HOX Gene Promoter Prediction and Inter-genomic Comparison: An Evo-Devo Study

Marla A. Endriga[1], Victoria Karenina R. de la Paz[1], Jezreel Marie G. Sazon[1], Elisa L. Co[1] and Custer C. Deocaris[2*]

[1]Department of Biology, College of Arts and Sciences, University of the Philippines Manila, 1000 Ermita, Manila, Philippines
[2]Department of Anatomy, Yonsei University College of Medicine, Seoul, South Korea

Abstract

Homeobox genes direct the anterior-posterior axis of the body plan in eukaryotic organisms. Promoter regions upstream of the Hox genes jumpstart the transcription process. CpG islands found within the promoter regions can cause silencing of these promoters. The locations of the promoter regions and the CpG islands of *Homeo sapiens sapiens* (human), *Pan troglodytes* (chimpanzee), *Mus musculus* (mouse), and *Rattus norvegicus* (brown rat) are compared and related to the possible influence on the specification of the mammalian body plan. The sequence of each gene in Hox clusters A-D of the mammals considered were retrieved from Ensembl and locations of promoter regions and CpG islands predicted using Exon Finder. The predicted promoter sequences were confirmed via BLAST and verified against the Eukaryotic Promoter Database. The significance of the locations was determined using the Kruskal-Wallis test. Among the four clusters, only promoter locations in cluster B showed significant difference. HOX B genes have been linked with the control of genes that direct the development of axial morphology, particularly of the vertebral column bones. The magnitude of variation among the body plans of closely-related species can thus be partially attributed to the promoter kind, location and number, and gene inactivation via CpG methylation.

Keywords:

Introduction

Humans are believed to have descended from the same ancestors as the apes, making the latter one of the closest relatives to humans in terms of evolution. Orthologous proteins are present in humans and chimpanzees, with 29% of them being identical and the rest typically differing only by two amino acids. Between the genomes of the two species, a mean rate of 1.23% single nucleotide substitutions occur, and 1.6% of this corresponds to the divergence between the species. Meanwhile, compared to another mammalian species, 40% of the human genome can be aligned to that of the mouse, representing orthologous sequences which remained from a common ancestor [1].

The anterior-posterior axis and the proper number and placement of segment structures of eukaryotic organisms during early embryonic development are controlled by clusters of Homeobox (Hox) genes. The basic body regions are laid out initially by the *Pax* homeobox genes in the somites, which are regulated by signals from the notochord and the neural tube [2]. The expression of *Pax3* is modulated by BMP-4 (bone morphogenetic protein 4) and the *Wnt* protein family, which ventralizes the mesoderm, confining it to muscle precursors.

Hox genes also specify positional identity, as evident in the differences in the vertebrae [2]. This indicates that a cell or group of cells in the embryo obtains a unique state according to its position at a given time during development.

Genes in each hox cluster are expressed in a temporal and spatial order that reflects their position on the chromosome. Similar sequences of certain gene sets are present in the genomes of other eukaryotic organisms, such as the mouse and chimpanzees, suggesting a high level of conservation in the homeotic domain and hence, a role in cell differentiation and embryonic patterning [3].

The Transcription Promoter Region (TPR) flanks Transcription Start Sites (TSS) and couples with the General Transcription Factors (GTFs) and Pre-Initiation Complex (PIC) during transcription. Thus, the biochemical

Corresponding address:
Custer C. Deocaris, PhD,
Department of Anatomy,
Yonsei University College of Medicine,
Seoul, South Korea
Email : cdeocaris@gmail.com

environment necessary for transcription is attained and the process begins [4]. Within the TPRs are dinucleotide clusters of CpGs, formally defined by Gardiner-Garden and Frommer [5] as a DNA region of about 200 bp with a high G+C content and with an Observed CpG/ Expected CpG ratio greater than or equal to 0.6. Methylation of a CpG site leads to repression of the gene, thus the state of CpG islands affects processes such as gene silencing, X-chromosome inactivation, silencing of intergenomic parasites, genomic imprinting, and carcinogenesis. Methylated cytosines have been mutational hotspots and have contributed to CpG depletion during the course of mammalian evolution. Around 40% of mammalian genes have CpG islands [6].

This study computationally predicts the location of the mammalian promoter regions and the CpG windows of Hox clusters A, B, C, and D of *Homo sapiens sapiens* (human), *Pan troglodytes* (chimpanzee), *Mus musculus* (mouse), and *Rattus norvegicus* (rat) and determines if there are significant statistical differences in the locations of these. The promoters present in the hox genes of each species are also identified and insights on possible factors that play a role in the specification of the mammalian body plan are gleaned.

Experimental Procedures

Sequences of homeobox genes of clusters A-D of *Homo sapiens sapiens* (human), *Pan troglodytes* (chimpanzee), *Mus musculus* (mouse), and *Rattus norvegicus* (rat) were retrieved from Ensembl (http://www.ensembl.org/index.html). Then these sequences as well as those 10 kilobasepairs (kbp) upstream and 10 kbp downstream of each gene were obtained in FASTA format. These were inputted to FirstEF (http://rulai.cshl.edu/tools/FirstEF/; 7) and gene promoters were predicted. The top-ranking predicted promoter for each species was considered for further analysis.

The predicted promoters were verified against the Eukaryotic Promoter Database (http://www.epd.isb-sib.ch/; 8) using advanced BLAST (Basic Local Alignment Search Tool). The promoter type and gene description were obtained and a Kruskal-Wallis test at a significance level of 0.05 was performed on the promoter location data.

Results

Identified HOX genes. A total of 141 homeobox gene sequences were retrieved from Ensembl (http://www.ensembl.org/index.html): the chimpanzee has 32 hox genes; humans have 41; mice have 39, and rats have 29. The hox genes present in the species considered are summarized in Table 1.

Promoter and CpG island locations. Many promoters present in one hox cluster of a species are also found in the equivalent cluster of other species considered (Figure S.8.1), especially in Hox cluster B. These include MAGEE1, Rn cytochrome C som, Gg histone H1-c10, ANXA1, SPAT, ALDOA E1P1, ALDOA E3P2, ALDOA E4P3, ALDOA E4P4,DDAX 17, to name a few. All are involved in processes such as transcriptional regulation, DNA binding, polymorphism, acetylation, apoptosis, and phosphorylation.

In Figure S.8.1B, it is seen that the number of promoter regions increased for each species while the number of CpG islands in the entire cluster decreased, compared to Figure S.8.1A. In Figures S.8.1C and S.8.1D, on the other hand, it is evident that the density of the promoter regions and CpG locations in each cluster are relatively lower compared to that in Figure S.8.1B. The temporal and spatial property exhibited by the Hox genes are evident in the maps shown. They generally show a relatively sharp anterior border and a less defined posterior border, and particular sets of expressed Hox genes characterize almost every region of the anterior-posterior axis (2).

Statistical Analysis. Promoter locations of the four species relative to the Hox clusters were compared using the Kruskal-Wallis test (α = 0.05). It was found that only promoter locations in Hox Cluster B are significantly different (Table 2), thus, the locations of the promoter regions of *Homo sapiens, Pan troglodytes, Mus musculus,* and *Rattus norvegicus* in Clusters A, C and D are similar.

CpG island locations were also predicted. As evident from Figure S.8.1, overlaps in promoter regions and CpG windows exist. This suggests that Hox gene expression is not solely regulated by promoters but also by other factors which may not have been included in the analyses done. Results indicate that there is no significant difference among the CpG island locations

Table 1. Homeobox genes in the four mammalian species considered

Species/ Hox gene	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13
Pan trogolodytes	✓	✓	✓	✓	✓		✓		✓		✓		✓
Homo sapiens	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓
Mus musculus	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓
Rattus norvegicus		✓			✓		✓			✓			✓

Species/ Hox gene	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13
Pan trogolodytes	✓	✓	✓	✓		✓			✓				✓
Homo sapiens	✓	✓	✓	✓	✓	✓	✓	✓	✓				✓
Mus musculus	✓	✓	✓	✓	✓	✓	✓	✓				✓	✓
Rattus norvegicus	✓			✓	✓	✓		✓	✓				✓

Species/ Hox gene	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
Pan trogolodytes				✓	✓	✓		✓	✓		✓	✓	✓
Homo sapiens				✓	✓	✓		✓	✓	✓	✓	✓	✓
Mus musculus				✓	✓	✓		✓	✓	✓	✓	✓	✓
Rattus norvegicus				✓	✓	✓				✓		✓	✓

Species/ Hox gene	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
Pan trogolodytes	✓		✓	✓				✓	✓	✓	✓		✓
Homo sapiens	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓
Mus musculus	✓		✓	✓				✓	✓	✓	✓	✓	✓
Rattus norvegicus	✓		✓	✓				✓	✓	✓		✓	✓

✓ = gene is present in the species

Table 2. Results of the Kruskal-Wallis Statistical test (α = 0.05) on the Promoter locations in Hox clusters of the four species considered. The number of degrees of freedom in each case is 3.

Cluster/Parameter	Adjusted H	p-value
A	0.617	0.892
B	13.62	0.00347
C	0.158	0.984
D	0.492	0.921

Table 3. Results of the Kruskal-Wallis Statistical test (α = 0.05) on the CpG island locations in Hox clusters of the four species considered. The number of degrees of freedom in each case is 3.

Cluster/Parameter	Adjusted H	p-value
A	1.302	0.729
B	7.293	0.063
C	0.463	0.927
D	0.319	0.956

of clusters A to D of *Homo sapiens, Pan troglodytes, Mus musculus,* and *Rattus norvegicus* (Table 3).

This supports the contention that there are other various transcription regulators, as well as microRNAs and the cellular cycle, present in the hox genes, which can affect promoter inactivation.

Discussion

Promoter locations. Almost all bilateral animals have similar presentation and expression of Hox genes (9). As Hox clusters have genes that are comparable in sequence and in chromosomal position, the promoter sequences of the different species were examined. Bengani and colleagues' (10) *in silico* analysis of the upstream/intronic sequences of the homeobox genes of the mouse, chimpanzee, and human revealed novel motifs lacking binding sites for known transcription factors. They predicted that these could be positions of chromatic modifying complexes involved in epigenetic regulation.

Hox gene expression is known to direct morphological development such as body patterning. In this study, it was found that promoter location from different species appeared to have no significant difference from one another. The diversity of body plans of these species may therefore be caused by components other than those tested, such as the Hox-regulated enhancers. Mutations at these enhancers may play a role in directing the fate of genes regulated by Hox. From experiments which alter the Hox binding sites of enhancers, Capovilla et al. (11) observed the resulting binding affinity. The presence of non-Hox proteins in the mutant binding sites of the enhancer caused reduced response to the Hox-regulator. Testing the similarities of in vivo experiments to wild-type enhancers in mutant embryos is thus recommended.

DNA Methylation. CpG islands, although still poorly understood, have been recognized as one of the key players in genetic regulation during normal development and cell differentiation (12). About 50-70% of CpG islands are found in promoter regions and near transcription start sites (13). Zhang et al. (14) reported a positive spike in the GC content near these sites, and a negative spike near the stop site, hence the belie that the methylation of CpG

islands in vertebrate genomes is relevant to gene expression (15).

In this study, no significant difference among the CpG locations in the homeobox genes of the four species was found. It has been reported previously that during evolution, there have been only minimal changes in the G+C content of closely-related species. Sakaki et al. (16) showed that the genomic difference between humans and chimps are at a mere 1.23% at the nucleotide level. The human chromosome 21 (HSA21) was compared against the chimpanzee PTR22. It is one of the most studied human chromosomes since it is a representative of the human genome, having repetitive and duplicated structures and uneven distributions of G + C content with a high correlation to density. The G + C content for both species was estimated at only 41%, with modern humans displaying a slight increase during evolution, while that of the chimpanzees stayed constant. They also showed that genes with high sequence divergence of associated CpG islands were more likely to have changed their expression. Additionally, Vinogradov (15) showed that there is a weak correlation between the maximum level of gene expression and promoter CpG island, compared to the GC content of intronic sequence and third codon position of the coding sequence, which has the strongest correlation. This is due to the broader definition of promoter CpG islands that may likely include Alu-associated CpG islands. Vinogradov (15) suggested that Alu repeats can also have regulatory elements.

Curradi et al. (17) reported that transcription repression happens through direct interference with the binding of transcription factors to DNA. Transcriptional regulators that cannot bind methylated recognition elements only become capable of repression after chromatin assembly. A few methylated cytosines can inhibit a flanking promoter but a required number of modified sites is needed for repression. When methylation does not reach sufficient levels to establish the inactivated chromatin structure, histone deacytelation causes gene repression, where a repressive chromatin environment is formed. Transcriptional repression does not always require methylation of the promoter, and promoter modification does not always lead to greater repressive effects because there is competition between transactivators and methyl-binding proteins. Curradi et al. (17) further proposed three main important factors that contribute to transcriptional repression: distance of methylation sites from the promoter regions, length of the modified sequence, and density of the methylated cytosines. Additionally, methylation at specific critical CpG sites and the abundance of transcription factors contribute to transcription repression.

The mechanism for DNA methylation involves the transfer of methyl to DNA, a process that involves DNA methyltransferases. There are three methyltransferases that maintain and establish methylation in mammals: Dnmt1, Dnmt3a and Dnmt3b. The last two are important for *de novo* methylation while the first one is for maintenance (18). The mammalian genome contains around $3x10^7$ residues of 5-methylcytosine, and most are within the 5'-m^5CG-3' dinucleotides. The primary methyl donor is the S-adenosylmethionine, also called SAM or AdoMet (13). The target cytosine is pulled from the DNA helix and is pocketed deep into the active site of the enzyme. Once there, catalytic cysteine thiolate forms an intermediate state with the Carbon-6 of the cytosine ring; reactive carbons 4 and 5 form an enamine that attacks the methyl group and transfers it to carbon 5. Proton abstraction from carbon 5 leads to the reformation of the double bonds in C-5 and C-6 and to the release of the enzyme (19).

In mammals, the bulk of DNA methylation happens at the many repetitive sequences that are considered as "junk DNA" (19). The methylation process also increases the coding capacity of the genome, and reversible methylation and demethylation is involved in the regular development of the embryo. There is also irreversible promoter silencing that appears to be restricted to organisms with modified bases (13).

There have been attempts to fully map CpG islands in the genomes of mammals. Illingworth et al. (20) opine that CpG mapping is still unstable and depends upon the implementation of the software used in predicting the islands due to the variations in the CpG regions. A situation worth considering would be that of a short CpG region. Although it fails to fulfill the set island criteria, the same region may fulfill the criteria for the small and seemingly unrelated changes in a few neighboring nucleotides. Thus, they have suggested the use of numerical scores that could quantify the strength of a CpG region.

Other factors that affect transcription. Bestor (19) suggested that Transcription Factors (TFs)

contribute to transcriptional repression. The binding of TFs has the capacity to determine the fate of the organism's segments by competing for the establishment of an inactive promoter conformation. Methylation alone is not sufficient to cause inactivation (17).

MicroRNAs (miRNAs) are also being studied because Hox genes are possible miRNA targets (9). Many miRNA-Hox interactions have been proposed but only the mouse Hoxb8 transcript is validated to be a miRNA target. Kawasaki and Taira (21) tested the expression of Hoxb8 in the presence of miR-196 (a miRNA), and observed the decreased level of Hoxb8. Other miRNAs are also being studied for their influence on Hox gene expression as they are encoded with Hox gene clusters. These include the mir-10a located near Hoxb4 and mir10b near Hoxd4. The possible influence of the miRNAs located within the Hox cluster to axial patterning is being looked into, since it has been observed that the expression pattern of miRNAs are similar to that of their adjacent Hox gene, suggesting coordinated regulation.

Conclusion

The promoter regions and CpG window locations of HOX genes in humans, chimpanzees, mice, and rats vary significantly from one another in cluster B but not in clusters A, C and D. This supports the link between the involvement of HOX B genes in the development of the axial morphology and the differentiation in the body plans of closely-related species, particularly in mammals, which have extreme variations in body patterns. Since there is a high level of conservation of the HOX genes among different species, factors such as the presence of transcription factors, miRNAs, and other enhancers and silencers may have greater roles in the development of the mammalian body plan and in regulating the expression of the HOX genes.

References

[1] Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P. Lander ES (Mouse Genome Sequencing Consortium). *Nature* 2002; **420**:520-562

[2] Wolpert L, Beddington R, Jessel T, Lawrence P, Meyerowitz E, Smith J. Principles of development. 2nd ed. Oxford University Press, USA, 2002.

[3] Holland PWH, Booth AF and Bruford EA. *BMC Biol* 2007; **5**:4

[4] Bajic VB, Brent MR, Brown RH, Frankish A, Harrow J, Ohler U, Solovyev VJ, and Tan SL. *Genome Biol* 2006; **7**: S3

[5] Gardiner-Garden M and Frommer M. *J Mol Biol* 1987;**196** (2):61-82

[6] Takai D and Jones PA. *Proc Natl Acad Sci USA* 2002; **99**(6): 3740-3746

[7] Davuluri RV, Grosse I and Zhang MQ. *Nature Genet* 2001; **29**:412-417

[8] Schmid CD, Perier R, Praz V and Bucher P. *Nucleic Acids Res* 2006; **34**:D82-85

[9] Pearson JC, Lemons D and McGinnis W. *Nature Rev Genet* 2005; **6**, 893-904

[10] Bengani H, Ganapathi M, Singh GP, Brahmachari V. *J Exp Zool B Mol Dev Evol.* 2007; **308** (4):384-95

[11] Capovilla M, Kambris Z, Botas J. *Development*. 2001; **128** (8):1221-30

[12] Shen L, Kondo Y, Guo Y, Zhang J, Zhang L, Ahmed S, Shu J, Chen X, Waterland RA, and Issa JPJ. *PLOS Genet* 2007; **3** (10):2023-36

[13] Siedlecki P and Zielenkiewicz P. *Acta Biochimica Polonica* 2006; **53**(2):245-256

[14] Zhang MQ, Xuan F, Wang J and Chen G. *Genome Biol* 2005; **6**:R72

[15] Vinogradov AE. *Trends Genet* 2005; **21** (12):639-43

[16] Sakaki Y, Fujiyama A, Toyoda A, Watanabe H, Hattori M, and Taylor TD. *International Congress Series* 2002; **1246**:183-187

[17] Curradi M, Izzo A, Badaracco G, and Landsberger N. *Molecular and Cellular Biology* 2002; **22** (9):3157-73

[18] Feng YQ, Desprat R, Fu H, Olivier E, Lin CM, Lobell A, Gowda SN, Aladjem MI, and Bouhassira EE. *PLOS Genet* 2006; **2** (4):e65

[19] Bestor T. *Human Molec Genet* 2000; **9** (16):2395-2402

[20] Illingworth R, Kerr A, DeSousa D, Jørgensen, Ellis HP, Stalker J, Jackson D, Clee C, Plumb R, Rogers J, Humphray S, Cox T, Langford C and Bird A. *PLOS Biol* 2008; **6** (1):e22

[21] Kawasaki H and Taira K. *Nucleic Acids Symp Ser* 2004; **48** (1): 211-2