Como forma de controlo, foram efectuados os mesmos procedimentos para os restantes organismos. A detecção de repetições idênticas dos mesmos aminoácidos noutros organismos, por si só não é significativa, uma vez que falta saber se os genes onde as repetições são detectadas, são ortólogos. Para esse efeito, foram seleccionados os organismos a partir da base de dados KEGG Orthology ("KEGG: Kyoto Encyclopedia of Genes and Genomes,").

Uma vez determinadas as repetições de codões, avançou-se para a detecção de repetições de aminoácidos, pois os resultados variam, obtendo-se um número mais elevado de repetições destes últimos, como seria de esperar.

Os genes com maior número de repetições no organismo Schizosaccharomyces pombe estão apresentados na tabela 2.

Tabela 2 – Lista de genes com repetições de codões >=10 em Schizosaccharomyces pombe

Gene ID	Aminoácido	Posição inicial	Nº de repetições
>SPBC30B4.01C	SER (S)	466	52
>SPBC146.01	GLN (Q)	769	33
>SPCC553.10	SER (S)	433	15
>SPBC30B4.01C	SER (S)	655	14
>SPAC13F5.02C	GLU (E)	931	13

Os genes seleccionados para o estudo estão indicados na coluna respectiva. Embora o gene SPBC30B4.01C apresente um número de repetições superior (52), foi descartado por não apresentar, a partir do organismo Candida albicans, genes ortólogos na cadeia evolutiva. Dessa forma o estudo centrou-se nos três genes mais representativos no organismo ancestral, para os quais existem genes ortólogos em todos os organismos referidos anteriormente. Da literatura, importa referir que os três genes estão relacionados com o respectivo gene ortólogo no ser humano com: SPBC146.01 (MAML in Homo sapiens) - mucoepidermoid carcinomas, benign Warthin tumors and clear cell, hidradenomas (Afrouz Behboudi, 2005), SPCC553.10 (DSPP in Homo sapiens) –dentine disorders and others, Hypophosphatemia, que são doenças genéticas degenerativas (Lorenz-Depiereux, et al., 2006; MacDougall, et al., 1997), e SPAC13F5.02C (TAF7 in Homo sapiens, uma proteína complexa que desempenha um papel central na regulação do promotor, dando resposta a vários activadores e repressores (Dephoure, et al., 2008).

3. ADAS: Algoritmo de Detecção Automática de Sequências

O algoritmo implementado para descoberta de subsequências em genes dos vários organismos em estudo, que sejam idênticos ou muito próximos da sequência encontrada em cada um dos genes, teve como princípio a divisão da sequência inicial em sequências mais pequenas. Estas são subdivididas sempre que não seja detectada a presença de uma sequência idêntica à sequência inicial.

Após a subdivisão da sequência original em várias subsequências, inicia-se o processo de detecção da subsequência maior que se encontra integralmente no gene em estudo. Desta forma, garante-se que ao ser encontrada uma subsequência da sequência inicial, essa estará provavelmente na região onde a conservação terá ocorrido. Posteriormente, efectua-se o varrimento por “Brute force” da zona onde essa subsequência é encontrada num intervalo de aproximadamente metade do tamanho da

sequência inicial, quer à esquerda quer à direita da sequência detectada, garantindo assim que a zona de vizinhança dessa sequência será maior em tamanho, do que a sequência inicial. Para cada subsequência encontrada durante o varrimento da zona de vizinhança dessa sequência, é aplicado o algoritmo da distância de Levenshtein (Levenshtein, 1966), guardando a sequência que tiver menor distância (por inserção, remoção ou alteração). Essa será à priori a subsequência do gene que melhor se adaptará à sequência inicial.

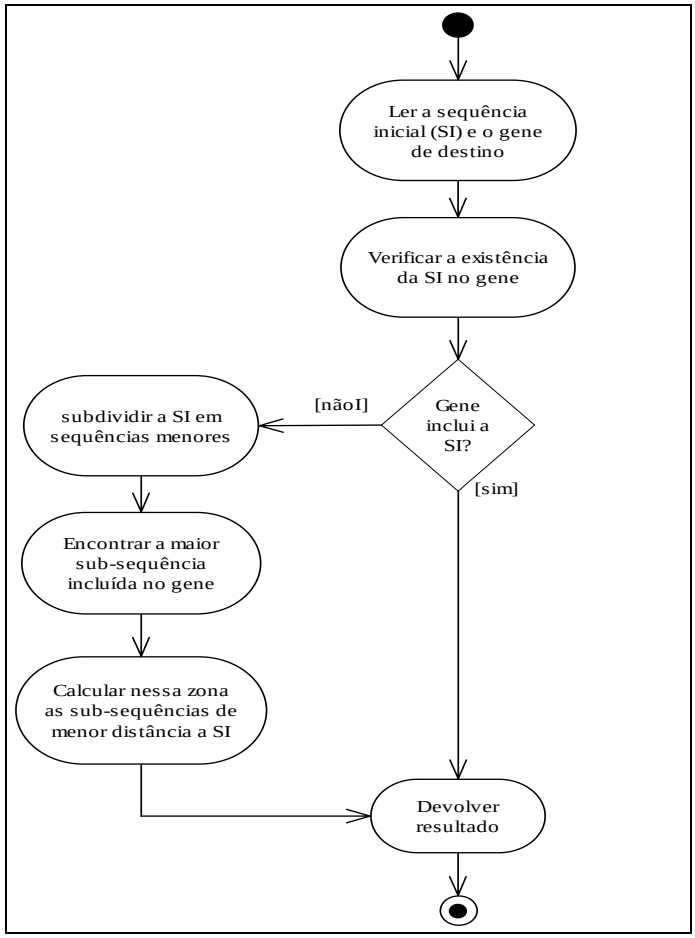


Figura 1 – Diagrama de actividades de especificação do algoritmo ADAS

Atendendo ao facto de que pode existir mais do que uma sequência com nível de similaridade igual, o algoritmo percorre todo o gene, aplicando a mesma técnica sempre que uma sequência nessas condições for encontrada. No final, são apresentadas as subsequências encontradas, cuja distância à sequência inicial seja mínima. Caberá ao investigador decidir, de acordo com os critérios biológicos, qual a sequência que deverá seleccionar.

4. Campo de aplicação

Neste estudo foram aplicadas, para além de técnicas e modelos já devidamente validados, nomeadamente GeneSPLIT (Lousado, Moura, Santos, & Oliveira, 2008; Moura, et al., 2007) , novas metodologias para detecção de padrões de repetições (codões e aminoácidos), através de um algoritmo dinâmico, em que apenas é fornecida à partida a cadeia de aminoácidos que se pretende comparar. De notar que o algoritmo pode ser também aplicado com qualquer sequência, independentemente de ser com repetições ou não, abrindo o leque de possibilidades para a descoberta de outros padrões.

Para cada uma das sequências dos três genes em estudo, foram encontradas as sequências dos outros genes ortólogos que melhor se aproximam da sequência inicial, estando esses valores apresentados na tabela 3.

Numa análise mais pormenorizada, salienta-se o facto de a repetição de 33 resíduos de glutamina (GLN or Q) que surge no gene SPBC146.01 do organismo Schizosaccharomyces pombe, não ser detectável nos organismos Aspergillus fumigatus, Candida albicans, Sacharomyces cerevisiae e Arabidopsis thaliana. No entanto, nos organismos Drosophila melanogaster, Mus musculus, e em particular Homo sapiens, essa repetição conserva-se, chegando a expandir-se neste último para 34 numa primeira sequência, seguida de 27 repetições desse mesmo aminoácido (ver Figura 3). É conhecido que o codão CAG, está associado a algumas doenças neurodegenerativas do ser humano (Ferro, et al., 2002). Acontece porém, que os codões que descodificam a glutamina em Schizosaccharomyces pombe são maioritariamente CAA, enquanto que no Homo sapiens são, no gene ortólogo (MAML), maioritariamente CAG, havendo uma abundância de CAG na região traduzida, interrompida por alguns codões de outros aminoácidos.

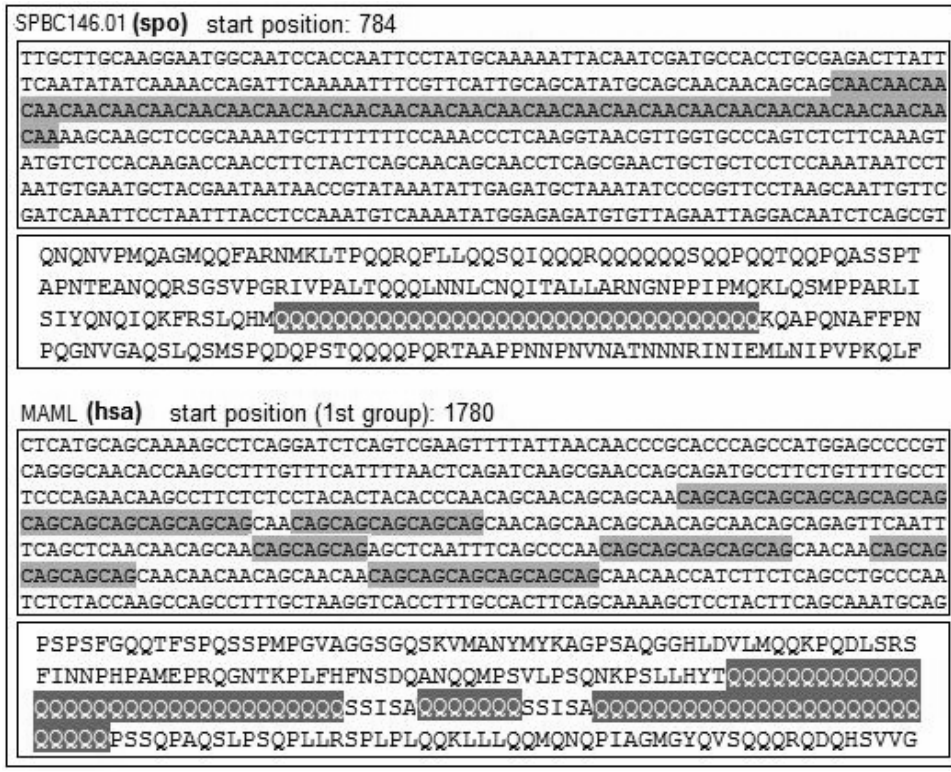


Figura 2 – Comparação entre codões e aminoácidos de dois genes ortólogos, de *Schizosaccharomyces pombe (spo)* e *Homo Sapiens (hsa)*.Para cada gene, estão representados os codões com o fundo cinzento e texto a preto e os respectivos aminoácidos com cor branca no texto e o fundo cinzento.

Podemos portanto discutir o impacto que o aumento do aminoácido glutamina teve na evolução deste gene, e por outro lado, evidenciar a alteração que ocorreu ao longo dessa evolução, uma vez que houve mudança ao nível dos codões sinónimos (CAA to CAG) que pode em última análise, ter influência sobre a saúde pública, uma vez que a presença de CAGs neste gene, está associada na literatura a doenças extremamente graves para os humanos.

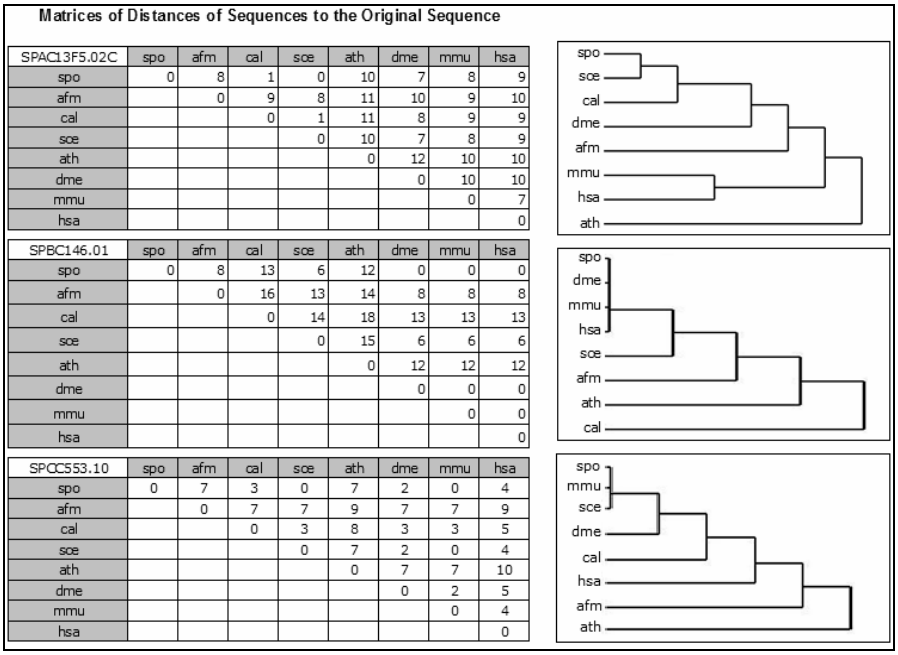
Tabela 3 – Os melhores resultados obtidos para cada organismo nos respectivos genes ortólogos do gene original em Schizosaccharomyces pombe a partir do algoritmo ADAS.

GeneID		**SPAC13F5.02C**	spo
Base String		EEEEEEEEEEEE	
Distância aos ortólogos	8	GEYYEEDEYYDDE	afm
	1	EEEEEEEEEEENEE	cal
	0	EEEEEEEEEEEE	sce
	10	EPDLNPELVQRVE	ath
	7	EEREDETEKESPN	dme
	8	EDEEDVNILDTEE	mmu
	9	EEDINIIDTEEDLE	hsa

GeneID		**SPBC146.01**	spo
Base String		QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ	
Distância aos ortólogos	8	QQQQQQQQQSQQQQQQQSQQNQAMLQQQRVQQ	afm
	13	NQQQLSQIPNQQQQQQQQQQQVPQSQPHASQQ	cal
	6	QQMQHLQQLKMQQQQQQQQQQQQQQQQQQQQQ	sce
	12	QQFQQRQMQQQQLQARQQQQQQQLQARQQAAQLQQ	ath
	0	QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ	dme
	0	QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ	mmu
	0	QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ	hsa

GeneID		**SPCC553.10**	spo
Base String		SSSSSSSSSSSSSS	
Distância aos ortólogos	7	GSASTSSSSTGTVSS	afm
	3	SSSSTSSSSSSTPS	cal
	0	SSSSSSSSSSSSSS	sce
	7	SSSSSFSFGTSANSG	ath
	2	SSSSSSSSSSTSSK	dme
	0	SSSSSSSSSSSSSS	mmu
	4	SSDSSDSSDSSSSSDSS	hsa

A aplicação desenvolvida para o efeito de validação do modelo e do algoritmo, permite realizar clustering entre sequências encontradas, tendo por base as distâncias entre elas. (Figura 3). Essa ferramenta permite que se analise o grau de proximidade entre os organismos em estudo, no que às sequências encontradas diz respeito, podendo tirar-se conclusões também relativamente à evolução dos genes entre pares de organismos.


Matrices of Distances of Sequences to the Original Sequence

SPAC13F5.02C	spo	afm	cal	sce	ath	dme	mmu	hsa
spo	0	8	1	0	10	7	8	9
afm		0	9	8	11	10	9	10
cal			0	1	11	8	9	9
sce				0	10	7	8	9
ath					0	12	10	10
dme						0	10	10
mmu							0	7
hsa								0

SPBC146.01	spo	afm	cal	sce	ath	dme	mmu	hsa
spo	0	8	13	6	12	0	0	0
afm		0	16	13	14	8	8	8
cal			0	14	18	13	13	13
sce				0	15	6	6	6
ath					0	12	12	12
dme						0	0	0
mmu							0	0
hsa								0

SPCC553.10	spo	afm	cal	sce	ath	dme	mmu	hsa
spo	0	7	3	0	7	2	0	4
afm		0	7	7	9	7	7	9
cal			0	3	8	3	3	5
sce				0	7	2	0	4
ath					0	7	7	10
dme						0	2	5
mmu							0	4
hsa								0

Figura 3 – Matrizes de distâncias entre as sequências detectadas e respectivos clusters

5. Conclusão

A integração de algoritmos específicos em ferramentas computacionais com a biologia deu origem a uma área recente de investigação - a bioinformática. Essas ferramentas, no presente caso a implementação do algoritmo apresentado - ADAS, veio permitir verificar que a conservação de genes ao longo da árvore filogenética, genes ortólogos, não é homogénea, permitindo aos biólogos retirar conclusões que até à data eram desconhecidas.

É do conhecimento da comunidade científica que as repetições nos genes ortólogos estão relacionadas com a evolução das espécies, sendo por si só um importante campo de observação e de investigação, uma vez que essa conservação, com maior ou menor relevância, nomeadamente ao nível da expressão e função do gene, tem um grande impacto ao nível do processo de especiação. Ao debruçarmo-nos sobre um aspecto em particular, neste estudo, as repetições de determinados codões, em várias espécies eucariotas, verificou-se que um dos genes com maior número de repetições no organismo ancestral, no caso Schizosaccharomyces pombe, evoluiu até ao eucariota superior – Homo sapiens, de uma forma não uniforme, uma vez que organismos de posição intermédia na árvore filogenética reprimiram essa sequência, tendo a mesma

surgido nos seres eucariotas superiores novamente, só que desta forma associada não unicamente ao codão CAA, mas principalmente associada ao codão CAG, com as implicações que daí resultam ao nível da conservação das espécies.

REFERÊNCIAS BIBLIOGRÁFICAS

- AFROUZ Behboudi, M. W., Ludmila GORUNOVA, Joost J. van den OORD, Fredrik MERTENS, Fredrik ENLUND, Göran STENMAN. (2005). Clear cell hidradenoma of the skin - a third tumor type with a t(11;19)-associated TORC1-MAML2 gene fusion. *Genes, Chromosomes and Cancer, 43*(2), 202-205.
- DEPHOURE, N., ZHOU, C., VILLÃ©N, J., BEAUSOLEIL, S. A., BAKALARSKI, C. E., ELLEDGE, S. J., et al. (2008). A quantitative atlas of mitotic phosphorylation. *Proceedings of the National Academy of Sciences, 105*(31), 10762-10767.
- FERRO, P., CATALANO, M. G., DELL'EVA, R., FORTUNATI, N., & PFEFFER, U. (2002). The androgen receptor CAG repeat: a modifier of carcinogenesis? *Molecular and Cellular Endocrinology, 193*(1-2), 109-120.
- FREED, K. A., COOPER, D. W., BRENNECKE, S. P., & MOSES, E. K. (2005). Detection of CAG repeats in pre-eclampsia/eclampsia using the repeat expansion detection method. *Mol. Hum. Reprod., 11*(7), 48 - 87.
- FU, Z., & JIANG, T. (2008). Clustering of main orthologs for multiple genomes. *J Bioinform Comput Biol, 6*(3), 573-584.
- Gabriela MOURA, M. P., Raquel SILVA, Isabel MIRANDA, Vera AFREIXO, Gaspar DIAS, Adelaide FREITAS, José L OLIVEIRA, and Manuel AS SANTOS. (2005). Comparative context analysis of codon pairs on an ORFeome scale. *Genome Biology, 6*(3).
- GORDON, D. B., NEKLUDOVA, L., MCCALLUM, S., & FRAENKEL, E. (2005). TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. *Bioinformatics, 21*(14), 3164-3165.
- HERISHANU, Y. O., PARVARI, R., POLLACK, Y., SHELEF, I., MAROM, B., MARTINO, T., et al. (2009). Huntington disease in subjects from an Israeli Karaite community carrying alleles of intermediate and expanded CAG repeats in the HTT gene: Huntington disease or phenocopy? *Journal of the Neurological Sciences, 277*(1-2), 143-146.
- JONES, N. C., & PEVZNER, P. A. (2006). Comparative genomics reveals unusually long motifs in mammalian genomes. *Bioinformatics, 22*(14), e236-242.
- KEGG: Kyoto Encyclopedia of Genes and Genomes. from http://www.kegg.com
- LEVENSHTEIN, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl, 10*(8), 707–710.
- LORENZ-DEPIEREUX, B., BASTEPE, M., BENET-PAGES, A., AMYERE, M., WAGENSTALLER, J., MULLER-BARTH, U., et al. (2006). DMP1 mutations in autosomal recessive hypophosphatemia implicate a bone matrix protein in the regulation of phosphate homeostasis. *Nat Genet, 38*(11), 1248-1250.
- LOUSADO, J. P., MOURA, G. R., SANTOS, M. A. S., & OLIVEIRA, J. L. (2008). Exploiting Codon-Triplets Association for Genome Primary Structure Analysis, *Biocomputation, Bioinformatics, and Biomedical Technologies, 2008. BIOTECHNO '08. International Conference on* (pp. 155-158). Bucharest, Romania: IEEE Xplorer.
- MACDOUGALL, M., SIMMONS, D., LUAN, X., NYDEGGER, J., FENG, J., & GU, T. T. (1997). Dentin Phosphoprotein and Dentin Sialoprotein Are Cleavage Products Expressed from a Single Transcript Coded by a Gene on Human Chromosome 4. Dentin phosphoprotein dna sequence determination. *J. Biol. Chem., 272*(2), 835-842.

- MOURA, G., LOUSADO, J., PINHEIRO, M., CARRETO, L., SILVA, R., OLIVEIRA, J., et al. (2007). Codon-triplet context unveils unique features of the Candida albicans protein coding genome. *BMC Genomics, 8*, 444.
- PAUL, S. (2007). Polyglutamine-Mediated Neurodegeneration: Use of Chaperones as Prevention Strategy. *Biochemistry (Moscow), 72*(4), 359-366.
- PEARSON C.E., N. E. K., CLEARY J.D. . (2005). Repeat instability: mechanisms of dynamic mutations. [Review]. *Nat Rev Genet., 6*(10), 729-742.
- PEARSON, W. R., WOOD, T., ZHANG, Z., & MILLER, W. (1997). Comparison of DNA Sequences with Protein Sequences. *Genomics, 46*(1), 24-36.
- SHER ALI, S. A., Nasreen Z. EHTESHAM, Md. ASIM AZFER, Uday HOMKAR,, & Rajesh GOPAL, S. E. H. (1998). Analysis of the evolutionarily conserved repeat motifs in the genome of the highly endangered central Indian swamp deer Cervus duvauceli branderi. *GENE, 223*, 361–367.
- STOYE, J. (1997). *Divide-and-Conquer Multiple Sequence Alignment.* Universität Bielefeld.
- TATUSOVA, T. A., & MADDEN, T. L. (1999). BLAST 2 S, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters, 174*(2), 247-250.