A Probabilistic Approach to the Assessment of Phylogenetic Conservation in Mammalian *Hox* Gene Clusters

Nikola Stojanovic[1] **Ken Dewar**[2,3]

[1] *Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, Texas, USA*
[2] *Department of Human Genetics, McGill University, and*
[3] *Genome Quebec Innovation Centre, Montreal, Quebec, Canada*
Email: nick@cse.uta.edu, ken.dewar@mcgill.ca

ABSTRACT: With the increasing availability of mammalian genome sequences it became possible to use large scale phylogenetic analysis in order to locate potentially functional regions. In this paper we describe a new probabilistic method for the characterization of phylogenetic conservation in mammalian DNA sequences. We have used this method for the analysis of *Hox* gene clusters, based on the alignment of 6 species, and we constructed a map of *Hox* indicating short and long conserved fragments and their positions with respect to the known locations of *Hox* genes and other elements, sometimes showing surprising layouts.

1 INTRODUCTION

The utility of comparative sequence analysis has been evident for many years, as the conservation patterns reveal the homology between genomic segments, as well as the effects of functional constraints on mutations [4]. A popular theory is that conserved regions did not succumb to the evolutionary drift due to the effect of deleterious mutations, so sequence alignments became important for locating functional loci in DNA [9]. For larger regions, such as gene exons, pairwise alignments already serve well, but for subtler signals one needs multiple sequences. Until just a few years ago large scale studies of multiple alignments were not possible in eukaryotes, mammals in particular, but with the advancement of sequencing projects this situation is rapidly changing. Projects directed towards targeted sequencing of genomic regions for the purpose of the analysis of conservation [17], [16] have already started.

Early attempts to identify functional DNA elements based on phylogenetic conservation were largely heuristic, but there were efforts to statistically characterize them with respect to the background [8]. Local background is important, as it has been known for a while that regions like *Hox* gene clusters may be protected from evolutionary drift by some yet unknown mechanism [3]. Thus if a region of good conservation appears unlikely in its environment, that is a strong indication that it is important. A "likely" region can still be functional, but it does not stand out clearly enough to suggest function based solely on the conservation. However, before addressing this issue we need to characterize what "local" means. Concerning the rate of change between homologous sequences it should clearly be an area over which this rate does not vary much.

We have applied our method to the analysis of mammalian *Hox* clusters. We have chosen them because they have been extensively studied, so the locations of genes and many other elements are known, and also because of their good phylogenetic conservation, which involves features other than simple conservation of sequence motifs, like the conservation of intergenic spacing in paralogous clusters and an apparent resiliency to the insertion of transposable elements [7]. There are 4 *Hox* clusters in mammals (labeled A, B, C and D), spanning about 100–200 Kb each, and containing a total of 39 genes in human, in 13 groups of paralogs (labeled 1 through 13). They are ordered in the same way in each cluster, although not every cluster contains the full set of 13 genes. The function of the paralogs is only partially redundant, as the loss of one cannot be completely compensated by the others [5].

Hox genes, named after a common *homeodomain* motif, produce transcription factors which regulate the formation of the anterior–posterior axis of an animal during early embryonic development, acting on a large number of downstream genes. Since this axis is common throughout the evolution, *Hox* clusters are well conserved, often over regions much longer than expected under a simple model of coding sequence and transcriptional regulation. This was instrumental for our purpose, because we could build reliable multiple alignments, and also expect that the nature and positions of functional elements could not have varied much. However, the main goal of our work was the characterization of overall phylogenetic sequence conservation in *Hox* clusters, rather than a search for individual functional elements. The latter task can be better achieved by projects like ENCODE [16], since *HoxA* is already one of its target regions, and it would likely spark further analysis of its paralogs.

2 METHODS

We have constructed long alignments of all 4 *Hox* clusters and their surrounding regions (of about 500 Kb each, measured in human sequence) from 6 mammals representing 3 distinct groups: two primates (human and baboon), two ungulates (cow and pig) and two rodents (mouse and rat), using Multi-LAGAN software [2]. Because of the state of the sequences at the time when these alignments were built we had to restrict our analysis only to a large contiguous high–quality interval in each cluster, but the resulting fragments included a majority of *Hox* genes.

We started by fragmenting the alignments into large blocks where the conservation rate appeared constant. This was done iteratively, expanding the small seed blocks until the application of the Central Limit Theorem indicated that the neighboring ones are unlikely to draw from the same distribution, at 0.99 or higher significance. The seed length has been set to 50, as we wanted it as short as possible and below this sample size the application of the CLT may be unreliable.

The alignment columns were scored using the weighted parsimony algorithm [12], although we employed a somewhat naive model. It has been established that rodent evolution rate is faster than that of primates [10], but the relative positions of rodent and ungulate branches on the evolutionary tree, with respect to primates, are still controversial (all 3 groups are at about equal distance of 80–100 Myr). We have applied a model under which rodents are closer to primates, as there appears to be accumulating evidence in support of this hypothesis. Insertions and deletions, reflected as gaps in the alignment, were treated as any other substitutions, even if a chain of gaps likely corresponds to a single evolutionary event. This way the entire alignment was represented as an array of scores, divided into blocks of the initial seed length, to be further refined. In each iterative step we calculated the means and sample variances of the neighboring regions, then used the mean of the larger sample as the true mean, and the smaller sample for the calculation of the confidence interval. These steps were repeated until there was no change in the total number of blocks. Once it has been determined that neighboring blocks were unlikely to feature the same conservation rate, further refinement was done in order to establish the most likely boundary, by moving it until it optimally distributed the columns closer to one of the two means.

Intuitively, large blocks of constant conservation correspond to genome loci with the same mutation rates. This can be due to different concentration of long and short functional DNA elements or due to some other mechanism protecting specific domains. After such blocks have been determined, we proceeded to identify the outliers. The expectation was that these blocks would roughly correspond to gene exons, as they would be the only known elements that would warrant long blocks of consistent good conservation.

Knowing the background conservation rate, it is possible to isolate shorter regions significant within their own environment. Since the lower values for individual alignment column scores obtained through the application of parsimony indicated better conservation, we modified them by subtracting them from the average local background divergence. That gave the best score to the most conserved columns, and only these scoring better than the mean remained positive. However, we now assigned an infinite negative score to gap–containing columns — while some significant areas might be lost because of this strategy, it also protected us from dealing with blocks in which all but one sequence featured a gap. We used the modified scores in order to isolate the full runs of columns, by applying an algorithm we adapted from Bentley [1]. We define full runs as the maximal intervals scoring higher than any of their subintervals. Our algorithm (unpublished) locates the full runs in $O(n)$ time, where n is the size of the score array, i.e. the number of columns in the alignment.

We have calculated the mean and the variance for each of the located regions, and used them for the comparison with these of all background environments. However, the located regions may not be significant in their own surroundings, so they needed to be further evaluated. Because they were generally short (up to a few dozen bases), we used the Student t test. Due to the decrease in variance when the average is taken over the longer intervals (and the increase in the degrees of freedom), longer ones may be more likely to pass the significance threshold, although in purely random setting they would be also less likely to stand out. This corresponds well with their presumed biological meaning, however the quality of the background conservation introduces a semantic bias. Blocks with a significant mean should thus be considered by that measure only, while the significance test should be applied to these discovered in poor background conservation areas.

Our assumption was that if the areas of constant conservation rates do not capture the exons of *Hox* genes,

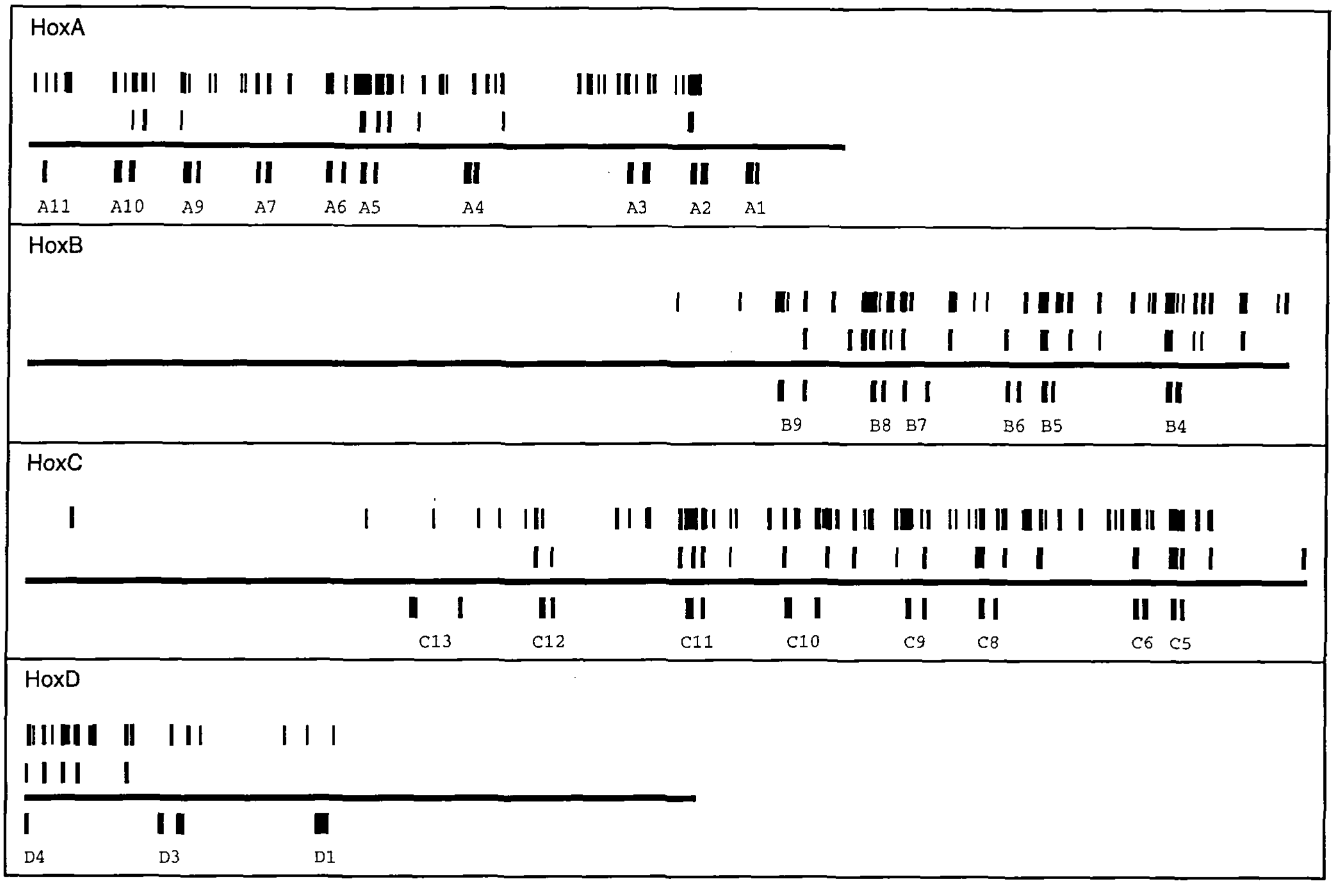


Figure 1: Blocks of maximal sequence conservation in all 4 *Hox* clusters. Thick horizontal lines indicate the range that has been analyzed, with position of *Hox* genes indicated below (all *Hox* genes have two exons). The bars right above the line indicate the positions of larger (50–400 bp) environments conserved at, or below, 0.1 average level, and the bars above them indicate the positions of shorter (25–100 bp) regions at the same level of conservation.

at least their fragments would show up as long significant blocks. Shorter intervals would indicate possible transcription factor binding sites and other functional elements, but their actual prediction would require further work, such as clustering or Bayesian analysis. The fact that a region is distinguished from its environment still need not imply that it has a function, and we have thus limited our study to the annotation of the sequence sites according to how unusual they are, leaving the actual determination of functionality to expert estimates and laboratory analysis.

3 RESULTS

The primary motivation for doing this work originated from an informal observation that the overall conservation patterns in our alignments did not appear to fit well with our expectations. If an alignment is biologically correct (and a mathematical optimum under a good scoring scheme would presumably come close to that), one would expect that gene exons would stand out more–or–less clearly, while the right regulatory sequences would be dotted with clusters of conserved transcription factor binding motifs. Of course, because of the inter–species genetic variations and the lack of DNA sequence specificity of most regulatory proteins this rarely happens, but a reasonably close alignment layout is intuitive [15].

Discarding the opening and closing gaps in incomplete sequences, we have selected for analysis only parts of the alignments that exhibited reasonable sequence and layout quality throughout. In *HoxA* that was from the second exon of *HoxA11* gene through about 11 Kb 3' to *HoxA1* gene, including the 3', but not the 5' end of the cluster. In *HoxB* it was from about 7 Kb 3' to *HoxB13* to about 13 Kb 3' to *HoxB4*, thus missing several genes at both 5' and 3' end of the cluster. Interestingly, due to the deactivation of 3 genes between *HoxB13* and *HoxB9* this left us with a large intergenic region at the opening end. In *HoxC* we have selected the area between about 48 Kb 5' to *HoxC13* to about 15 Kb 3' of *HoxC5*. This included the 5' end of *HoxC* cluster, with a significant starting intergenic area, but excluded its 3' end, missing the *HoxC4* gene. In *HoxD* we had the least sequence to work with — our fragment included the last 264 bp of

	500–1000 5’	200–500 5’	0–200 5’	Exons	Introns	0–1000 3’	Intergenic
HoxA, long	2	1	1	2	0	1	4
HoxA, short	5	10	14	28	10	15	54
HoxB, long	1	2	5	9	0	1	10
HoxB, short	4	7	7	3	8	14	29
HoxC, long	1	2	4	7	4	2	8
HoxC, short	3	4	9	8	13	15	51
HoxD, long	0	0	0	1	0	0	(4)
HoxD, short	0	0	0	0	2	4	(22)

Table 1: Number of short and long regions of average conservation 0.1 substitution per site, or better, falling into each distinct genomic domain. The intergenic region numbers for *HoxD* have been parenthesized because of the Ensembl gene prediction at the location where many of these regions have been found.

	500–1000 5’	200–500 5’	0–200 5’	Exons	Introns	0–1000 3’	Intergenic
HoxA	0.067	0.315	0.616	0.223	0.066	0.077	0.057
HoxB	0.115	0.342	0.788	0.639	0.071	0.145	0.024
HoxC	0.104	0.202	0.609	0.521	0.089	0.105	0.035
HoxD	0	0	0	0.061	0.026	0.066	(0.027)

Table 2: Fractions of the total number of alignment columns in each distinct genomic domain contained in the regions of minimal length 25 bp, with average conservation 0.1 or better. The intergenic data for *HoxD* have been parenthesized because of the Ensembl gene prediction at the location where many of these regions were found.

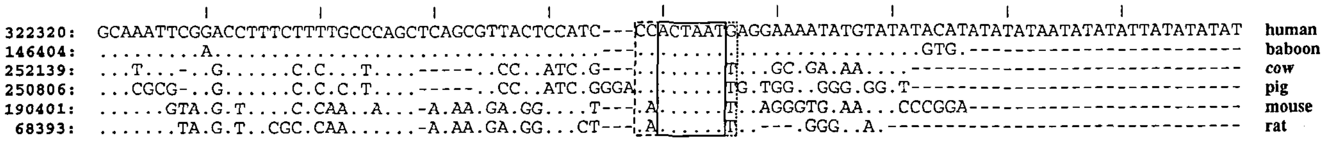
the intron of *HoxD4* (thus missing 6 *Hox* genes, plus 1 exon) until about 46 Kb 3’ of *HoxD1*. At the 3’ end of the *HoxD* cluster we thus had a large segment of intergenic sequence, however there is an Ensembl [6] prediction of another gene (XP_496612.1) in that area. Overall, this gave us a good blend of *Hox* environments in which any patterns should be clearly visible.

The initial breakup of the alignments into areas of constant conservation rate was somewhat surprising, dividing them into a large number of blocks of 250 bp on average, which could not be further merged. This was primarily due to very low sample variances, and that confirmed the known fact that genomic sequences are far from random, even outside genes. However, since intervals of this size can capture exons, we were content with this division, especially as it did not substantially change with large increases in the significance threshold. Using this division we have also located shorter full runs standing out in these environments.

We first looked at all segments, either large constant–rate environments or shorter regions of minimal length 25 bp, featuring a parsimony score of at most 0.1 substitutions per site. Minimal length was set at 25 because it is unlikely that an individual element would be this long (with transcription factor binding sites of 5–25 bp, and miRNAs of about 22 bp), and we still wanted to analyze the trends. The distribution of these areas is shown in Figure 1. As it can be seen from the picture, the layout of these regions was slightly indicative of the concentration at the anterior end, and some studies have indicated [14] that the mechanisms of regulation may be considerably different between groups of *Hox* genes, and that *cis*–acting elements are more likely to be found in the close proximity of anterior genes, with posterior ones being regulated in increasingly complex and spatially distant ways.

As it is difficult to see from the figure where these regions are exactly located, we have tabulated their distribution over several distinct genomic domains, including 5’ regulatory regions, exons, introns, 3’ sequences and intergenic sequences in Table 2. As we have mapped the *Hox* genes by the beginnings of their coding sequences, and in *Hox* they are always located in the first exon, the immediate 5’ sequence always contained the untranslated regions, with the promoter and the associated elements being more distant. Because of the varying sizes of the regions, their counts were not very informative, so we have measured the percentage of the columns contained in the regions with the mean less than 0.1 and shown the results in Table 2.

The layout of these columns is somewhat surprising. It shows the highest density not in gene exons, as expected, but at their immediate 5’ loci, normally containing the UTRs. This phenomenon has also been noted by other



```
322320:  GCAAATTCGGACCTTTCTTTTGCCCAGCTCAGCGTTACTCCATC---CCACTAATGAGGAAAATATGTATATACATATATATAATATATATTATATATAT  human
146404:  .........A..................................---.........................GTG.------------------------  baboon
252139:  ...T...-..G......C.C...T....------..CC..ATC.G---.........T...GC.GA.AA....----------------------------  cow
250806:  ...CGCG-..G......C.C.C.T....------..CC..ATC.GGGA.........TG.TGG..GGG.GG.T----------------------------  pig
190401:  ......GTA.G.T....C.CAA..A...-A.AA.GA.GG....T---.A.......T..AGGGTG.AA...CCCGGA-----------------------  mouse
 68393:  .......TA.G.T..CGC.CAA......-A.AA.GA.GG...CT---.A.......T..---.GGG..A.------------------------------  rat
```

Figure 2: A region from the *HoxA* cluster standing in a stark contrast with its environment. Solid line box encloses the area of perfect conservation in all species. Dashed and dotted lines show the areas of rodent and primate differential conservation, respectively. Dots indicate the same letter as in the row 1 (human).

studies, on other gene clusters [11]. The conservation density drops as one moves away from the genes, however the fact that the density is measured over regions of minimal length 25 is somewhat puzzling. One can argue that some of these regions actually represent clusters of regulatory elements, since not every included column is required to maintain the same high conservation rate, but the conservation is still too good for this scenario. Overall, from the high block conservation of the 5' UTRs and promoter regions one can hypothesize that some yet unknown mechanism protects these entire areas from too many mutations, imposing a much wider constraint on the sequence than just on the functional elements.

Surprisingly, no regions of high overall conservation have been found in the small part of the *HoxD* cluster we analyzed. Further inspection has shown that both the exons of these genes and the corresponding 5' sequences were indeed conserved, although not at the stringent 0.1 substitution average level. More puzzling was the concentration of high conservation in an apparently intergenic region, but they almost all lie at or around the site of the Ensembl gene prediction, providing additional evidence for its correctness. However, many highly conserved regions have been found in the intergenic regions of other *Hox* clusters, too. Some of them contain functional elements, although it is an open question why they are so long. Recent studies [18] have found several miRNA genes within *Hox* clusters, important for gene regulation at the post–transcriptional level, and distal regulatory sequences are common in the genome.

In addition to its capacity to identify general trends, our program is capable of finding small isolated regions when they stand in a contrast with their environment, as depicted in Figure 2. We have looked at the regions of minimal length 5 whose cumulative score was exceeding the mean of their environment. As expected, many such regions have been located, and we used the Student t test in order to estimate their significance. Roughly half were significant above the 0.99 level, again confirming the fundamental non-random nature of genomic sequences. Some of these regions were clustering, but there was only an occasional match with experimentally confirmed functional sites (dataset compiled from the literature by Laura Elnitski, unpublished). As these regions were more likely to stand out only in the areas of poor general conservation, we have not attempted to plot them on a chart similar to that of Figure 1.

4 DISCUSSION

In this paper we have presented an argument that the functional constraints on DNA sequences may be enforced by a mechanism broader than a simple prohibition of mutations within functional elements. The overall conservation patterns, both in the background and in the contiguous areas scoring better than the background indicate consistent good conservation in sequences upstream of the translation start sites, and often better than within the coding sequences themselves. In addition, a large number, if not a majority, of both long and short intervals that score better than their local environment do so with high significance.

In many respects, this is a work in progress. We still need to find a good way of integrating data from various background conservation levels and long and short outlier regions, along with their significance. There are existing tools, like the Multi–PipMaker [13] that perform similar tasks, and also provide an intuitive graphical representation. However, the PipMaker software works from gap–free pairwise alignments towards the integration into a multiple alignment, while our approach takes the other direction. In addition, PipMaker leaves the deduction of the significance to the user. However, the significance is the center-point of our approach, and any way of presenting the information must include it.

We would like to perform further systematic analysis by varying thresholds for the mean score of the conservation, and analyze the trends. In addition, our treatment of gaps in sequences, while practical, is not satisfactory. The uncritical inclusion of gaps often leads to artifacts, but their exclusion creates problems, too. No matter how uncomfortable they are to work with, gaps in alignments are presumed to reflect the natural process of nucleotide insertion and deletion, and as such they should be fully included in the analysis.

5 ACKNOWLEDGMENTS

The authors would like to thank Dr. Laura Elnitski, who has compiled a large data set of experimentally confirmed functional sites in *Hox* which we used for cross–comparisons with our data. A part of this work has been done while the authors were at the Whitehead Institute, when it had been financed from the Human Genome Project and other grants to WI Center for Genome Research. NS is currently supported by UTA REP grant 14 7487 62, and KD is partially supported by funds from Genome Canada and Genome Quebec.

REFERENCES

[1] J. Bentley. *Programming Pearls*. Addison–Wesley, 1986.

[2] Michael Brudno, Chuong B. Do, Gregory M. Cooper, Michael F. Kim, Eugene Davydov, Eric D. Green, Arend Sidow, and Serafim Batzoglou. Lagan and Multi–LAGAN: efficient tools for large–scale multiple alignment of genomic dna. *Genome Res.*, 13:721–731, 2003.

[3] L. Duret and P. Bucher. Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.*, 7:399–406, 1997.

[4] Ross Hardison. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.*, 16(9):369–372, 2000.

[5] G.S.B. Horan, E.N. Kovacs, R.R. Behringer, and M.S. Featherstone. Mutations in paralogous Hox genes result in overlapping homeotic transformations of the axial skeleton: evidence for unique and redundant function. *Dev. Biol.*, 169:359–372, 1995.

[6] T. Hubbard, D. Andrews, M. Caccamo *et al.* Ensembl 2005. *Nucleic Acids Res.*, 33:D447–D453, 2005.

[7] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[8] Jia Li and Webb Miller. Significance of inter–species matches when evolutionary rate varies. In *Proceeding of the Sixth Annual International Conference on Computational Biology*, pages 216–224. ACM Press, 2002.

[9] Webb Miller, K.D. Makova, A. Nekrutenko, and Ross C. Hardison. Comparative genomics. *Annu. Rev. Genomics Hum. Genet.*, 5:15–56, 2004.

[10] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.

[11] Monique Rijnkels, Laura Elnitski, Webb Miller, and Jeffrey M. Rosen. Multispecies comparative analysis of a mammalian–specific genomic domain encoding secretory proteins. *Genomics*, 82:417–432, 2003.

[12] D. Sankoff and R.J. Cedergren. Simultaneous comparison of three or more sequences related by a tree. In D. Sankoff and J.B. Kruskal, editors, *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*, pages 253–264. Addison–Wesley, 1983.

[13] S. Schwartz, Z. Zhang, K.A. Frazer, A. Smith, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. PipMaker — A web server for aligning two genomic DNA sequences. *Genome Res.*, 10:577–586, 2000.

[14] J. Sharpe, S. Nonchev, A. Gould, J. Whiting, and R. Krumlauf. Selectivity, sharing and competitive interactions in the regulation of Hoxb genes. *EMBO J.*, 17:1788–1798, 1998.

[15] N. Stojanovic, L. Florea, C. Riemer, D. Gumucio, J. Slightom, M. Goodman, W. Miller, and R. Hardison. Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Res.*, 27:3899–3910, 1999.

[16] The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306:636–640, 2004.

[17] J.W. Thomas, J.W. Touchman, R.W. Blakesley *et al.* Comparative analyses of multi–species sequences from targeted genomic regions. *Nature*, 424:788–793, 2003.

[18] S. Yekta, I.H. Shih, and David P. Bartel. Microrna–directed cleavage of HOXB8 mRNA. *Science*, 304:594–596, 2004.