Journal of Biology
J. Biol.
BioMed Central

Minireview

Conserved elements within open reading frames of mammalian Hox genes

Joost M Woltering* and Denis Duboule*†

Addresses: *National Research Centre 'Frontiers in Genetics', Department of Zoology and Animal Biology, University of Geneva, Sciences III, Quai Ernest-Ansermet 30, 1211 Geneva 4, Switzerland. †National Research Centre 'Frontiers in Genetics', School of Life Sciences, Ecole Polytechnique Fédérale, 1015 Lausanne, Switzerland.

Correspondence: Denis Duboule. Email: Denis.Duboule@unige.ch

Published: 6 February 2009

Journal of Biology 2009, **8**:17 (doi:10.1186/jbiol116)

The electronic version of this article is the complete one and can be found online at http://jbiol.com/content/8/2/17

Abstract

A recent study in *BMC Evolutionary Biology* shows that many of the open reading frames in mammalian Hox genes are more conserved than expected on the basis of their protein sequence. The presence of highly conserved DNA elements is thus not confined to the noncoding DNA in neighboring regions but clearly overlaps with coding sequences. These findings support an emerging view that gene regulatory and coding sequences are likely to be more intermingled than once believed.

More conserved than conserved

Comparisons between vertebrate genomes reveal a range of highly conserved sequences located within noncoding regions [1,2]. These conserved elements are typically 50 to 300 nucleotides long and were initially identified by alignments of orthologous loci. More recently, whole-genome sequence comparisons have provided rather exhaustive accounts of such elements, which were identified using criteria of various stringencies and (expectedly) are referred to by different terminologies [1-4]. For example, the best-conserved class of elements, named 'ultraconserved elements', contains sequences 200 bp long at least and identical among human, mouse and rat [3]. Another study describes 'ultraconserved regions', that is, sequences showing both a minimum of 95% identity over 50 bp between human and mouse, and also some homology with their counterparts in the fish *Fugu* [4]. As these criteria are somewhat arbitrary, and because unambiguous functional criteria allowing for a more relevant classification of these sequences are still lacking, we shall consider all these sequences as a whole and, for the sake of simplicity, refer to them as 'conserved elements'.

Conserved elements are preferentially associated with either genes encoding transcription factors or genomic loci important for development [1]. Ever since they were discovered, it has been assumed that the function of these elements is, primarily, to regulate the expression of neighboring gene(s), as short-range regulators or long-range enhancers, to help establish the complex and dynamic expression patterns of these genes [1,5,6]. However, the biological relevance of these elements is still elusive and several instances in which conserved elements were removed from the mouse genome *in vivo* failed to clearly support this hypothesis (for example [6]). While conserved elements are mainly present in noncoding regions, several studies have identified a significant number overlapping with coding regions (for example [3]). Interestingly, these latter elements are usually excluded from global analyses, probably because their identification and interpretation represent an additional level of complexity. Coding regions are indeed expected to exhibit a significant degree of nucleotide sequence conservation, due to strong constraints on the corresponding amino acid sequences, and whenever coding and regulatory sequences overlap, the nucleotide sequence becomes informative regarding two independent processes, each associated with its own set of constraints.

A strategy to identify protein-coding DNA regions under evolutionary constraints other than that of generating a faithful amino acid sequence is to look at the balance between so-called silent, or synonymous, nucleotide substitutions (that is, those that do not modify the amino acid sequence) and nonsynonymous substitutions (those that have an impact on protein sequence and, likely, protein function). When the unique task of a given nucleotide sequence is to encode a protein sequence, synonymous substitutions are, expectedly, under near-neutral selection, whereas nonsynonymous substitutions will be mostly under purifying (negative) selection or, much more rarely, positive selection for improved function. In contrast, if a strong constraint on the nucleotide sequence is added, such as the presence of consensus binding sites for regulatory proteins, then synonymous substitutions may also be under negative selection. Screening for coding sequences with a bias in the proportion of synonymous substitutions can thus be informative in this respect, and it has been shown that many open reading frames (ORFs) indeed display such a decreased rate of synonymous substitutions. Interestingly, this observation often involves transcription factors and genes with developmental functions; that is, a sample of genes comparable to those in which conserved elements are found outside protein-coding regions.

Conserved elements within Hox gene ORFs

In recent work published in *BMC Evolutionary Biology*, Lin and colleagues [7] investigated mammalian Hox genes for a bias in synonymous versus nonsynonymous substitutions. Hox genes encode a family of homeobox-containing transcription factors involved in many developmental processes during embryogenesis. While they are best known for their role in patterning the main body axes, Hox genes are also necessary for organogenesis. Mammals have 39 Hox genes, which are organized in four genomic clusters (HoxA, B, C and D) with 13 paralogous groups (Hox1 to Hox13), which are the result of two successive genome duplications that accompanied the transition towards vertebrates. Both the integrity of these loci and their syntenic relationships have been highly conserved during evolution [8]. Notably, Hox gene clusters are associated with many noncoding conserved elements, located both within and outside the clusters themselves, which are believed to participate in the transcriptional control of these genes during development.

Lin and colleagues analyzed the Hox ORFs using a sliding-window strategy to identify regions devoid of synonymous substitutions in pairwise alignments between human, mouse, dog, cow and opossum loci [7]. Using a 120-bp sliding window, each possible sequence of 120 consecutive bases within the ORFs (that is, nucleotides 1-120, nucleotides 2-121, nucleotides 3-122, and so on) was independently analyzed for synonymous substitutions. Each window that did not contain any such substitutions is considered part of a conserved element and where multiple ‘empty’ windows overlap they are grouped in the same element. This way, conserved elements of 120 bases or longer will be detected. Interestingly, Lin and colleagues report differences in the rates of synonymous substitution even between closely related paralogous genes, as illustrated by the alignment of the murine and human *HoxC6* and *HoxB6* DNA sequences (Figure 1, in which only synonymous substitutions are visualized). The *HoxC6* DNA sequence displays regions of extended conservation at the nucleotide level, whereas variations in *HoxB6* sequences are as expected on the basis of a degenerate genetic code. These data strongly suggest the presence of an additional constraint acting over the *HoxC6* DNA sequence that is different from that imposed by the mere production of the corresponding protein.

The authors point out that the presence of such conserved elements within coding regions seems to be rather specific to eutherian mammals; they are not found in chicken or platypus, and are mostly absent from the opossum genome. Lin *et al.* substantiate their findings by showing that such conserved elements cannot be identified in Hox gene ORFs when other taxa at comparable evolutionary distances are compared (for example, different *Drosophila* species or the two sequenced pufferfish species). It appears that these elements became constrained, and hence stabilized, in early placental mammals. Consequently, and also because one would expect the emergence of an internal reproductive system to be accompanied by the recruitment of specific enhancers for Hox genes, Lin *et al.* suggest that these conserved elements might be related to the evolution of the placenta. Testing of this hypothesis will have to await careful functional analysis using mouse molecular genetics.

The role of conserved elements: why such conservation?

As it is unlikely that conserved elements merely correspond to ‘cold spots’ - that is, places where a decreased mutation rate (rather than purifying selection) results in no sequence variation [9] - a critical challenge now is to understand the function(s) of these elements, and hence the mechanisms

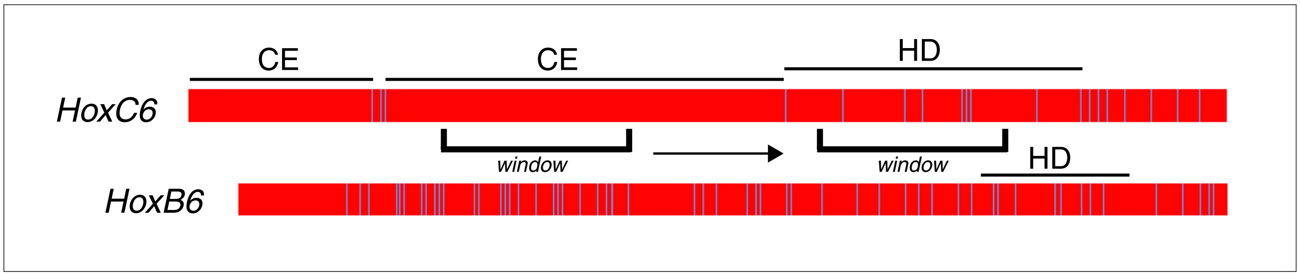

Figure 1
Schematic diagram of synonymous substitutions between human and murine *HoxC6* and *HoxB6* nucleotide sequences. This diagram shows that many more synonymous substitutions (blue bars) are present in *HoxB6* than in *HoxC6*. The two conserved elements (CEs) identified in *HoxC6* by Lin *et al*. [7] are indicated, as well as the position of the homeodomain (HD). The sliding-window strategy is visualized by the positioning of a 120-bp window within a CE as well as over the homeodomain, which is not a CE because it does not contain stretches of 120 consecutive bases devoid of synonymous substitutions. The sequence encoding the homeodomain, at the amino acid sequence level one of the most conserved features of Hox genes, still contains multiple synonymous substitutions in both *HoxC6* and *HoxB6*, whereas the 5' region of *HoxC6*, which encodes a domain of the protein without any clearly defined function, is virtually 100% conserved. It should be noted that the HoxB6 protein is overall slightly less conserved than HoxC6, between mouse and human, and contains five nonsynonymous nucleotide substitutions (which are not indicated here), whereas HoxC6 is fully conserved at the amino acid level.

constraining their high degree of sequence conservation. Because many conserved elements display enhancer activity [1-5], it seems reasonable to assign them a role in transcriptional regulation, and there is no particular reason why those elements identified by Lin and colleagues should belong to a fundamentally different class. Their location within ORFs would support the emerging view that regulatory and coding sequences are more intermingled than anticipated [10]. There are, however, some difficulties with this conventional interpretation. First, the interactions between transcription factors and their binding sites are notoriously promiscuous and can thus hardly offer an explanation, by themselves, for the high purifying selection observed for conserved elements [1]. Second, several sequences carrying specific enhancer potential do not show any obvious interspecies conservation. Finally, some conserved elements can be deleted *in vivo* without any apparent effects [1]. Does this mean that DNA sequences strongly conserved during evolution might not necessarily be of functional importance - and *vice versa*?

In this context, a critical parameter to consider is the heuristic values of the various readouts, which, by definition, are biased by current views of transcriptional regulation. For example, many conserved elements are located several hundred kilobases from the genes they are believed to regulate, whereas they sometimes 'ignore' genes located nearby. In such cases, the mechanisms by which conserved elements contact their target promoters at the appropriate times and places during development are still poorly understood. Also, such regulatory sequences may be involved in higher-order chromatin structure or even in the three-dimensional organization of chromosomes; these kinds of processes are arguably difficult to document in a classical transgenic assay, which is normally designed to study more local interactions between enhancers and promoters. Standard transgenic assays do not tell us, for instance, about the capacity of particular sequences to mediate DNA looping in order to confer transcriptional activity on target genes located at considerable distances in the right cells at a precise time. In this context, conserved elements located within ORFs, or in the vicinity of transcription units, could serve as 'docking sites' for sequences located further away. Transgenic assays are also limited whenever repressive sequences are considered.

The same limitations hold true in phenotypic analyses and it is possible that many conserved elements are involved in regulating genes in places and at times such that their effects escape notice. It is also conceivable that compensatory mechanisms exist, which make the actual function of some conserved elements impossible to assess using current genetic tools. Finally, the apparent lack of effect of removing some of the elements *in vivo* could reflect the redundancy of some transcription regulatory circuitry. Although the evolution of compensatory and/or redundant mechanisms for their own sake is difficult to envisage, such properties may have emerged as a result of other constraints associated with developmental processes. It is even possible that regulatory redundancy should be considered, in turn, as increasing the potential for evolvability by stabilizing critical expression domains, thereby allowing greater flexibility in evolving novel regulation. In any case, these mechanisms could increase the robustness of developing systems under a broad range of physiological conditions, which is difficult to test experimentally. Whether or not the intriguing conserved sequences reported by Lin *et al.* could be instrumental in any of these processes remains to be demonstrated.

Acknowledgements

JMW is supported by an EMBO long-term fellowship. The laboratories are supported by funds from the University of Geneva, the Federal Institute of Technology (EPFL) in Lausanne, the Swiss National Research Fund, the National Research Center (NCCR) 'Frontiers in Genetics' and the EU programs 'Cells into Organs' and 'Crescendo'. Due to journal policy, we have only sparingly referenced the literature and apologize to those whose work we were unable to specifically mention.

References

1. Elgar G, Vavouri T: **Tuning in to the signals: noncoding sequence conservation in vertebrate genomes.** *Trends Genet* 2008, **24:**344-352.
2. Dermitzakis ET, Reymond A, Antonarakis SE: **Conserved non-genic sequences - an unexpected feature of mammalian genomes.** *Nat Rev Genet* 2005, **6:**151-157.
3. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome.** *Science* 2004, **304:**1321-1325.
4. Sandelin A, Bailey P, Bruce S, Engström PG, Klos JM, Wasserman WW, Ericson J, Lenhard B: **Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes.** *BMC Genomics* 2004, **5:**99.
5. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM: **In vivo enhancer analysis of human conserved non-coding sequences.** *Nature* 2006, **444:**499-502.
6. Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM: **Deletion of ultraconserved elements yields viable mice.** *PLoS Biol* 2007, **5:**e234.
7. Lin Z, Ma H, Nei M: **Ultraconserved coding regions outside the homeobox of mammalian Hox genes.** *BMC Evol Biol* 2008, **24:**260.
8. Duboule D: **The rise and fall of Hox gene clusters.** *Development* 2007, **134:**2549-2560.
9. Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET, Hirschhorn JN: **Conserved noncoding sequences are selectively constrained and not mutation cold spots.** *Nat Genet* 2006, **38:**223-227.
10. Tümpel S, Cambronero F, Sims C, Krumlauf R, Wiedemann LM: **A regulatory module embedded in the coding region of *Hoxa2* controls expression in rhombomere 2.** *Proc Natl Acad Sci USA* 2008, **105:**20077-20082.