Divergence of Conserved Non-Coding Sequences: Rate Estimates and Relative Rate Tests

Günter P. Wagner[†,*], Claudia Fried[‡], Sonja Prohaska[‡], and Peter F. Stadler[‡,§]

[†]Department of Ecology and Evolutionary Biology
Yale University, New Haven, CT 06405-8106, USA

[‡]Bioinformatik, Institut für Informatik, Universität Leipzig, Kreuzstraße 7b, D-04103 Leipzig, Germany

[§]Institut für Theoretische Chemie und Molekulare Strukturbiologie Universität Wien, Währingerstraße 17, A-1090 Wien, Austria

[*]Author for correspondence

Abstract.

In many eukaryotic genomes only a small fraction of the DNA codes for proteins but the non-protein coding DNA harbors important genetic elements directing the development and the physiology of the organisms, like promoters, enhancer, insulators and micro-RNA genes. The molecular evolution of these genetic elements is difficult to study because their functional significance is hard to deduce from sequence information alone. Here we propose an approach to the study of the rate of evolution of functional non-coding sequences at a macro-evolutionary scale. We identify functionally important non-coding sequences as Conserved Non-Coding Nucleotide (CNCN) sequences from the comparison of two outgroup species. The so identified CNCN sequences are then compared to their homologous sequences in a pair of ingroup species and monitor the degree of modification these sequences suffered in the two ingroup lineages. We propose a method to test for rate differences in the modification of CNCN sequences among the two ingroup lineages, as well as a method to estimate their rate of modification. We apply this method to the full sequences of the HoxA clusters from six gnathostome species: a shark, *Heterodontus francisci*, a basal ray finned fish, *Polypterus senegalus*, the amphibian, *Xenopus tropicalis*, as well as three mammalian species, human, rat and mouse. The results show that the evolution rate of CNCN sequences is not distinguishable among the three mammalian lineages, while the Xenopus lineage has a significantly increased rate of evolution. Furthermore the estimates of the rate parameters suggest that in the stem lineage of mammals the rate of CNCN sequence evolution was more than twice the rate observed within the placental mammal clade, suggesting a high rate of evolution of cis-regulatory elements during the origin of amniotes and mammals. We conclude that the proposed methods can be used for testing hypotheses about the rate and pattern of evolution of putative cis-regulatory elements.

1. Introduction

A major mode of developmental gene evolution is based on the modification of cis-regulatory elements (Arnone and Davidson, 1997; Carroll *et al.*, 2001; Davidson, 2001; Stern, 2000; Wray *et al.*, 2003). Binding sites for transcription factors are usually short and variable and are thus hard to identify unambiguously, in particular if the transcription factors involved are not known *a priori* (Tautz, 2000; Ludwig *et al.*, 2000; Dermitzakis *et al.*, 2003). Non-coding sequences, however, can contain islands of strongly conserved segments, so-called *phylogenetic footprints* (Tagle *et al.*, 1988). In a number of cases it has been shown that these phylogenetic footprints are indicative of functional cis-regulatory elements (Tagle *et al.*, 1988; Manen *et al.*, 1994; Leung *et al.*, 2000; Chiu *et al.*, 2002; Blanchette and Tompa, 2002; Santini *et al.*, 2003), reviewed by Duret and Bucher (1997) and Fickett and Wasserman (2000). Hence it is possible in principle to gain insights into the extent and the phylogenetic timing of major changes in the cis-regulatory elements of a gene by studying the phylogenetic pattern of non-coding sequence conservation. In a recent study we have presented an efficient computational tool, the `tracker` program, to simultaneously survey the orthologous intergenic regions in multiple large gene clusters (Prohaska *et al.*, 2004a). This technique allowed us to demonstrate that footprint patterns contain sufficient phylogenetic information e.g. to resolve the orthology of shark and mammalian *Hox* clusters (Prohaska *et al.*, 2004b).

The quantitative analysis of dynamical aspects of footprint loss and acquisition, however, is complicated by the fact that we cannot independently observe individual regulatory DNA regions. Instead, phylogenetic footprinting always detects regulatory elements in pairs of sequences. As a consequence, even very simplistic models of footprint loss lead to rather sophisticated inference and test methods as we shall see in this contribution. We will focus on two questions: (i) How can we detect rate differences in footprint modification in two different lineages? (ii) Can we determine periods in evolutions with exceptionally large or small footprint loss?

2. Data Acquisition

Sequence data of *HoxA* clusters were downloaded from Genbank: *Homo sapiens* HsA = reverse complement (r.c.) of AC004080, ***AC010990*** r.c. (overlaps 200nt with ***AC004080***), and ***AC004079*** (pos. 75001-end, r.c., overlaps 200nt with ***AC010990***), as in Chiu *et al.* (2002); *Heterodontus francisci* HfM = ***AF479755***; *Polypterus senegalus* PsA = ***AC132195*** and ***AC12632*** as in Chiu *et al.* (2004); *Mus musculus* MmA ***NT_039343*** r.c.; *Rattus norvegicus* Rn = ***NW_043751***; *Xenopus tropicalis* XtA = ***AC145789*** (downloaded 14/Aug/2003).

Conserved non-coding sequences are detected using the `tracker` program (Prohaska *et al.*, 2004a). Very briefly, this approach is based on `BLAST` (Altschul *et al.*, 1990) for the initial search of all pairs of input sequences restricted to homologous intergenic regions. The resulting list of pairwise sequence alignments is then assembled into groups of partially overlapping regions that are subsequently passed through several filtering steps and finally aligned using the segment based multiple alignment tool

DIALIGN2 Morgenstern (1999). The final output of the program is the list of these aligned “footprint cliques”, see Electronic Supplement[1].

The alignments of all footprint cliques are concatenated and padded with gap characters where data are missing, i.e., where a footprint detected between some sequences does not have a counterpart in others. Consequently, all gap characters are treated as “missing data” in the further analysis, i.e., as unknown nucleotides rather than as deletion. Conserved positions between groups of sequences are counted as specified in equ.(1) below. In order to take missing data into account we discount columns with gaps in the relevant sequences by a factor of 1/4 for each gap, data are summarized in Tables 1 and 2.

3. A Model

Consider the tree in Fig. 1. Suppose we have a set CNCN-positions Ω of footprint cliques at $Q = \text{lca}(X(AB))$ and set $q = |\Omega|$. We assume that CNCN are lost according to a simple exponential decay law. Furthermore, suppose the rate λ is everywhere the same, with exception of the lineage from $P = \text{lca}(AB)$ to B.

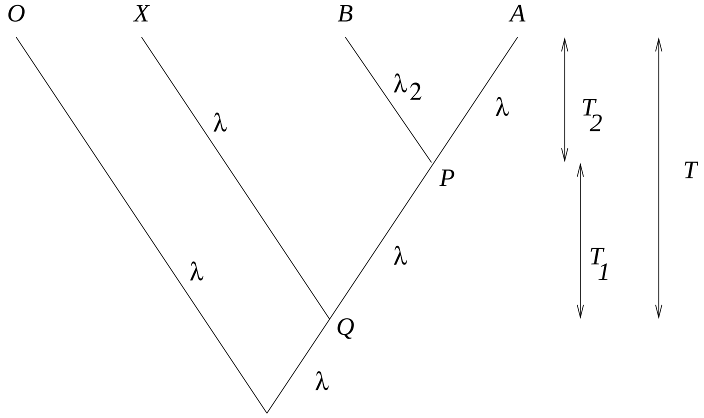

Figure 1. Each test for the rate of change in CNCN sequences is based on the comparison of four sequences O, X, A, and B. O and X are outgroup sequences, which serve for the detection of conserved non-coding sequences. The additive evolutionary distance between O and X is assumed to be long enough to randomize sequences which are not under stabilizing selection. Following Tagle *et al.* (1988) we only accept outgroup sequences with at least 250 Mio years of additive evolutionary time between them. A and B are the two ingroup sequences and Q is the most recent common ancestor of X and (A, B) that existed at a time T, and P is the most recent common ancestor of A and B which existed as a time T_2 before the present. We test whether the rate of modification along the branch $P - B$ is different from that along the branch $P - A$, where it is assumed that the rate of the evolution along $P - A$ is the same as in the rest of this tree.

[1] URL: http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/04-007/.

Given an outgroup O we may consider all those CNCN that appear in O and in at least one of the three species A, B, and X. The measurable parameters are then

$$\begin{aligned}
c_{XA} &= & |O \cap X \cap A| & = |(O \cap X) \cap (O \cap A)| \\
c_{XB} &= & |O \cap X \cap B| & = |(O \cap X) \cap (O \cap B)| \\
c_{AB} &= & |O \cap X \cap A \cap B| & = |(O \cap X) \cap (O \cap A) \cap (O \cap B)| \\
c_{A\vee B} &= & |O \cap X \cap (A \cup B)| & = c_{XA} + c_{XB} - c_{AB} \\
u &= & |O \cap (X \cup A \cup B)| & = |(O \cap X) \cup (O \cap A) \cup (O \cap B)|
\end{aligned} \tag{1}$$

Given the model in Fig. 1 we can readily express the observable footprint counts in terms of the model parameters

$$\begin{aligned}
c_{XA}/q &= e^{-2\lambda T} \\
c_{AB}/q &= e^{-\lambda T} e^{-\lambda T_1} e^{-\lambda T_2} e^{-\lambda_2 T_2} = e^{-2\lambda T} e^{-\lambda_2 T_2} \\
c_{XB}/q &= e^{-\lambda T} e^{-\lambda T_1} e^{-\lambda_2 T_2} \\
u/q &= 1 - (1 - e^{-\lambda T}) \left[(1 - e^{-\lambda T_1}) + e^{-\lambda T_1} (1 - e^{-\lambda T_2})(1 - e^{-\lambda_2 T_2}) \right]
\end{aligned} \tag{2}$$

A short computation yields

$$\begin{aligned}
e^{-\lambda T} &= \sqrt{c_{XA}/q} \\
e^{-\lambda_2 T_2} &= \frac{c_{AB}}{c_{XA}} \\
e^{-\lambda T_2} &= \frac{c_{AB}}{c_{XB}} \\
e^{-\lambda T_1} &= \frac{c_{XB}}{c_{AB}} \sqrt{c_{XA}/q}
\end{aligned} \tag{3}$$

The last line of equ.(2) then becomes

$$\frac{u}{q} = 2\sqrt{\frac{c_{XA}}{q}} + \frac{c_{XB}}{c_{XA}} \sqrt{\frac{c_{XA}}{q}} - \frac{c_{AB}}{c_{XA}} \sqrt{\frac{c_{XA}}{q}} - \frac{c_{XA}}{q} - \frac{c_{XB}}{q} + \frac{c_{AB}}{q} \tag{4}$$

Multiplying with q and solving for $\sqrt{q}$ yields

$$q = c_{XA} \left(\frac{u + c_{A\vee B}}{c_{XA} + c_{A\vee B}} \right)^2 \tag{5}$$

The variance of an exponential process with decay constant a and initial value b is

$$s^2 = b e^{-at} (1 - e^{-at}) \tag{6}$$

We are interested in the variance σ^2 of the difference of the loss rates along PA and PB, which equals twice the variance of the exponential process along one of the lineages. Thus

$$\begin{aligned}
\sigma^2 &= 2q e^{-\lambda T_1} e^{-\lambda T_2} \left(1 - e^{-\lambda T_2}\right) = 2q e^{-\lambda T} \left(1 - e^{-\lambda T_2}\right) \\
&= 2\sqrt{c_{XA} q} \left(1 - \frac{c_{AB}}{c_{XB}}\right) = 2 c_{XA} \left(\frac{u + c_{A\vee B}}{c_{XA} + c_{A\vee B}} \right) \left(1 - \frac{c_{AB}}{c_{XB}}\right)
\end{aligned} \tag{7}$$

The number of CNCN exclusively lost along PA is $m'_A = c_{A \vee B} - c_{XB}$; for PB we have $m'_B = c_{A \vee B} - c_{XA}$. Thus $m'_A - m'_B = c_{XA} - c_{XB}$. The test statistics is therefore

$$z = \frac{c_{XA} - c_{XB}}{\sigma} \tag{8}$$

Equation (8) gives a test statistic which assumes that the loss of conservation at each nucleotide position is stochastically independent. This assumption, however, is not plausible assuming that the elementary event in the evolution of an enhancer is the loss or gain of a transcription factor binding site. Typically, transcription factor binding sites are between 5 and 20 nucleotide positions long, but have various degrees of degeneracy. Evolutionary changes in the number and kind of transcription factor binding sites thus induces a stochastic dependency among the nucleotide positions compared here. To account for this stochastic dependency we scale the predicted sampling variance with the average length of contiguous CNCN sequence elements in our data, $\bar{\ell}$. This value is typically between $4 \leq \bar{\ell} \leq 6$ and thus at the same scale as many known transcription factor binding motives. The resulting test statistic then is

$$z' = z/\sqrt{\bar{\ell}} \tag{9}$$

which in normally distributed with variance 1.

4. Estimating Footprint Loss Rates

As outlined in the previous section the number of shared and unique CNCN positions among the four taxa X, O, A and B can be interpreted in terms of the parameters of an exponential loss model. In particular it is possible to derive expressions for λT and λT_2. If, in addition, we have independent estimates for the time of divergence of the taxa compared we could, from all possible four taxa comparisons, estimate the loss rate λ along the lineage from the most recent common ancestor $\mathrm{lca}(A, B)$ of A and B and one of the two taxa, A or B. While this exercise is computationally straightforward, the interpretation of the so obtained numbers needs careful attention to estimation biases.

One problem with the raw estimates obtained from solving the equations of the model for the parameters λT is that the CNCN detected in the comparison between the two outgroup species, O and X, contain spurious CNCN positions. That are nucleotide positions which are identical between the sequences of O and X but are only identical by chance rather than due to purifying selection. While `tracker` and other alignment procedures are designed to identify significant stretches of conserved sequence there is a possibility that at the borders of conserved sequence blocks spurious sites are included in the count of CNCN sites. There is no objective way to eliminate them from the sequence alignment, but it is possible to determine their influence on the estimates of the rate parameters.

Let $\lambda_o(T)$ be the rate parameter observed from a comparison in which the most recent common ancestor of A and B lived T years before the present, and let us assume the loss of CNCN positions is time homogeneous. Then this estimated rate is determined by the true rate λ_c as well as by the number of spurious CNCN positions. The true CNCN position evolve at a rate λ_c, but the spurious sites randomize much

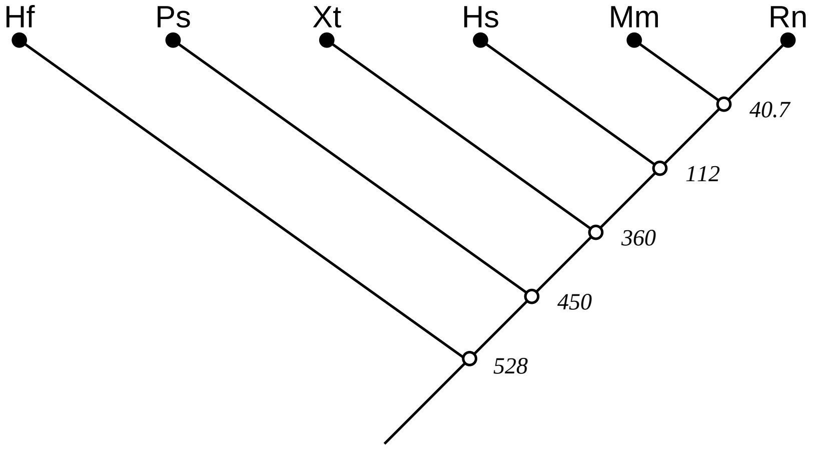

Figure 2. Phylogenetic tree of the taxa used in this study and their divergence times in Myr (Kumar and Hedges, 1998). Abbreviations:
Hf *Heterodontus francisci*, Ps *Polypterus senegalus*, Xt *Xenopus tropicalis*, Hs *Homo sapiens*, Mm *Mus musculus*, Rn *Rattus norvegicus*

quicker than the true CNCN sites. Over the timescales we consider in this paper these spurious positions randomize instantaneously, and thus contribute an additive term to the true rate to give the observed rate

$$\lambda_o(T)T = \lambda_c T + C\,. \tag{10}$$

Hence the observed rate λ_o is predicted to be a linear function of $1/T$ with a slope which depends on the logarithm of the fraction of spurious CNCN, and an intercept equal to an estimate of the true rate λ_c:

$$\lambda_o(T) = \lambda_c + C/T\,. \tag{11}$$

That means that if we have λ estimates for at least two time points we can do a linear regression of the observed rate parameters. The intercept is then a corrected rate estimate λ_c, and the slope an estimate of the fraction of spurious CNCN sites in the alignment of sequences from O and X.

5. The *HoxA* Clusters of Gnathostomes

We applied the method described above to a data set containing the full *HoxA* cluster sequences of three mammal species, human, rat and mouse, as well as an amphibian, *Xenopus tropicalis*, the basal ray finned fish bichir, *Polypterus senegalus* (Chiu *et al.*, 2004), as well as the shark *Heterodontus francisci* (Kim *et al.*, 2000; Chiu *et al.*, 2002), Fig. 2. In Tab. 1 a subset of the tests done for this data set is presented.

The comparison of the three mammalian species shows that the rate of modification of CNCN positions is similar, leading to the retention, r, of about 35% of the CNCN detectable in the outgroup species. The CNCN retention rate is the same whether shark and bichir are used as outgroups or bichir and frog. The z' statistic for differences in the rate of modification of CNCN is between 0.32 and 0.97 and are all far from significant.

Table 1. Summary of relative rate tests for the rate of modification of ancestrally conserved non-coding sequences. The taxa compared are indicated in the columns O, X, A, and B; q is the estimated total length of CNCN positions in the most recent common ancestor of X and (A, B); c_{XA} and c_{XB} are the numbers of CNCN positions shared between O, X, and A or B, respectively. The $r(A)$ and $r(B)$ values are the fraction of CNCN sites still conserved in A or B. z' is the test statistic and $P(z')$ is the two tailed type one error rate for rejecting the null hypothesis that $\lambda = \lambda_2$, based on the standard normal distribution. The average length of contiguous conserved sequences is $\bar{\ell} = 5.11$ with O = HfM and X = PsA and $\bar{\ell} = 4.48$ with O = PsA and X = XtA.

O	X	A	B	q	c_{XA}	$r(A)$	c_{XB}	$r(B)$	z'	$P(z')$
HfM	PsA	MmA	RnA	10668.87	3760.50	0.352	3746.00	0.351	0.238	0.810
HfM	PsA	HsA	MmA	10723.63	3808.50	0.355	3760.50	0.351	0.761	0.446
HfM	PsA	HsA	RnA	10750.50	3808.50	0.354	3746.00	0.348	0.995	0.318
PsA	XtA	MmA	RnA	10168.96	3599.75	0.354	3554.50	0.350	0.934	0.352
PsA	XtA	HsA	MmA	10124.07	3601.75	0.356	3599.75	0.356	0.035	0.968
PsA	XtA	HsA	RnA	10046.03	3628.75	0.361	3593.50	0.358	0.609	0.542
HfM	PsA	HsA	XtA	10843.97	3816.75	0.352	3576.25	0.330	2.535	0.011
HfM	PsA	RnA	XtA	10785.83	3757.50	0.348	3576.25	0.331	1.815	0.067
HfM	PsA	MmA	XtA	10804.83	3760.50	0.348	3559.25	0.329	2.064	0.039

The comparison of the Xenopus sequence with the three mammalian data sets shows that the rate of modification of CNCN in the Xenopus lineage is higher than in the mammalian lineage. The Xenopus lineage retains about 33% of the CNCN detected in the comparison of shark and bichir sequences, while the mammalian lineages retain about 35%. All these differences are significant with the comparison between Xenopus and human being significant at the 0.011 level, the comparison with mouse at the 0.039 level, while the comparison with rat is marginally significant at the 0.067 level. Hence it seems that the Xenopus lineage experiences a higher rate of modifications of CNCN positions than the mammalian lineage.

The results from the new test were compared to the Tajima relative rate test (Tajima, 1993), which can also be applied to the kind of data analyzed here (see Appendix). In Table 2 the results for the Tajima test of the same data as in Table 1 are summarized. The results are consistent with the ones from the z'-statistic (Table 1), confirming that the Xenopus lineage evolves faster than the mammalian.

None of the comparisons of mammalian *HoxA* clusters is significant but all the comparisons between Xenopus and the mammals are significant at least at the 5% level.

The rate parameter estimated from the model for the three different mammalian lineages vary depending on the outgroup taxa used. The rate parameters are consistently smaller the more distant the most recent common ancestor of the compared taxa is. This effect is anticipated based on the arguments put forward in section 4. The problem is that the comparison of the two outgroup species O and X will identify a number of spurious CNCN, which are identical in O and X due to chance. These CNCN then enter the estimation of the rate of evolution since the most recent common ancestor of A and B and inflate the rate estimate.

Table 2. Same data as in Table 1 but analyzed with the Tajima relative rate test (see Appendix). The Tajima statistics contrasts the difference between c_{XA} and c_{XB} in a χ^2-statistic with one degree of freedom. Note that the results are qualitatively consistent with those of the new test. None of the comparisons of mammalian *HoxA* clusters is significant but all the comparisons between Xenopus and the mammals are significant at least at the 5% level.

O	X	A	B	c_{AB}	c_{XA}	c_{XB}	Tajima	P
HfM	PsA	MmA	RnA	3531.69	3760.50	3746.00	0.0928	> 0.7
HfM	PsA	HsA	MmA	3531.00	3808.50	3760.50	0.8885	> 0.3
HfM	PsA	HsA	RnA	3519.81	3808.50	3746.00	1.4868	> 0.1
PsA	XtA	MmA	RnA	3400.62	3599.75	3554.50	1.2947	> 0.1
PsA	XtA	HsA	MmA	3387.06	3601.75	3599.75	0.0021	> 0.9
PsA	XtA	HsA	RnA	3347.38	3601.75	3554.50	1.0799	> 0.3
HfM	PsA	MmA	XtA	3039.62	3760.50	3559.25	6.3892	< 0.05
HfM	PsA	RnA	XtA	3014.62	3746.00	3559.25	5.3487	< 0.05
HfM	PsA	HsA	XtA	3074.44	3808.50	3559.25	9.9745	> 0.01

To correct for this effect we performed a linear regression of rate estimates over the inverse of the time T since the most recent common ancestor of A and B. First we analyzed the rate estimates for the mammalian data with all possible combinations of outgroup species. The intercept was 0.218, but the data revealed a deviation from linearity in the plot of the residuals over $1/T$. The regressions were thus repeated for data points using either only more distant (360 and 112 Mio years) or only the more recent most recent common ancestors (112 and 40.7 Mio years). The rate estimates are 0.153 ± 0.071 for the more recent time points and 0.378 ± 0.067 for the more distant time points. These results suggest that there is systematic rate variation in the evolution of mammalian lineages such that the rate of modification of CNCN is considerably higher in the stem lineage of amniotes and mammals than among placental mammals.

The slope of the regression equation (11) over $1/T$ allows an estimate of the fraction of spurious CNCN positions. These values suggest that only between 2 and 5% of the CNCN entering these calculations are spurious and thus do not greatly affect the variance used in calculating the z' statistic for the relative rate test.

6. Discussion

In this paper we describe a method for detecting rate heterogeneity in the evolution of putative cis-regulatory elements. Rate heterogeneity can be detected both between two lineages as well as along the same lineage over different time frames. The approach detects putative cis-regulatory elements through their conservation among two outgroup species and records the rate of modification of CNCN sequences along two ingroup lineages. This approach comes with advantages as well as disadvantages. The advantage being that one does not have to rely on notoriously noisy predictions of transcription factor binding sites to assess the presence of cis-regulatory sites. The phylogenetic conservation of non-coding sequences is taken as evidence for the functional importance of non-coding DNA sequences (Tagle *et al.*, 1988; Manen *et al.*, 1994; Duret and Bucher, 1997; Leung *et al.*, 2000; Fickett and Wasserman, 2000; Chiu *et al.*, 2002; Blanchette and Tompa, 2002; Santini *et al.*, 2003; Ghanem *et al.*, 2003). The disadvantage of this approach is that it is known that functionally conserved cis-regulatory elements can quickly loose their sequence similarity and would thus not be detectable as conserved non-coding sequences (Ludwig, 2002; Phinchongsakuldit *et al.*, 2004). On the other hand, there are examples of functionally conserved enhancers which also retain sequence conservation over long evolutionary distances (e.g. (Shashikant *et al.*, 1998)). The reasons for the differences of sequence conservation of functionally conserved cis-regulatory elements is unknown but may be related to such general factors as population size and mutation rate (Carter and Wagner, 2002). We thus propose that the method presented in this paper should be primarily used in a hypothesis testing framework. Below we outline a few scenarios in which the proposed test might be useful.

One situation in which the method might be useful is to test the following hypothesis. It is plausible that the adaptation of a gene to a new function is not limited to the coding region of the gene, but also affects the cis-regulatory elements determining the location, timing and the level of expression. While it is relatively routine to detect selection in coding regions (Liberles *et al.*, 2001), adaptive evolution of cis-regulatory elements is hard to detect in general, but see (Kohn *et al.*, 2004). But one may test the following hypothesis: if the coding regions of a group of genes is under directional selection in one lineage, say B, but not in another lineage, say A, then the cis-regulatory elements will also evolve quicker in lineage B than in lineage A. This hypothesis could be tested by comparing the rate of modification of CNCN sequences in the these two lineages.

Another hypothesis testable by the proposed approach is that cis-regulatory elements of duplicated genes diverge asymmetrically, i.e. that one of the duplicates diverges faster than the other. This has been shown to be the case for coding sequences (e.g. (Wagner, 2002)), but has to our knowledge not been demonstrated for cis-regulatory elements. Another hypothesis is that putative cis-regulatory elements evolve faster when the expression patterns of the genes in the same genomic region undergo evolution. A limited result along these lines has been presented in the example data set analyzed in this paper, i.e. the *HoxA* cluster sequences of gnathostomes. The results suggest that, in the stem lineage of mammals and amniotes, the rate of CNCN sequence evolution is more than twice as high than among the placental mammals, human, mouse and rat. This result is preliminary due to limited taxon sampling, but is consistent with the idea that body-plan evolution involves major re-wiring of transcriptional regulation of developmental genes (Davidson, 2001).

The usefulness of the proposed method strongly depends on the extent of taxon sampling. The example data set analyzed for this paper consists of the complete sequences of *HoxA* clusters from six species. The continuing efforts to sequence the genomes from representatives of major clades will certainly increase the number of taxa that can be included in a comparative study of their non-coding sequences. Data sets from many different species will have considerable statistical and cladistic power if analyzed with appropriate statistical tools.

Acknowledgement

This work supported by NSF grant INB-0321470 to GPW and the DFG Bioinformatics Initiative BIZ-6/1-2.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

Arnone MI, Davidson EH, 1997. The hardwiring of development: Organization and function of genomic regulatory systems. Development 124:1851–1864.

Blanchette M, Tompa M, 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Research 12:739–748.

Carroll SB, Grenier JK, Weatherbee SD, 2001. From DNA to Diversity. Malden, MA: Blackwell Science.

Carter AJ, Wagner GP, 2002. Evolution of functionally conserved enhancers can be accelerated in large populations: a population-genetic model. Proc R Soc Lond B Biol Sci 269:953–960.

Chiu Ch, Amemiya C, Dewar K, Kim CB, Ruddle FH, Wagner GP, 2002. Molecular evolution of the HoxA cluster in the three major gnathostome lineages. Proc Natl Acad Sci USA 99:5492–5497.

Chiu CH, Dewar K, Wagner GP, Takahashi K, Ruddle F, Ledje C, Bartsch P, Scemama JL, Stellwag E, Fried C, Prohaska SJ, Stadler PF, Amemiya CT, 2004. Bichir *HoxA* cluster sequence reveals surprising trends in rayfinned fish genomic evolution. Genome Res 14:11–17.

Davidson E, 2001. Genomic Regulatory Systems. San Diego: Academic Press.

Dermitzakis ET, Bergman CM, Clark AG, 2003. Tracing the evolutionary history of *Drosophila* regulatory regions with models that identify transcription factor binding sites. Mol Biol Evol 20:703–714.

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

Fickett JW, Wasserman WW, 2000. Discovery and modeling of transcriptional regulatory regions. Current Opinion in Biotech 11:19–24.

Ghanem N, Jarinova O, Amores A, Qiaoming L, Hatch G, Park BK, Rubenstein JLR, Ekker M, 2003. Regulatory roles of conserved intergenic domains in vertebrate *Dlx* bigene clusters. Genome Res 13:533–543.

Kim CB, Amemiya C, Bailey W, Kawasaki K, Mezey J, Miller W, Minosima S, Shimizu N, P. WG, Ruddle F, 2000. Hox cluster genomics in the horn shark, *heterodontus francisci*. Proc Natl Acad Sci USA 97:1655–1660.

Kohn MH, Fang S, Wu CI, 2004. Inference of positive and negative selection on the 5' regulatory regions of drosophila genes. Mol Biol Evol Dec 5 2003 [Epub ahead of print].

Kumar S, Hedges B, 1998. A molecular timescale for vertebrate evolution. Nature 392:917–920.

Leung JY, McKenzie FE, Uglialoro AM, Flores-Villanueva PO, Sorkin BC, Yunis EJ, Hartl DL, Goldfeld AE, 2000. Identification of phylogenetic footprints in primate tumor necrosis factor-α promoters. Proc Natl Acad Sci USA 97:6614–6618.

Liberles DA, Schreiber DR, Govindarajan S, Chamberlin S, Benner SA, 2001. The adaptive evolution database (TAED). Genome Biol 2:Research0028.

Ludwig MZ, 2002. Functional evolution of noncoding DNA. Curr Op Genet Devel 12:634–639.

Ludwig MZ, Bergman C, Patel NH, Kreitman M, 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403:564–567.

Manen J, Savolainen V, Simon P, 1994. The *atpB* and *rbcL* promoters in plastid DNAs of a wide dicot range. J Mol Evol 38:577–582.

Morgenstern B, 1999. `DIALIGN 2`: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15:211–218.

Phinchongsakuldit J, MacArthur S, Brookfield JFY, 2004. Evolution of developmental genes: Molecular microevolution of enhancer sequences at the *Ubx* locus in *Drosophila* and its impact on developmental phenotypes. Mol Biol Evol 21:348–363.

Prohaska S, Fried C, Flamm C, Wagner G, Stadler PF, 2004a. Surveying phylogenetic footprints in large gene clusters: Applications to Hox cluster duplications. Mol Phyl Evol 31:doi:10.1016/j.ympev.2003.08.009. In press.

Prohaska SJ, Fried C, Amemiya CT, Ruddle FH, Wagner GP, Stadler PF, 2004b. The shark HoxN cluster is homologous to the human HoxD cluster. J Mol Evol 58:212–217.

Santini S, Boore JL, Meyer A, 2003. Evolutionary conservation of regulatory elements in vertebrate *Hox* gene clusters. Genome Res 13:1111–1122.

Shashikant C, Kim CB, Borbley MA, Wang WC, Ruddle FH, 1998. Comparative studies on mammalian *Hoxc8* early enhancer sequence reveal a baleen whale-specific deletion of a cis-acting element. Proc Natl Acad Sci 95:15446–15451.

Stern DL, 2000. Evolutionary developmental biology and the problem of variation. Evolution 54:1079–1091.

Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT, 1988. Embryonic epsilon and gamma globin genes of a prosimian primate (galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 203:439–455.

Tajima F, 1993. Simple methods for testing molecular clock hypothesis. Genetics 135:599–607.

Tautz D, 2000. Evolution of transcriptional regulation. Curr Opin Genet Dev 10:575–579.

Wagner A, 2002. Asymmetric functional divergence of duplicate genes in yeast. Mol Biol Evol 19:1760–1768.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA, 2003. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20:1377–1419.

Appendix

Tajima's relative rate test (Tajima, 1993) concerns the rates of evolutions along the terminal edges PA and PB where $P = \mathrm{lca}(A, B)$. In order to measure the rate of CNCN loss between P and A, however, we need an outgroup X since only the numbers defined in equ. (1) can be obtained directly from the data. Let

$$\begin{aligned} m_A &= c_{A\vee B} - c_{XA} = c_{XB} - c_{AB} \\ m_B &= c_{A\vee B} - c_{XB} = c_{XA} - c_{AB} \end{aligned} \tag{12}$$

be the numbers of CNCN that are present in X and are lost along PA but not along PB, and *vice versa.* The original Tajima statistics assumes that each residue compared is stochastically independent, which is not likely in the case of loss of conservation in putative cis-regulatory sequences, since the elementary evolutionary event is likely the loss of a transcription factor binding site. We thus correct for stochastic dependency in the same way as we did for the z'-statistic proposed in this paper by dividing the χ^2-value, which is a variance, by the average length of contiguous conserved sequences. In order to test whether m_A and m_B are significantly different we therefore consider the test statistics

$$\chi^2 = \frac{(m_A - m_B)^2}{(m_A + m_B)\bar{\ell}}\,. \tag{13}$$

Since we have a single degree of freedom χ^2 is significant at 95% level of $\chi^2 > 3.841$.