Proceedings of the 2007 IEEE Symposium on Computational
Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Comparison of Human and Mouse Pseudogenes

Pooja[1] and Jagath C. Rajapakse[1,2,3], *Senior Member, IEEE,*
[1]BioInformatics Research Centre, School of Computer Engineering, Nanyang Technological University, Singapore
[2] Singapore-MIT Alliance, Singapore
[3] Biological Engineering Division, Massachesetts Institute of Technology, USA
Email: asjagath@ntu.edu.sg

Abstract— **Pseudogenes are formed by either gene duplication or retrotransposition and yet unknown to express any RNA or produce any protein, which may be due to some defects in their structure. Because of the non-functional nature, pseudogenes are considered important resources in the study of evolutionary history and phylogenetic comparison of genomes. Psuedogenes whose structure is similar to the normal genes, pose problems in gene annotation and interfere with PCR or hybridization experiments. In this paper, we compare structural and functional properties of pseudogenes of human and mouse.**

It was found that 3,277 pseudogenes of humans which had conserved regions in pseudogenes of mouse shared the same number of chromosomes. Human and mouse pseudogenes are very similar to each other based on the effective codon usage and fraction of codons having guanine or cytosine at the third codon position. However, the proportion of GC content and lengths of base pairs were different. The parent genes or proteins which have more number of pseudogenes may be considered to be evolving more quickly showing more variability. Further ribosomal proteins, binding proteins and receptors had more number of pseudogenes than the other proteins.

Index Terms—**pseudogenes, effective number of codons, comparative genomics, orthologues.**

I. Introduction

Within the human genome, a little is known of non-coding DNA sequences. If these sequences do not have any function, then the questions arise why these are evolutionarily preserved and why pseudogenes (the faulty replicates of normal genes) occur in a genome. Despite major advances in the field of bioinformatics, these questions remain unanswered. Comparative genomics compares the genomes of one organism with one or more other organisms. It helps us investigate how the overall structure of genes and genomes has evolved and how these findings can be related to gene expression and gene regulation. This new and emerging field of research can be used to study the pseudogenes, also known as non-functional genes. Pseudogenes look very similar to the genes but with some defects in structure, that prevents them from functioning normally like genes. They are created either by reverse transcription of mRNAs of functional genes by long interspersed nuclear element 1 (LINE1) giving rise to processed pseudogenes or by gene duplication giving rise to unprocessed pseudogenes [1][2][3]. Bescause of their close similarity to the functional genes, pseudogenes are considered important resources in the study of evolutionary history and phylogenetic comparison of genomes and are hence considered as 'genomic fossils' [4][5][6][7].

Comparison of fruit fly and human genomes revealed that about 60 percent of genes are conserved between fruit fly and humans and two-thirds of human genes known to be involved in cancer have their counterparts in fruit fly [8]. Yet another surprising finding is that when a human gene associated with early-onset of Parkinson's disease (PD) is inserted, fruit flies display symptoms similar to those of PD patients [8]. This has raised the possibility of making use of these fruit flies as a model for testing therapies aimed at curing PD. Comparison of genomes of four different species of baker's yeast (S. *cerevisiae*, S. *paradoxus*, S. *mikatae* and S. *bayanus*) provided a good understanding of yeast genome [9]. The results enabled researchers to filter out the "noise" and focus on the real "signals" in genomes and revealed that only five percent of the human genome is functional. The above research assists in the extraction of important biological signals hidden in the noise of vast non-functional regions. Clinical response of dogs is often similar to humans; Ostrander et al. [10] have discussed the benefits of comparing the canine genome to the human genome and suggested the mapping and cloning of a gene for inherited narcolepsy in a colony of Doberman pinschers. One of the rare forms of narcolepsy was caused by mutation in the hypocretin-2 receptor gene [11]. Sleep disorders are very common in humans and hypocretin deficiency is associated with most cases of narcolepsy in humans. The narcolepsy gene is localized to a region of canine chromosome 12 and this corresponds to a gene poor region of human 6p21.

Zhang et al. [12] found that the number of processed pseudogenes in mouse is only half of that of humans and estimated that $\sim 60\%$ are lineage specific, created after the human and mouse diverged. They also concluded that the age distribution of processed pseudogenes of mouse closely resembles long interspersed repeats (LINEs) while that of humans resembles short interspersed repeats (SINEs). Human genes which have multiple retrotransposed pseudogenes seem to have homologues in mouse with multiple pseudogenes [13].

In this paper, we perform a comparative genomics study of pseudogenes of human and mouse, based on their structural and functional properties. Because of their similarity to genes, pseudogenes implicate gene annotation [14][15] and interfere with PCR or hybridization experiments of the genes [16][17]. Hence, understanding the structure help us differentiate them from their parent genes. Furthermore, this help understand the evolution of genes and the correseponding pseudogenes, and create phylogeny profiles of the genomes. The human

1-4244-0710-9/07/$20.00 ©2007 IEEE

genome is considered to be very close to the mouse genome [18], and mice are often used as useful models for studying human disease. Scientists can perform experiments on mice, not ethical on humans; the small size and short generation time of mice make it practical for scientists to experiment with [19].

Our comparison of the pseudogenes of human and mouse showed that the effective codon usage and fraction of codons having guanine or cytosine at their third codon position is similar in the pseudogenes of human and mouse. However, there was a significant difference in the proportion of GC content and length of pseudogenes of human and mouse. We analyzed functions of orthologues of the parents genes of human and mouse pseudogenes. Orthologues are a group of proteins having similar functions in different organisms. It is found that orthologous parent genes of psedogenes tend to code for ribosomal proteins, binding proteins, and receptors.

II. Structural Comparisons

Genome annotation is defined as the process of attaching biological information to sequences and could be in terms of structural or functional annotation. Structural annotation deals with identifying the genomic elements such as gene structure, coding regions, number of exons, number of introns, open reading frames (ORF), and their localizations.

A. Number of base pairs

The number of base pairs (bp) is often used as a measure of length of a DNA segment. The human genome consists of around 3.2 billion base pairs along forty-six DNA molecules contained in twenty-three pairs of chromosomes [20]. While, on the other hand the mouse genome consists of around 2.7 billion base pairs along forty DNA molecules contained in twenty pairs of chromosomes [18]. Genes can vary in size but their average length is about 3,000 base pairs in length. Human DNA contains an estimated 30,000 genes. Therefore, genetic or coding DNA makes up less than 5% of total human DNA and the remainder of the DNA is non-coding DNA [21].

B. Number of exons

Exon is defined as a region of a gene that contains instructions for making a protein. The exons are separated by introns in most genes; the intervening segments of DNA are removed during splicing. The maximum number of exons per gene for humans is 312 (TTN coding gene) and for mouse is 146 (NEB coding gene). Intronless genes are also found in human and mouse chromosomes [22].

C. GC content

GC content is a characteristic of a genome sequence. GC pairs in DNA are connected with three hydrogen bonds while the AT pairs are connected by two. GC pair is stronger and more resistant to denaturation by high temperatures as compared to AT pair. Knowing this statistic is important because genes in some organisms tend to be in more GC rich regions of the genome. Hence, understanding the GC distribution helps in assessing the gene potential of a region. GC content is defined as the proportion of GC base pairs in the DNA molecule or genomic sequence.

$$GC = \frac{\text{number of G or C nucleotides in the sequence}}{\text{total number of nucleotides in the sequence}} \quad (1)$$

The average GC content of human genome is 41%, while that of mouse is 42%. The distribution is tighter in mouse with lesser deviations. In case of human genome, the deviations can range from 33% to 56% [18].

D. Effective number of codons (N_c)

Much of the genetic code is degenerate, i.e., most of the amino acids can be encoded by more than one codon (triplet of nucleotides). In bacteria such as Escherichia *coli*, highly expressed genes tend to selectively use codons recognized by the most abundant tRNA species [23]. This similarity between codon usage bias (the probability that a given codon will be used to code for an amino acid over a different codon which codes for the same amino acid) and gene expression level is also observed in bacteria like Clostridium *perfringens* and Haemophilus *influenzae* [24][25][26]. Highly expressed genes, such as those encoding ribosomal proteins and histones in S. *pombe*, C. *elegans*, and D. *melanogaster*, have different biased patterns of codon usage [28]. However, in X. *laevis* and H. *sapiens*, codon usage in the genes encoding ribosomal proteins and histones were not significantly biased. This suggested that the primary factor influencing codon usage diversity in these species was not translation efficiency.

The effective number of codons N_c is a measure of overall codon bias and is similar to the effective number of alleles measure used in population genetics. Hence, the reported value of N_c varies between 20 and 61. Twenty when only one codon is effectively used for each amino acid and 61 when codons are used randomly. If there are no amino acids in a synonymous family or if the gene is too short or has extremely skewed amino acid usage, then N_c is not calculated.

E. GC_3 content of pseudogenes

Codons coding for same amino acid are synonymous and usually differ by one nucleotide in the third position. The different synonymous codons are used at different frequencies at different genomes [29]. In prokaryotes, a few theories have been suggested to account for this variation in synonymous codon usage. In bacteria such as Mycoplasma *genitalium*, the primary source of variation seems to be related to the use of GC-ending codons [30][31]. While, in bacteria such as Borrelia *burgdorferi* [32] and Treponema *pallidum* [33], there is a base usage skew between the leading and lagging strands of replication. Genes on the leading strand are seen to preferentially use GT-ending codons.

GC_3 is defined as the fraction of codons that are synonymous at the third codon position, which is that they have either a guanine or cytosine at their third codon position. It can be calculated as

$$GC_3 = \frac{\text{number of codons with G or C at 3rd position}}{\text{total number of codons}} \quad (2)$$

F. Conserved Regions

Pseudogenes of human and mouse can also be compared by comparing their sequences and looking for the conserved regions. These findings will further help in investigating their evolutionary descent and functions of the conserved regions of the pseudogenes which are presently considered as ”junk” pieces of DNA [34].

III. Functional Comparisons

Functional annotation deals with attaching biological information to genomic elements, including the biochemical and biological functions, genomic expression and interactions, etc. Pseudogenes are similar to normal genes but do not express any RNA or protein. They can be described as the crippled copies of known functional genes like LINEs and SINEs [35][36]. In both mouse and human genomes, similar types of genes give rise to many processed pseudogenes (also known as retropseudogenes). These functional genes tend to be housekeeping genes, which are seen to be highly expressed in the germ line. In particular, the ribosomal-protein genes form the largest sub-group [12].

The parent genes of the pseudogenes are functional genes and code for proteins. A protein which has more number of pseudogenes are evolving faster than the proteins with less or zero number of pseudogenes [12]. We compare the protein products of the parent genes of the pseudogenes of human and mouse and find their orthologues. This is achieved using the program CD-HIT [37][38] which produces a set of ’non-redundant’ sequences as output from a given set of sequences. Besides, this program also outputs a ’cluster’ file which contains sequence groups for each ’non-redundant’ sequence representative. We use the CD-HIT program to produce a set of closely related protein families of human and mouse genes from a given set of sequences of human and mouse pseudogenes.

IV. Experiments and Results

The pseudogene sequences for human and mouse genomes were obtained from a pseudogene resource [39]. The data contained 7868 known pseudogenes of human and 4476 known pseudogenes of mouse.

A. Structural comparisons

1) Number of base pairs: Since mouse genome is about 14% smaller than human genome [40] and the numbers of pseudogenes of human and mouse differ, the distribution of number of base pairs in each pseudogene in mouse is normalized by 1000bp (Figure 1). Mean length of human and mouse pseudogenes were $(\mu_l)_{\mathrm{human}}$ = 743.6bp and $(\mu_l)_{\mathrm{mouse}}$ = 726.8bp. *Student’s t distribution* was used to test the significance of the result at significance level $\alpha = 0.05$. There was a significant difference in length of pseudogenes of human and mouse ($p < 0.05$).

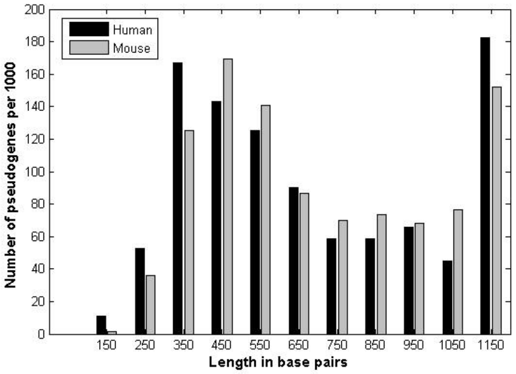

Fig. 1. Distribution of lengths of human and mouse pseudogenes (normalized by 1000 pseudogenes in human)

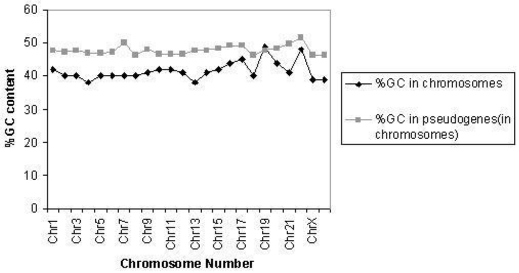

Fig. 2. The distribution of %GC content in human chromosomes and pseudogenes in respective chromosomes

2) Number of Exons: Numbers of exons of orthologues parent genes of human and mouse pseudogenes (orthologues of human and mouse pseudogenes were found in Sec III B) were calculated. It was observed that the mean number of exons per gene is 7.3 and 95.8% of orthologous genes had less than 15 exons per gene and 0.19% of orthologous genes were intronless genes.

3) GC content: The mean GC contents of human and mouse pseudogenes were $(\mu_{\mathrm{GC}})_{\mathrm{human}}$ = 47.569% and $(\mu_{\mathrm{GC}})_{\mathrm{mouse}}$ = 46.819%. GC content of human chromosomes compared to those of pseudogenes belonging to the respective chromosomes are shown in Figure 2. The GC content of human chromosomes was obtained from the dataset of Venter et al. [41]. As seen, GC contents were higher in pseudogenes, indicating their resemblence to genes. Figure 3 shows %GC content in pseudogenes of human and mouse (at the chromosome level, i.e., GC content of pseudogenes belonging to respective chromosomes). GC contents of human and mouse pseudogenes at chromosome level were significantly different ($p < 0.05$).

4) Effective number of codons: We use the CodonW package by John Peden for codon usage analysis [42] to obtain the effective number of codons (N_c) of sequences. The normalized

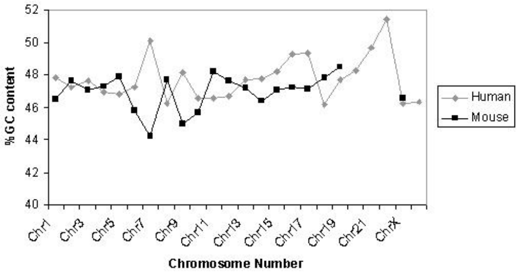

Fig. 3. The distribution of %GC content in pseudogenes of human and mouse at chromosome level

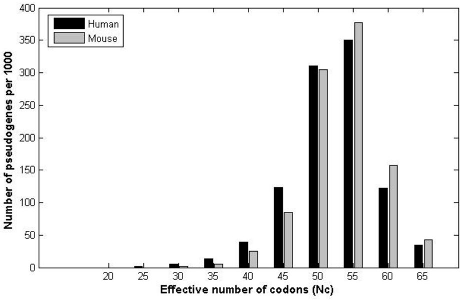

Fig. 4. Distribution of the number of codons of human and mouse pseudogenes (the number in humans were normalized per 1000 pseudogenes)

plot of the distribution of N_c is shown in Figure 4. Mean N_c values of human and mouse pseudogenes were $(\mu_c)_{\text{human}}$ = 49.59 and $(\mu_c)_{\text{mouse}}$ = 50.74 respectively. There was no significant difference in the effective number of codons among human and mouse pseudogenes.

5) GC_3 content of the pseudogenes: We obtained the GC_3 values of human and mouse pseudogenes from the CodonW package by John Peden [42]. The normalized plot is illustrated in Figure 5. The mean GC_3 values of human and mouse pseudogenes were $(\mu_3)_{human}$ = 0.477 and $(\mu_3)_{mouse}$ = 0.469 respectively. There is no difference in GC_3 contents of human and mouse pseudogenes at $p < 0.05$.

6) Conserved regions: We used Blastz [44] to compare the pseudogenes of human and mouse and find regions which were 70% conserved and had at least 100bp length [45]. 127,474 regions were seen to be conserved. These regions were further analyzed to study the synteny. 3,277 pseudogenes of human which had conserved regions in pseudogenes of mouse shared the same chromosome number.

B. Functional comparisons

Protein sequences of corresponding functional (parent) genes of pseudogenes of human and mouse were obtained from SwissProt protein database [46]. The paralogues (similar proteins in the same genome) in human and mouse protein sequences were removed at 90% sequence identity using the clustering program CD-HIT [37][38]. This program was further used to find the clusters of orthologous protein sequences in human and mouse at 70% sequence identity.

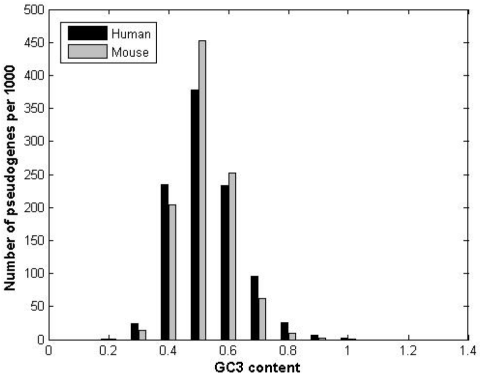

Fig. 5. The distribution of GC_3 contents of human and mouse pseudogenes (normalized per 1000 pseudogenes)

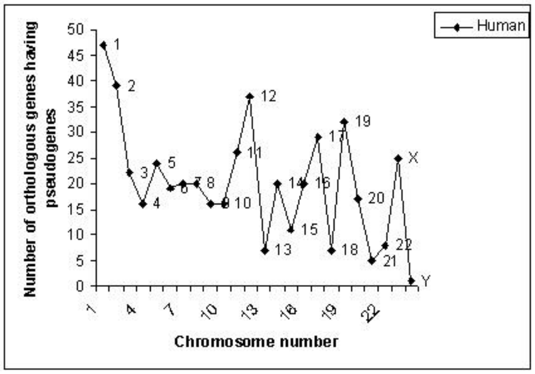

Fig. 6. Distribution of orthologous parent genes of pseudogenes of human and mouse in human chromosomes

Out of 3737 protein sequences of human and mouse, produced by parent genes of psuedogenes, there were 546 human and mouse orthologous proteins at 70% sequence identity. 34 protein sequences had paralogues in human genome and 36 protein sequences had paralogues in mouse genome. The 546 human and mouse orthologous protein sequences were further analyzed based on their functions and the notable propotions among them belonged to ribosomal proteins (14.8%), binding proteins (12.3%), receptors (2.7%), transferase (2%) and oxidoreductase (1.8%). The distribution of these 546 protein sequences in human chromosome is given in Figure 6. Chromosome 1 has maximum number of orthologous protein sequences and chromosome Y has the minimum number of orthologous protein sequences.

V. Discussion and Conclusion

Mouse genome is remarkably similar to human genome not only in the structure at the level of chromosomes but also at DNA sequence level [47]. In this work, we compared human and mouse pseudogenes which were found to be very similar to each other on effective codon usage and fraction of codons having guanine or cytosine at their third codon position. However, they had different proportion of GC content and lengths of base pairs. Human pseudogenes of respective chromosomes have higher GC contents than those of the respective chromosomes. Pseudogenes are very similar to the coding genes. Since the genes are rich in GC content [41], hence the higher GC content of pseudogenes can be explained.

Pseudogenes are very similar to the genes and interfere with PCR or hybridization experiments that are intended for the genes [16][17]. Therefore, identification of pseudogenes are important and our future study will focus on comparing structural compositions of genes, the pseudogenes and their parent genes. Pseudogenes are presently considered as ”junk” DNA [34]. However, there must be some reason behind the conservation of these ”junk” DNA over millions of years of evolution. The comparative genomic study of human and mouse pseudogenes can be further explored to find possible functions of the pseudogenes.

The ribosomal proteins, binding proteins, and receptors were the common proteins in human and mouse which are evolving more quickly than other proteins because they have higher number of pseudogenes. These finding could help in evolutionary and phylogenetic study of the two genomes. By including more number of species, we can obtian a clear understanding of not only the creation or possible functions of pseudogenes, but also the evolution of various genomes and species divergence.

References

[1] Maestre, J. et al. (1995) mRNA retroposition in human cells: processed pseudogene formation. *EMBO J.* 14, 6333-6338

[2] Feng, Q. et al. (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* 87, 905-916

[3] Esnault, C. et al. (2000) Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet*. 24, 363-367

[4] Petrov, D.A. et al. (1996) High intrinsic rate of DNA loss in Drosophila. *Nature* 384, 346-349

[5] Petrov, D.A. et al. (2000) Evidence for DNA loss as a determinant of genome size. *Science* 287, 1060-1062 7

[6] Zhang, Z. and Gerstein, M. (2003), The human genome has 49 cytochrome c pseudogenes, including a relic of a primordial gene that still functions in mouse. *Gene* 312, 61-72

[7] Zhang, Z. and Gerstein, M. (2003) Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res.* 31, 5338-5348

[8] Barbara J. Culliton, The Humanized Fly, *Genome News Network*, March 2000

[9] Manolis Kellis (Kamvysselis) et al., Yeast Comparitive Genomics, Whitehead Genome Center

[10] Elaine A. Ostrander, Kerstin Lindblad-Toh, Eric S. Lander, Fred Hutchinson Cancer Research Center, Whitehead/MIT Center for Genome Research

[11] Lin L, Faraco J, Li R, Kadotani H, Rogers W, Lin X, Qiu X, de Jong PJ, Nishino S, Mignot, E. The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene. *Cell* 98: 365-376, 1999

[12] Zhaolei Zhang, Nick Carriero and Mark Gerstein, Comparative analysis of processed pseudogenes in the mouse and human genomes. *TRENDS in Genetics* Vol.20 No.2 February 2004

[13] Zhang and Mark Gerstein, Large-scale analysis of pseudogenes in the human genome, 2004 *Current Opinion in Genetics & Development* 14:328335

[14] van Baren MJ, Brent MR: Iterative gene prediction and pseudogene removal improves genome annotation. *Genome Res* 2006, 16:678-685.

[15] Brent MR, Guigo R: Recent advances in gene structure prediction. *Curr Opin Struct Biol* 2004, 14:264-272.

[16] Guo, N. et al. (1998) The human ortholog of rhesus mannose-binding protein-A gene is an expressed pseudogene that localizes to chromosome 10. *Mamm. Genome* 9, 246-249

[17] Ruud, P. et al. (1999) Identification of a novel cytokeratin 19 pseudogene that may interfere with reverse transcriptase-polymerase chain reaction assays used to detect micrometastatic tumor cells. *Int. J. Cancer* 80, 119-125

[18] Robert H. Waterston et al., Initial sequencing and comparative analysis of the mouse genome, *Nature* 420, 520-562 (5 December 2002)

[19] Mouse Genetics Core Project, Washington University, St. Louis http://mgc.wustl.edu

[20] J. Craig Venter et al., The sequence of the Human Genome, *Science*, Vol 291, (16 February 2001)

[21] DNA Profiling Techniques, http://www.crimtrac.gov.au/dnatechniques.htm

[22] Meena Kishore Sakharkar, Bagavathi S. Perumal, Kishore R. Sakharkhar and Pandjassarame Kangueane, An analysis on gene architecture in human and mouse genomes, *In Silico Biology* 5, 0032 (2005), Bioinformation Systems e.V

[23] T. Ikemura, Codon usage and tRNA content in unicellular and multicellular organisms, *Mol. Biol. Evol.* 2 (1985), pp. 13-34

[24] P.M. Sharp, E. Bailes, R.J. Grocock, J.F. Peden and R.E. Sockett, Variation in the strength of selected codon usage bias among bacteria, *Nucleic Acids Res.* 33 (2005), pp. 1141-1153.

[25] H. Musto, H. Romero and A. Zavala, Translational selection is operative for synonymous codon usage in Clostridium perfringens and Clostridium acetobutylicum, *Microbiology* 149 (2003), pp. 855-863

[26] G. Perriere and J. Thioulouse, On-line tools for sequence retrieval and multivariate statistics in molecular biology, *Comput. Appl. Biosci.* 12 (1996), pp. 63-69

[27] Haruo Suzuki, Rintaro Saito and Masaru Tomita , A problem in multivariate analysis of codon usage data and a possible solution, *FEBS Letters* 579 (2005) 6499-6504

[28] Shigehiko Kanaya, Yuko Yamada, Makoto Kinouchi, Yoshihiro Kudo, Toshimichi Ikemura, Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis, *J Mol Evol* (2001) 53:290298

[29] P.M. Sharp, E. Cowe, D.G. Higgins, D.C. Shields, K.H. Wolfe and F. Wright, Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity, *Nucleic Acids Res* 16 (1988), pp. 8207-8211.

[30] J.O. McInerney, Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns, *Microbial Comp. Genomics 2* (1997), pp. 89-97.

[31] A.R. Kerr, J.F. Peden and P.M. Sharp, Systematic base composition variation around the genome of Mycoplasma genitalium, but not Mycoplasma pneumoniae, *Mol. Microbiol.* 25 (1997), pp. 1177-1179

[32] J.O. McInerney, Replicational and transcriptional selection on codon usage in Borrelia burgdorferi, *Proc. Natl. Acad. Sci.* USA 95 (1998), pp. 10698-10703

[33] B. Lafay, A.T. Lloyd, M.J. McLean, K.M. Devine, P.M. Sharp and K.H. Wolfe, Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases, *Nucleic Acids Res* 27 (1999), pp. 1642-1649.

[34] Gibbs W.W. (2003) ”The unseen genome: gems among the junk”, *Scientific American*, 289(5): 46-53. (A review, written for non-specialists, of recent discoveries of function within junk DNA.)

[35] Woodmorappe, J., Are pseudogenes ’shared’ mistakes between primate genomes? *CEN Tech. J.* 14(3): 55-71, 2000.

[36] Walkup, L.K., Junk DNA, *CEN Tech. J.* 14(2):18-30, 2000.

[37] Weizhong Li, Lukasz Jaroszewski and Adam Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* (2001) 17:282-283

[38] Weizhong Li, Lukasz Jaroszewski and Adam Godzik. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* (2002) 18: 77-82

[39] Pseudogene Resouce, http://www.pseudogene.org

[40] Kate Dalke, Mouse in the House, *Genome News Network*, Dec 2002

[41] J. Craig Venter et al., The sequence of the Human Genome, *Science*, Vol 291, (16 February 2001)

[42] John Peden, Correspondance Analysis of Codon Usage, *SourceForge*, 2005 http://codonw.sourceforge.net/culong.html

[43] Project Ensembl, http://www.ensembl.org

[44] Scott Schwartz, W. James Kent, Arian Smit, Zheng Zhang, Robert Baertsch, Ross C. Hardison, David, Haussler and Webb Miller, Human-Mouse Alignments with BLASTZ, *Genome Research*, Vol 13:103-107, 2003

[45] Giardine B, Elnitski L, Riemer C, Makalowska L, Schwartz S, Miller W (2003), GALA, a Database for Genomic Sequence Alignments and Annotations, *Genome Res*. 13: 732-741

[46] Swiss-Prot Protein Knowledgebase, TrEMBL, Computer Annotated supplement to Swiss-Prot, http://www.expasy.org/sprot/sprot-retrieve-list.html

[47] Edward R. Winstead, Humans and Mice Together at Last, *Genome News Network*, May 2002