An integrative approach for codon repeats evolutionary analyses

José Paulo Lousado*

Centro de Estudos em Educação,
Tecnologias e Saúde,
ESTGL, Instituto Politécnico de Viseu,
Campus Politécnico de Viseu, 3504-510 Viseu, Portugal
E-mail: jlousado@estgl.ipv.pt
E-mail: lousado@ua.pt
*Corresponding author

José Luis Oliveira

Department of Electronics, Telecommunications
and Informatics (DETI/IEETA),
University of Aveiro,
3810-193 Aveiro, Portugal
E-mail: jlo@ua.pt

Gabriela Moura and Manuel A.S. Santos

Department of Biology (CESAM),
University of Aveiro,
3810-193 Aveiro, Portugal
E-mail: gmoura@ua.pt
E-mail: msantos@ua.pt

Abstract: The relationship between genome characteristics and several human diseases has been a central research goal in genomics. Many studies have shown that specific gene patterns, such as amino acid repetitions, are associated with human diseases. However, several open questions still remain, such as, how these tandem repeats appeared in the evolutionary path or how they have evolved in orthologous genes of related organisms. In this paper, we present a computational solution that facilitates comparative studies of orthologous genes from various organisms. The application uses various web services to gather gene sequence information, local algorithms for tandem repeats identification and similarity measures for gene clustering.

Keywords: codon repeats; evolutionary analyses; orthologous genes; gene comparison; tandem repeats; data integration; web services; bioinformatics.

Reference to this paper should be made as follows: Lousado, J.P., Oliveira, J.L., Moura, G. and Santos, M.A.S. (2012) 'An integrative approach for codon repeats evolutionary analyses', *Int. J. Data Mining and Bioinformatics*, Vol. 6, No. 4, pp.369–381.

Biographical notes: J.P. Lousado is a Professor in the School of Technology and Management of Lamego, Polytechnic Institute of Viseu and currently a PhD student in the Department of Electronic, Telecommunications and Informatics at the University of Aveiro. His research interests include codon context analysis, genome analysis and data integration using bioinformatics.

J.L. Oliveira is an Associated Professor at the Department of Electronic, Telecommunications and Informatics at the University of Aveiro (UA). He is the coordinator of the bioinformatics group of UA (http://bioinformatics.ua.pt/) and scientific coordinator of the bioinformatics unit at BIOCANT (www.biocant.pt). His main research interests are in the area of distributed systems and computational applications for bioinformatics and biomedical informatics.

G. Moura is an Assistant Researcher at the Centre for Environmental and Marine Studies (CESAM) of the University of Aveiro. Her main research interests include the evaluation of the impact of the error of mRNA decoding in cell degeneration and evolution, through the development and use of comparative genomics, bioinformatics and biostatistics methodologies.

M.A.S. Santos is an Associated Professor at the Department of Biology of the University of Aveiro. He runs the RNA Biology Laboratory of the University of Aveiro, where a number of post-graduate students are dedicated to the structural and functional analysis of small RNAs, focusing mainly on their role in the evolution of life and gene expression control.

1 Introduction

The analysis of protein primary structures, as well as their evolution over time, has been a highly studied area from the point of view of the evolutionary chain. Many studies (Ali et al., 1998; George et al., 2006; Jones and Pevzner, 2006; Fu and Jiang, 2008) showed the relationship between some human genes and various illnesses, such as cancer (Pearson, 2007), neurodegenerative disorders, and others (Bowen et al., 2000; Brameier and Wiuf, 2007). Many other studies focus on certain parts of the genome that have been important for the survival of the human species (Pestova et al., 1994; Ferro et al., 2002; Freed et al., 2005; Herishanu et al., 2009). These refer specifically to the repetition of certain codons and/or amino acids and have allowed to predict possible diseases and identify useful treatments, namely patient oriented medication (Hsueh, 2006; Bogaerts et al., 2008; Mena et al., 2008; Tarini et al., 2009). Pearson and Cleary (2005) identified a set of genes with repetitions that are related to several diseases.

One way of determining how these repeated regions evolved is to track the orthologous genes from related species so that they can be aligned with the sequence from the human gene. Two types of homologous genes can be distinguished: orthologues, which are defined as genes in different species that have evolved from a common ancestor, and paralogues, which originated from duplication in the same species.

Under the context of determining the extent to which repetitions are present in orthologous genes in several organisms, a set of biological questions arises, such as how

did these amino acids sequences evolved over time? Are they under some set of negative pressure? Could this phenomenon have influenced speciation and the evolution of organisms?

The OMIM database (OMIM, 2009) can be used to identify the association between diseases and genes (Hamosh et al., 2005). Additionally, orthologous sequences can be retrieved from KEGG (Ogata et al., 1999; KEGG, 2010). However, querying and relating data from a set of orthologous genes (especially those with repetitions) are extremely time-consuming tasks, exacerbated by the fact that there is no integrative tool that allows performing these comparisons in an automated manner for a large number of organisms. As such, developing a bioinformatics application to perform these tasks in an autonomous and integrated way is important to allow rapid data analysis.

In Lousado et al. (2009) we presented an algorithm that allows the identification of codon repetition regions in genome sequences. In the present paper we have extended this work by developing an integrative software application that facilitates the comparison of disease-related genes from humans with the respective orthologous gene sequences from other organisms, so as to perform evolutionary studies. In order to illustrate this idea, we focused on determining whether existing repetitions that cause diseases in humans have propagated from less evolved organisms, and how that propagation occurred, that is, with a decrease or an increase in the number of repetitions along the evolutionary chain.

Special emphasis was given in this work to facilitating the retrieve of orthologous genes. For example, they are downloaded in a versatile format so that the data may be saved locally as text files for later use off-line. The goal was to achieve automatic analysis of amino acid repetitions in the various orthologous genes. The multi-window functionality introduced in the application is also important, since the user does not lose data from query to query, being able to open up to ten completely independent windows for each of the options. Moreover, even if windows are closed, they remain easily accessible from a drop-down menu.

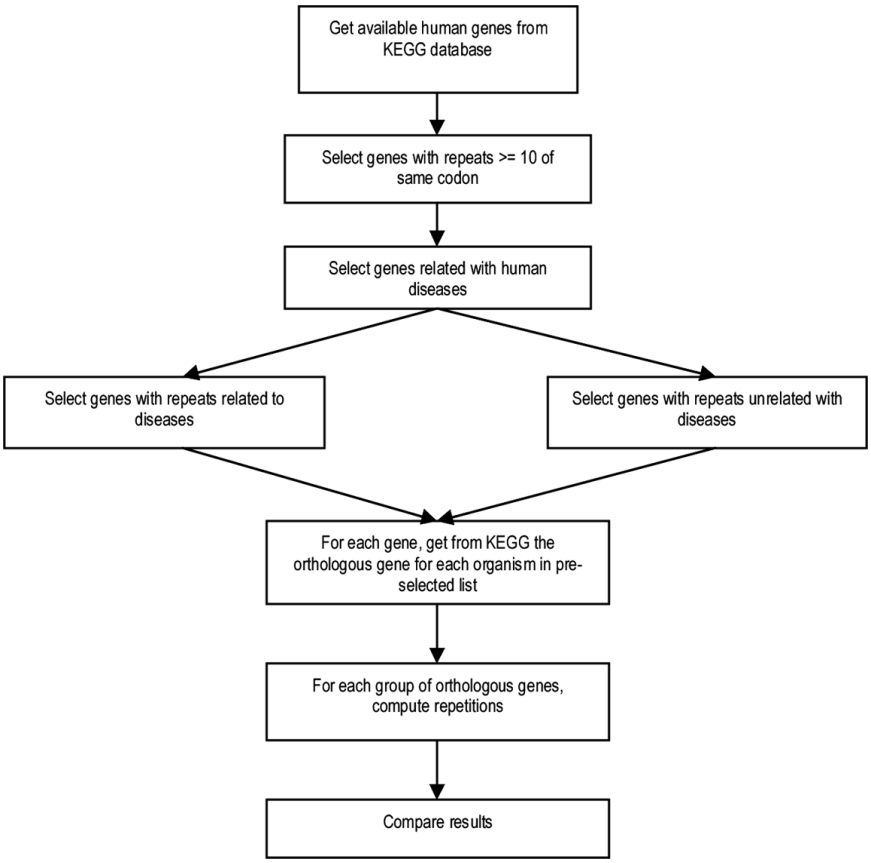
2 Methods

2.1 Integration workflow

For this approach we have used web services that are already available in several biological databases, such as KEGG (2010). This web-based technology facilitates data gathering, through a standardised programmatic interface. A software application can retrieve information on demand, as needed, avoiding downloading extensive data, typically through file transfer, and allowing just to extract a small part of the available data.

In order to carry out this work we have developed a standalone application, following a specific workflow (Figure 1). It starts with genes that have been previously identified as those implicated in diseases and iteratively constructs a relationship between them and their orthologous genes from various organisms, allowing to study codon/amino acid repetitions. Following the work presented by Jones and Pevzner (2006), we assume a default value of 10 as the minimum representative number of consecutive codon and/or amino acid that appear repeated.

Figure 1 Data integration workflow



The amino acid and codon data are extracted from KEGG reference database. From this database, the application identifies the genes that have at least 10 consecutive repeated codons – the predefined size threshold.

Once the genes and respective repetitive sequences are identified, a new phase is initiated to determine if the genes are associated to diseases. We use then the OMIM database to isolate gene-disease associations (Hamosh et al., 2005).

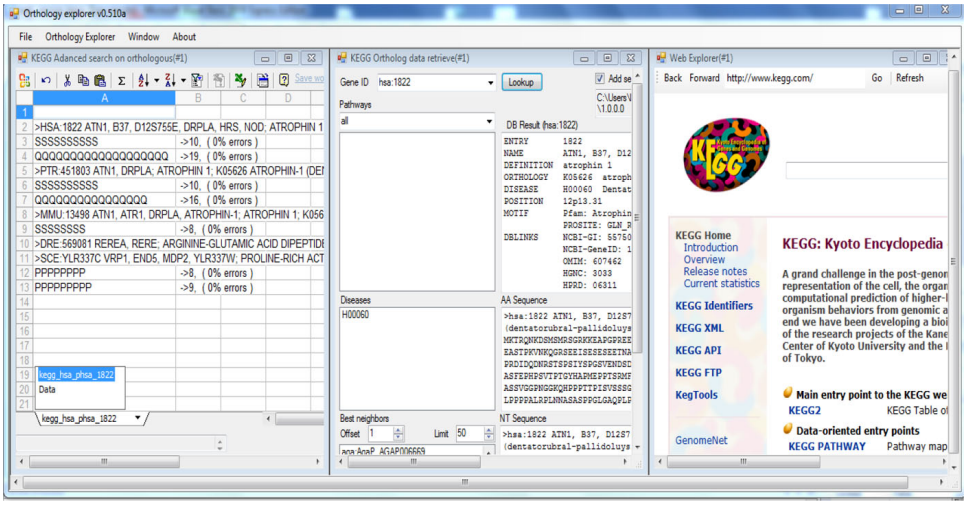
Using this information – genes associated to diseases and with repetitive sequences – the genes with repetitions responsible for diseases are isolated from the remaining genes in which the repetitions are not known to be related to diseases. The purpose of this separation is to create a control group of genes to validate the study. At the end, the results obtained from the test group can then be compared with the results from the control group.

For each gene in the set, the application looks for orthologous genes from a group of previously selected organisms. From that point, the process begins comparing the orthologous gene set creating a database of human genes and the respective orthologues found.

2.2 Implementation

The application was developed using the .NET platform integrated with Office Web Components. The use of KEGG web services, including integration with the .NET platform, required some modifications in the default parameters to avoid timeouts and transfer breaks. The user interface is depicted in Figure 2.

Figure 2 The application interface allows multiple windows enabling several simultaneous perspectives (see online version for colours)



The framework is made up of two main modules:

- "Orthologous data retrieval"
- "Orthologous advanced search".

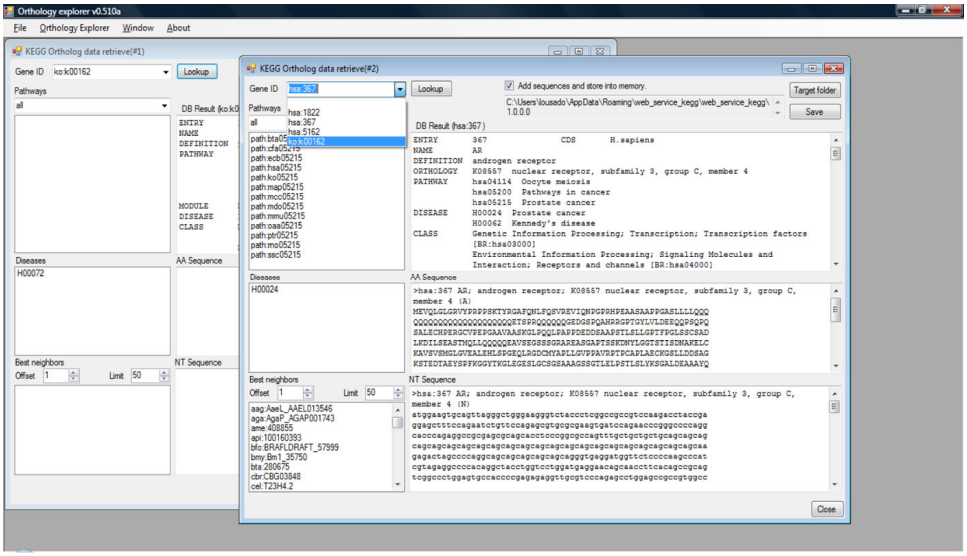
Additionally, the framework also incorporates a web explorer, pointing to KEGG web page by default. The user may create up to ten instances of each module (corresponding to ten separated frames), where each instance is able to access independent data. In order to manage these work frames, common visualisation features are available (tile, cascading, hide, copy, …) and all visualisation preferences can be kept along the several opened frames.

2.3 Data retrieval

Starting by a gene ID, a KEGG orthology identifier (ko) or a pathway, (e.g., hsa:367, ko:K08557 or path:hsa05215), the software allows to return the respective orthologous. The information collected is then displayed in the respective frame. The found orthologous genes are displayed on the left-hand list as well as the information on diseases related to that gene and pathways, if that is the case. The user may then access the orthologue by simply selecting the respective gene from the list on the left.

The data being viewed in the amino acid and nucleotide frames are held in memory by default. The user may save the data by simply selecting the respective option "Add sequences and store into memory" (Figure 3). By accessing the list of diseases, the window shows all available information including bibliographic references for each disease (Figure 3).

Figure 3 Two independent instances of KEGG orthologous data retrieval interface (see online version for colours)



2.4 Advanced search

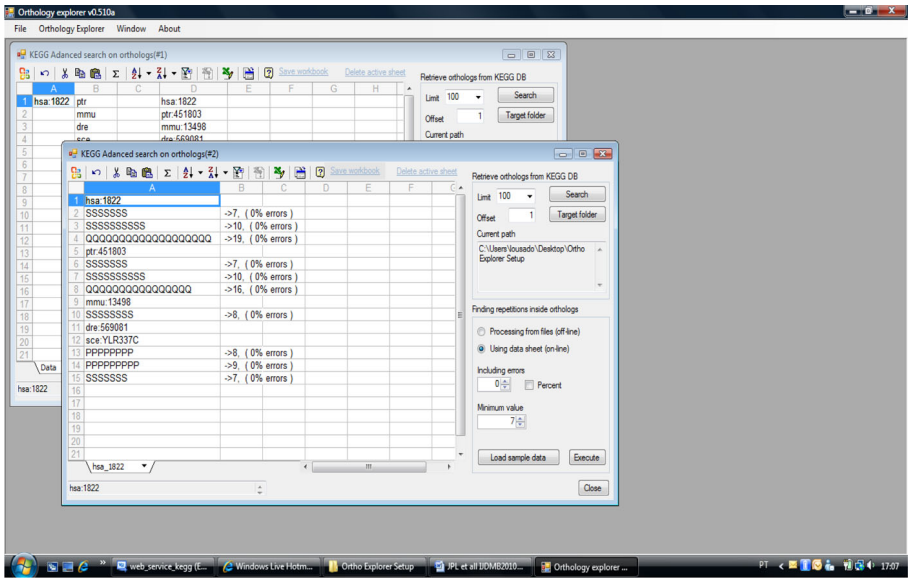
The module for orthologous advanced search is essentially an application of batch processing, that is, once the user creates the list of genes to be analysed and the list of organisms to be compared, the system will submit data to the KEGG database, automatically and iteratively. It extracts the information and saves the respective file in the folder that has been previously selected for that purpose.

As the data are being processed, a list of orthologous genes from the selected organisms is created. The gene order in the final list depends on the degree of similarity to the original sequence.

The tool also incorporates other functionalities, including the search for non-exact repetitions, i.e., some errors are allowed in the detection of tandem repeats. For the whole process, beside the web services connections, one can also resort on local data to conduct the analysis.

After processing, several spreadsheets are created with the results of the analysis on each of the orthologous genes of the set (Figure 4). These spreadsheets can alternatively be processed locally as a single file in XLS format.

Figure 4 Two independent instances of KEGG advanced search interface. The window behind refers to the filtering of orthologous genes. The window in front refers to the online search for repeats in orthologous genes (see online version for colours)



2.5 Web explorer

The application includes a web browser that gives complementary information about the genes under study and that allows filtering and parsing to retrieve more information from the gathered pages. We aim also to endow this module with annotation features so that more relevant concepts can be highlighted, facilitating the reading of the presented information.

3 Results

To test the application, we have conducted an analysis in accordance with the previously presented workflow (Figure 1). For this, we compared human genes under the referred conditions with the respective orthologous genes from several organisms.

Obtaining the results was almost immediate for the local source (offline), and it takes only a few seconds or minutes, depending on bandwidth and the amount of data, when we use the web server directly as a source of data (online). Since it integrates scattered data, whether via the Web (several sources) or by post-processing (offline files), the developed tool becomes a crucial ally for researchers, mainly in situations where massive data extraction is needed.

Table 1 presents a list of 15 human genes identified in the literature (Panzer et al., 1995; Hamosh et al., 2005; Pearson and Cleary, 2005) that have either codon or amino acid repetitions that are directly related with human diseases. This was our set of genes under study. Table 2 shows a list of 16 human genes with repeats, but with no described direct association with diseases. According to the workflow presented above, this list represents the control set for the presented study.

Table 1 List of human genes with amino acid repeats that are related to diseases

KEGG reference	*Gene reference*	*Amino acid*	*Start*	*Repeat count*
hsa:1822	ATN1	GLN/Q	1450	19
hsa:23054	NCOA6	GLN/Q	781	25
hsa:3064	HTT	GLN/Q	52	23
hsa:3209	HOXA13	ALA/A	112	14
hsa:3239	HOXD13	ALA/A	169	15
hsa:367	AR	GLN/Q	172	23
		GLY/G	1351	23
hsa:4287	ATXN3	GLN/Q	841	10
hsa:6310	ATXN1	GLN/Q	634	14
hsa:6311	ATXN2	GLN/Q	496	23
hsa:6314	ATXN7	GLN/Q	88	10
hsa:7546	ZIC2	ALA/A	1366	15
hsa:6908	TBP	GLN/Q	172	38
hsa:773	CACNA1A	HIS/H	6631	10
hsa:8202	NCOA3	GLN/Q	3730	29
hsa:860	RUNX2	GLN/Q	145	23
		ALA/A	217	17

Table 2 List of human genes with amino acid repeats that repeats are unrelated to diseases

KEGG reference	*Gene reference*	*Amino acid*	*Start*	*Repeat count*
hsa:4300	MLLT3	SER/S	445	42
hsa:93986	FOXP2	GLN/Q	454	40
hsa:84441	MAML2	GLN/Q	1762	34
hsa:4330	MN1	GLN/Q	1567	28
hsa:55589	BMP2K	GLN/Q	1378	27
hsa:84441	MAML2	GLN/Q	1915	27
hsa:9968	MED12	GLN/Q	6151	26
			6268	26
hsa:23524	SRRM2	SER/S	7594	25
hsa:64207	C14ORF4	GLN/Q	307	25
hsa:1602	DACH1	SER/S	418	24
hsa:51360	MBTPS2	SER/S	340	23
hsa:6595	SMARCA2	GLN/Q	646	23
hsa:84630	TTBK1	GLU/E	2245	23
hsa:27445	PCLO	PRO/P	7213	22
hsa:3778	KCNMA1	SER/S	115	22
hsa:55534	MAML3	GLN/Q	1456	21
			1885	18

In the next phase, we have selected 20 organisms randomly distributed along the evolutionary chain (Table 3). Then, using two separate instances of the "Advanced search" module (study and control), we have retrieved from the KEGG database all orthologous genes from the selected organisms.

Table 3 Organism list and phylogenetic classification.

Vertebrate	Mammals	Viviparous	Bos Taurus
			Canis familiaris
			Homo sapiens
			Mus musculus
			Pan troglodytes
		Marsupial	Monodelphis domestica
		Oviparous	Ornithorhynchus anatinus
		Bird	Gallus gallus
		Fish	Danio rerio
Invertebrate		Insect	Drosophila melanogaster
		Worm	#Caenorhabditis elegans
		Plant	Arabidopsis thaliana
		Fungus	Aspergillus fumigatus
			Kluyveromyces lactis
			Saccharomyces cerevisiae
			Schizosaccharomyces pombe
		Protozoan	Plasmodium falciparum
		Bacteria	Clostridium perfringens
			Mycobacterium tuberculosis

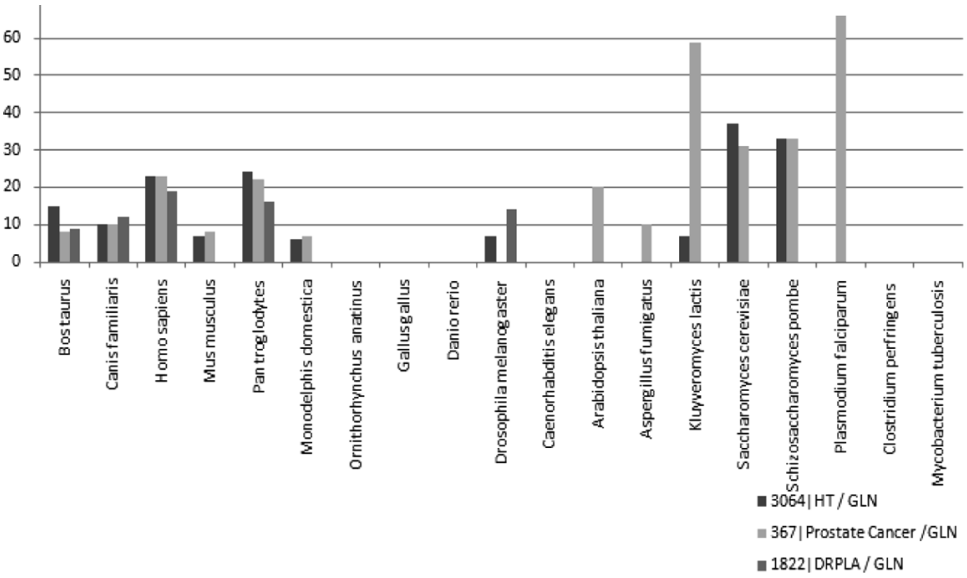
In order to extract the genes that are orthologous of those identified in Table 1 from the genomes of the organisms shown in Table 3, the application required approximately 15 min. This delay is due to the use web services and as such is always dependent on the amount of processing, data transfer and network efficiency. In this operation, 29 data files were created. Of these, 14 contained the orthologous genes, one file for each set of genes and their nucleotide sequences. Other 14 files contained identical data, but with their amino acid sequences. A log file was also created describing information about the success or failure in obtaining orthologues.

Figures 5 and 6 present two examples of post processing of the data using previous results. Figure 5 refers to genes with repeats that are responsible for diseases, whereas Figure 6, refers to the control gene set.

By comparing the results, we note that the Gln string (23 repeats) of the human gene 367, which is responsible for the emergence of prostate cancer, is also detectable in organisms rather distant on the evolutionary chain, namely in fungi. However, in intermediate species the string is not present and it appears again only in higher organisms (mammals). The same happens with the human gene 3064, responsible for

Huntington's disease (Herishanu et al., 2009). The Gln string from gene 1822, that is responsible for dentatorubral-pallidoluysian atrophy (DRPLA), a severe neuro-degenerative disease (Pearson, 2007), turns out to be only present in Drosophila melanogaster, as well as in most higher organisms.

Figure 5 The graph represents genes whose repetitions were found to be associated with human diseases. It presents a comparison between the repetition length from three human genes and their retrieved orthologous genes



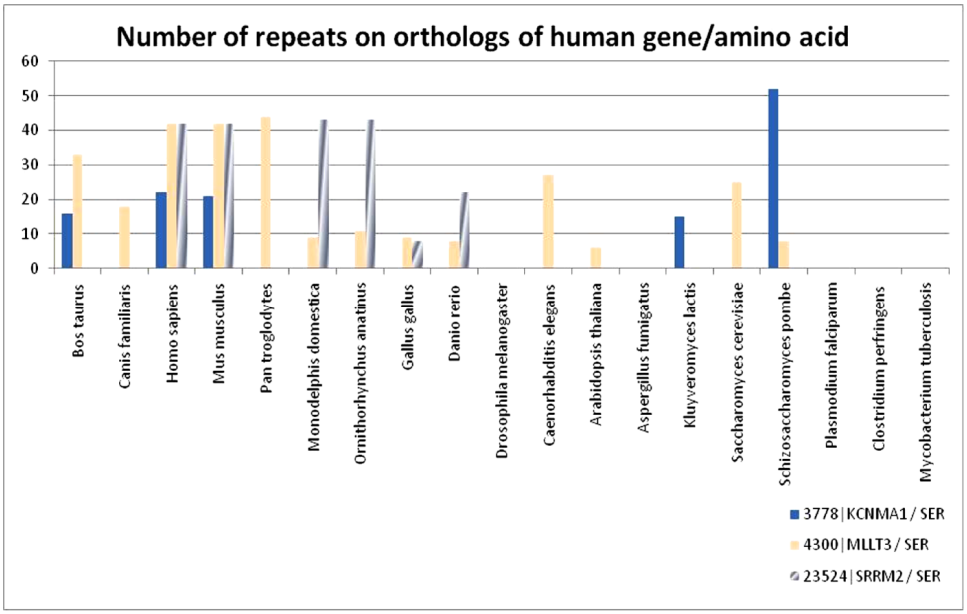
As for the control group, shown in Figure 6, we can observe, for instance, that the gene KCNMA1, responsible for human neurodegenerative diseases such as epilepsy and paroxysmal dyskinesis (Du et al., 2005), presents a great number of repetitions of Ser residues in *Schizosaccharomyces pombe* and *Kluyveromyces lactis*, two fungi, presenting no significant repetitions in intermediate species. The repetition pattern appears again in higher organisms. Interestingly, repetitions are not present in the organism Pan troglodytes, which is phylogenetically much close to the humans. Furthermore, repetitions of the human gene SRRM2, which is related with Parkinson's disease (Shehadeh et al., 2010), are present in *Homo sapiens*, *Mus musculus*, *Monodelphis domestica* and *Ornithorhynchus anatinus*, but do not appear in Pan troglodytes.

From the results, we can observe that the genes in Figure 5 have a similar behaviour in *Pan troglodytes* and *Homo sapiens*, which is not the case in the control group, with exception of the MLLT3 gene (Figure 6). It is perhaps interesting to note that for the genes in the test group the largest number of repeats occurs with Gln residues while in the control group, the largest number of repetitions occurs with Ser residues, suggesting a more deleterious effect occurring from Gln repetitions than from Ser ones.

With these results we cannot argue a clear distinction between the two groups and conclude that the repetitions responsible for diseases in humans have a different evolution from the repetitions that are not responsible for diseases. However, we cannot also assure that the disease associated to a particular gene in the control set is not caused by codon repetitions – simply there is yet no scientific evidence about that relation. Anyway, the

objective of this study was not to explore the biological meaning of the sample, but only to show the potential of the application regarding the integration of information for studying codon/amino acid tandem repeats in orthologous genes.

Figure 6 The graph represents genes whose repetitions could not be linked with diseases. It provides a comparison between the repetition length from three human genes and their retrieved orthologous genes (see online version for colours)



4 Conclusion

Codon or amino acid repeats have been associated with specific human diseases, and may play a variety of regulatory and evolutionary roles.

In this paper we presented a computation application that simplifies the study of genes with this type of pattern, along the evolutionary chain. To do so, the software extracts orthologous genes from public resources and performs a comparative analysis that shows how repeats have evolved over time within the species under study.

Using this methodology, shown some differences that occurred during the evolution process between orthologous genes of a group of genes. This evolution was not uniform and shows some differences between genes from the test and control groups. Future work should deal with extensive exploitation of these two groups, extending the study to other genes and other organisms, so that the full set of capabilities of the application can be explored.

Acknowledgements

J.P. Lousado was funded by the Instituto Politécnico de Viseu under the PROFAD Programme. G. Moura was funded by the FCT grant PTDC/BIA-BCM/72251/2006.

References

Ali, S., Ansari, S., Ehtesham, N.Z., Azfer, M.A., Homkar, U., Gopal, R. and Hasnain, S.E. (1998) 'Analysis of the evolutionarily conserved repeat motifs in the genome of the highly endangered central Indian swamp deer cervus duvauceli branderi', *GENE*, Vol. 223, Nos. 1–2, pp.361–367.

Bogaerts, V., Theuns, J. and Van Broeckhoven, C. (2008) 'Genetic findings in Parkinson's disease and translation into treatment: A leading role for mitochondria?', *Genes Brain Behav*., Vol. 7, No. 2, pp.129–151.

Bowen, T.A., Guy, C.A.A., Cardno, A.G.A.B., Vincent, J.B.C., Kennedy, J.L.C., Jones, L.A.A., Gray, M.A., Sanders, R.D.A., Mccarthy, G.A., Murphy, K.C.A., Owen, M.J.A.B. and O'donovan, M.C.A. (2000) 'Repeat sizes at cag/ctg loci ctg18.1, erda1 and tgc13-7a in schizophrenia', *Psychiatric Genetics*, Vol. 10, No. 1, pp.33–37.

Brameier, M. and Wiuf, C. (2007) 'Ab initio identification of human microrna based on structure motifs', *BMC Bioinformatics*, Vol. 8, No. 1, p.478.

Du, W., Bautista, J., Yang, H., Diez-Sampedro, A., You, S., Wang, L., Kotagal, P., Lüders, H., Shi, J. and Cui, J. (2005) 'Calcium-sensitive potassium channelopathy in human epilepsy and paroxysmal movement disorder', *Nature Genetics*, Vol. 37, No. 7, pp.733–738.

Ferro, P., Catalano, M.G., Dell'eva, R., Fortunati, N. and Pfeffer, U. (2002) 'The androgen receptor cag repeat: a modifier of carcinogenesis?', *Molecular and Cellular Endocrinology*, Vol. 193, Nos. 1–2, pp.109–120.

Freed, K.A., Cooper, D.W., Brennecke, S.P. and Moses, E.K. (2005) 'Detection of cag repeats in pre-eclampsia/eclampsia using the repeat expansion detection method', *Mol. Hum. Reprod.*, Vol. 11, No. 7, pp.481–487.

Fu, Z. and Jiang, T. (2008) 'Clustering of main orthologs for multiple genomes', *J. Bioinform Comput Biol*., Vol. 6, No. 3, pp.573–584.

George, R.A., Liu, J.Y., Feng, L.L., Bryson-Richardson, R.J., Fatkin, D. and Wouters, M.A. (2006) 'Analysis of protein sequence and interaction data for candidate disease gene prediction', *Nucl. Acids Res.*, Vol. 34, No. 19, p.e130.

Hamosh, A., Scott, A., Amberger, J., Bocchini, C. and Mckusick, V. (2005) 'Online Mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders', *Nucleic Acids Research*, Vol. 33, No. Database Issue, p.D514.

Herishanu, Y.O., Parvari, R., Pollack, Y., Shelef, I., Marom, B., Martino, T., Cannella, M. and Squitieri, F. (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*, Vol. 277, Nos. 1–2, pp.143–146.

Hsueh, W. (2006) 'Genetic discoveries as the basis of personalized therapy: Rosiglitazone treatment of Alzheimer's disease', *Pharmacogenomics J*., Vol. 6, No. 4, pp.222–224.

Jones, N.C. and Pevzner, P.A. (2006) 'Comparative genomics reveals unusually long motifs in mammalian genomes', *Bioinformatics*, Vol. 22, No. 14, pp.e236–e242.

KEGG (2010) *KEGG: Kyoto Encyclopedia of Genes and Genomes*, Obtained through the internet: http://www.kegg.com (accessed 20-03-2010)

Lousado, J., Oliveira, J., Moura, G. and Santos, M. (2009) 'Analysing the evolution of repetitive strands in genomes', *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*, Springer Berlin/Heidelberg, pp.1047–1054.

Mena, M.A., Rodriguez-Navarro, J.A., Ros, R. and De Yebenes, J.G. (2008) 'On the pathogenesis and neuroprotective treatment of Parkinson disease: What have we learned from the genetic forms of this disease?', *Curr. Med. Chem*., Vol. 15, No. 23, pp.2305–2320.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999) 'Kegg: Kyoto encyclopedia of genes and genomes', *Nucleic Acids Research*, Vol. 27, No. 1, p.29.

OMIM (2009) *OMIM, Online Mendelian Inheritance in Man*, Obtained through the internet: http://www.ncbi.nlm.nih.gov/omim/).

Panzer, S., Kuhl, D.P. and Caskey, C.T. (1995) 'Unstable triplet repeat sequences: a source of cancer mutations?', *Stem Cells*, Vol. 13, No. 2, pp.146–157.

Pearson (2007) *Repeat Disease Database*, Obtained through the internet: http://www.cepearsonlab.com/rdd.php (accessed 9-02-2009).

Pearson, C.E.N.E.K. and Cleary J.D. (2005) 'Repeat instability: mechanisms of dynamic mutations', *Nat Rev Genet.*, Vol. 6, No. 10, pp.729–742.

Pestova, T.V., Hellen, C.U. and Wimmer, E. (1994) 'A conserved aug triplet in the 5' nontranslated region of poliovirus can function as an initiation codon in vitro and in vivo', *Virology*, Vol. 204, No. 2, pp.729–737.

Shehadeh, L., Yu, K., Wang, L., Guevara, A., Singer, C., Vance, J. and Papapetropoulos, S. (2010) 'Srrm2, a potential blood biomarker revealing high alternative splicing in Parkinson's disease', *PLoS One*, Vol. 5, No. 2, p.e9104.

Tarini, B.A., Singer, D., Clark, S.J. and Davis, M.M. (2009) 'Parents interest in predictive genetic testing for their children when a disease has no treatment', *Pediatrics*, Vol. 124, No. 3, pp.e432–e438.