Update TRENDS in Genetics Vol.21 No.10 October 2005

Research Focus

The evolution of noncoding DNA: how much junk, how much func?

Cristian I. Castillo-Davis

Department of Statistics, Harvard University, Cambridge, MA 02138, USA

Comparative sequence analysis on a genomic scale has opened the door for the systematic analysis of *cis*-acting regulatory DNA. It is now possible to begin to answer basic questions such as, how much meaningful noncoding sequence is in the genome? How strong is natural selection on functional noncoding sequences in different species? Two recent articles have capitalized on the comparative genomic approach in an attempt to answer these questions with surprising results.

Introduction

Semantics, the study of meaning, has a long tradition in the humanities and asks the basic question, What is the relationship between symbolic representation, often language, and meaning? This question is currently being pursued in more material terms within biology with respect to the relationship between noncoding DNA and biological meaning. This question is a timely one because it is becoming increasingly clear that changes in gene expression – mediated primarily by changes in noncoding DNA – have a tremendous impact on organism phenotype. These include effects on drug response [1], disease susceptibility [2–4] and the evolution of novelties between species [5]. In particular, regions of noncoding DNA that bind to transcription factors that act to repress or activate gene expression, called *cis*-acting regulatory sequences (see Glossary), are thought to have a specifically important role in affecting transcriptional regulation. For example, it has been recently shown that the majority of genome-wide differences in gene expression between species appear to be due to changes in *cis*- not *trans*-acting factors [6] and, in humans, it is thought that variation in *cis*-regulatory sequences and not protein-coding sequences might underlie many complex, non-mendelian, diseases [7].

Despite the importance of *cis*-acting elements such as transcription-factor-binding sites (TFBS) and other DNA-binding sites, few are well-characterized owing to the traditionally laborious methods required to study them. Standard approaches such as deletion analysis and comprehensive mutant analysis are time consuming [8] and can usually be performed only one gene at a time. In addition, because such experiments are often performed under only one condition, it is difficult to be certain that even exhaustive functional study of a particular noncoding region is truly comprehensive. For example, if a given transcription factor (TF) binds to a specific noncoding site only under a particular condition (e.g. stress) an experimental assay, even a high-throughput one, performed under normal conditions will probably not identify the binding site [9].

For these reasons, it has been extremely difficult to generate general estimates about how much functional regulatory sequence is contained in different genomes by extrapolating from experimental studies. Even the most basic parameters of *cis*-regulatory control sequences are largely unknown. For example, what fraction of a typical 'promoter' is functional? How far do regulatory regions extend both upstream and downstream of a gene?

Two new studies have cleverly sidestepped these difficult problems by using available genome sequences of several closely related species and evolutionary theory to acquire estimates of the amount of functional regulatory DNA in noncoding sequences. What is exciting about these studies is that they demonstrate that functional noncoding DNA can be revealed without experimental intervention, relying instead on the fact that evolution has already carried out the requisite mutagenesis experiments – and under different natural conditions. Moreover, the results of these studies suggest how future experimental analyses can be improved, moving us towards a better understanding of transcriptional regulation. After an outline of some of the methods and major results of each article, I will discuss the conclusions of each in the context of future comparative evolutionary genomic experiments, and some exciting discoveries that might lie in the near future.

Glossary

***cis*- versus *trans*-acting elements or factors:** in the context of gene regulation, these terms originate from early experiments where certain sequences could restore normal gene expression when placed in a plasmid (e.g. those coding for a transcription factor), whereas others (e.g. transcription-factor-binding sites) did not. Because the TFBS sequences had to be *in cis* (physically on the same chromosome) as the gene being studied to confer function they were called *cis*-acting elements. Elements that were 'diffusible' (i.e. TFs) were henceforth called *trans*-acting factors.

N_e (effective population size): most population genetics theory has been developed using the simplifying assumption that population sizes are large in size and that individuals mate randomly. However, real populations often violate these assumptions; for example, not all individuals might be of reproductive age, the ratio of males to females might not be equal and populations might be structured unevenly. Mathematical adjustment of the census population size to its equivalent size in a 'well-behaved' population (that can be used as input for standard models) is known as the effective population size or N_e.

Corresponding author: Castillo-Davis, C.I. (ccastillo-davis@stat.harvard.edu).
Available online 11 August 2005

Conservation and degradation in noncoding DNA

Taking the comparative evolutionary approach discussed earlier, Chin and colleagues [10] attempted to asses the complexity of noncoding sequences in the genome of *S. cerevisiae* (baker's yeast). To estimate the level of functional noncoding sequence in putative regulatory regions, the authors first sought to establish the rate of neutral nucleotide substitution in the genome by examining the number of substitutions that have occurred in fourfold silent sites between four closely related yeast species (excluding genes with strong codon bias). Fourfold silent sites refer to those third bases of codons in a protein where a nucleotide substitution will not change the amino-acid specified (i.e. they are mutationally 'silent' with respect to protein sequence and are likely to be selectively neutral).

Armed with these data, the authors first ruled out any spatial correlation in mutation rates to insure that their measure could be applied uniformly across the genome. Next, they built a hidden Markov model (HMM) that enabled them to infer, probabilistically, blocks of high conservation and low conservation in the promoter regions of the four species alignments. According to the HMM, regions of high conservation accounted for ~34% of the total promoter sequence analyzed – ~90 bp total for a typical promoter, mostly made up of 8–20-bp blocks. Because typical TFBS are on the order of 8–30 bp, this provides a reasonable upper limit of approximately ten TFBS per promoter region per gene. Interestingly, this estimate of the amount of functional noncoding DNA in yeast promoters is somewhat comparable to that found in mammals with different methods to which I now turn.

Using a similar yet distinct comparative genomic approach, Keightley and colleagues [11] focused on the fraction of conserved noncoding sequences 6-kb upstream (5′) and downstream (3′) of protein-coding genes (80% from translation start and stop sequences) in two different pairs of mammals: human and chimpanzee in addition to mouse and rat. To measure the level of 'selective constraint' in these regions, the authors compared observed with expected rates of nucleotide substitution using expected rates of sequence change derived from fast-evolving intronic sequences. Similar to fourfold silent sites, these intronic sequences were assumed to be evolving neutrally between species and served as a benchmark for the evolution of 5′ and 3′ noncoding DNA between species. The first intron and domains flanking splice sites were excluded because these are known to contain functional elements.

When Keightley *et al.* examined levels of selective constraint in noncoding regions (averaged over all genes), several striking patterns emerged (Figure 1). First, levels of selective constraint appeared to drop precipitously after ~2 kb, upstream and downstream of genes in both the hominid and murid comparisons. This suggests that many functional *cis*-regulatory elements are likely to be concentrated in a 2-kb window surrounding most genes. Although this important result does not add to our understanding of long-range regulatory elements that act hundreds of kilobases from their targets, it does help to define more objectively the limits of the basic 'regulatory promoter' region in eukaryotes. Because 5′ and 3′ untranslated regions (UTRs) are included in the regions analyzed, purifying selection observed in this study might not be limited to regulatory motifs but might also include other factors such as mRNA stability.

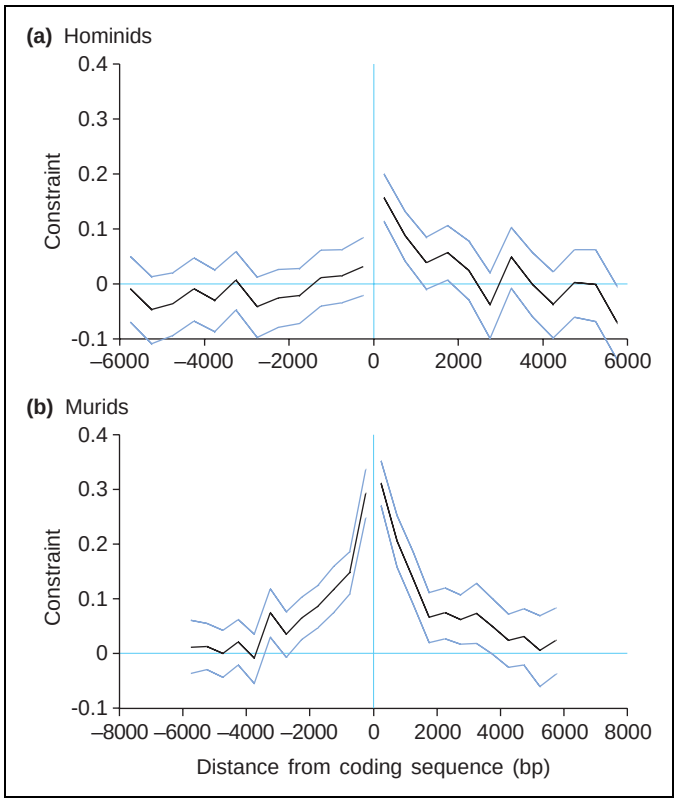

Figure 1. Average selective constraints in **(a)** hominid (human and chimpanzee) and **(b)** murid (mouse and rat) noncoding sequences. These were measured 6-kb upstream and downstream of protein-coding genes (*x*-axis) using the method of Keightley *et al.* [11]. Increasing constraint, shown on the *y*-axis, indicates increased conservation. Grey lines indicate bootstrap 95% confidence intervals. Reproduced, with permission, from Ref. [11].

Surprisingly, Keightley *et al.* also found that the level of selective constraint differed by at least an order of magnitude in the two lineages, with hominids exhibiting less selective constraint in noncoding regions compared with murids. Not only were selective constraints lower in hominids, they were for the most part absent. This led the authors to conclude that there is widespread 'degradation' of gene control regions in humans.

What can explain this dramatic difference in selective constraints in murids versus hominids? The authors rule out several possibilities including greater rates of sequencing error in the hominid data, a greater number of pseudogenes in the hominid data set and increased adaptive evolution in the hominid lineage. Instead they conclude that the fixation of many mildly deleterious mutations is likely to be responsible for the decreased levels of constraint in hominid regulatory regions. But how is this possible?

It is well-known to population geneticists that the probability of fixing a new mutation is dependent, not only on the magnitude and sign (beneficial or deleterious) of the mutation but also on the population size of the species in which the mutation occurs. In situations where the population size is large, the fate of a new beneficial or deleterious mutation is largely determined by the

selective value of the mutation, with a deleterious mutation more likely to be purged and beneficial mutation more likely to become fixed. By contrast, in small populations, the fixation probability of a negative or positive mutation is governed chiefly by stochastic forces such as genetic drift. Thus, the probability of fixing a new mutation by chance, even a deleterious mutation, is greater in a smaller population.

Given the smaller effective population size (N_e) of hominids relative to murids, ~20 000 versus ~600 000 [11], respectively, Keightley *et al.* argue that selection might be ineffective in purging deleterious mutations in the 5′ and 3′ region of genes in hominids. Therefore, a build up of slightly deleterious mutations in these regions would be manifested as a significant drop in apparent selective constraint. Although the authors note that this degradation will not continue indefinitely (thankfully!), they suggest it might explain why most human mendelian diseases are caused by changes in protein-coding regions where selective contraints are higher.

Population genetics and history

Interestingly, the invocation of N_e as an explanation for patterns of regulatory evolution is not entirely new. In 2002, Carter and Wagner [12] proposed a population genetic model to explain the observation of relatively low rates of evolution in vertebrate enhancers compared with the rates observed in *Drosophila*. In this model, slightly deleterious mutations are followed by compensatory mutations leading to an increased rate of evolution in *cis*-regulatory elements. Carter and Wagner showed that population size dramatically affects the rate of fixation of pairs of compensatory mutations. In particular, there is a 'sweet spot' of intermediate population size where the rate of compensatory mutations can greatly exceed the neutral mutation rate, leading to a high turnover of nucleotides in *cis*-regulatory elements.

Given an estimated per nucleotide mutation rate of $\sim 10^{-9}$ per generation (important to factor in the rate of introduction of new mutations) and an estimated N_e of 10 000–30 000 for humans and chimpanzees [11], primates appear to fall within this sweet spot (like flies), whereas murids do not ($N_e = 450\,000$–$810\,000$). Although further analysis is required to confirm this hypothesis, it is clear, from the work of Keightley *et al.* and others, that population genetic parameters cannot be ignored in the context of comparative genomics and in studies of noncoding sequences in particular. Models of noncoding evolution that explicitly take into account the effect of population history and structure on sequence change, promise to enhance our ability to identify functionally important genomic regions correctly.

The importance of alignment

Intimately related to the study of functional sequence evolution, but often unmentioned, is the process of sequence alignment – a particularly difficult task when noncoding sequences are involved. Rates of change inferred between species under different alignment schemes can differ dramatically. This is important because, in a poor alignment, the comparison of a substantial number of non-homologous nucleotides will make any subsequent evolutionary analysis meaningless. Problems associated with alignment ambiguity were largely avoided in the studies discussed above by using closely related species [i.e. that diverged <20 million years ago (Mya)]. However, Keightley and colleagues went a step further by using a stochastic model-based alignment procedure whereby empirical rates of insertion/deletion and nucleotide substitution in each lineage were used to generate the final alignments [13].

Considering the importance of sequence alignment relative to elucidating functional *cis*-regulatory elements, greater attention must be paid to generating realistic null models of how noncoding sequences change over time such as the model discussed in Ref. [13]. By combining null models of noncoding sequence evolution with 'positive' models of functional sequence change, our ability to uncover efficiently the functional elements responsible for *cis*-mediated transcriptional regulation will be greatly improved. It would be particularly exciting if we could reveal the combinatorial logic of transcriptional regulation by comparing more distantly related species where enough time has passed for changes in the spacing and orientation of TFBS and other elements to occur. However, such analysis awaits the release of more eukaryotic genomes and the creative efforts of wet-laboratory and dry-laboratory workers alike.

Concluding remarks

The comparative evolutionary approach epitomized by the work of Keightley *et al.* and Chin *et al.* demonstrates the potential of computational studies to reveal large-scale patterns of *cis*-regulatory architecture. In yeast, species with compact genomes, ~30% of upstream sequences are likely to be functional in gene regulation. In mammals, species with much larger genomes, it appears that functional elements are more diffuse than in yeast but are clustered mostly within 2-kb surrounding protein-coding sequences. These observations help to paint a general picture of noncoding conservation and structure in the genome and are likely to be extremely helpful in focusing future detailed investigations. In addition, the dramatic difference in functional constraints between murids and hominids discovered by Keightley *et al.* highlights the importance of population genetic parameters in understanding both the noncoding sequence change and the dynamics of *cis*-regulatory evolution. The future incorporation of demographic, genealogical and phylogenetic information into increasingly sophisticated models of noncoding sequence change promises to bring with it a plethora of exciting biological discoveries. With luck, they will continue to help us better distinguish the *junk* from the truly *func* (functional).

Acknowledgements

I thank Jun S. Liu for support and inspiration, Christina Muirhead, Sarah Kingan and Daniel Hartl for thoughtful discussion, and Daniel Weinreich, Christian Landry, Scott Ribich, Laurence Hurst and an anonymous reviewer for valuable comments on the article. This work was supported by NIH grant R01-HG02518-01 to J.S. Liu.

References

1 Marcuello, E. *et al*. (2004) Single nucleotide polymorphism in the 5′ tandem repeat sequences of thymidylate synthase gene predicts for response to fluorouracil-based chemotherapy in advanced colorectal cancer patients. *Int. J. Cancer* 112, 733–737
2 Ueda, H. *et al*. (2003) Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* 423, 506–511
3 Lazzaro, B.P. *et al*. (2004) Genetic basis of natural variation in *D. melanogaster* antibacterial immunity. *Science* 303, 1873–1876
4 Tournamille, C. *et al*. (1995) Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat. Genet.* 10, 224–228
5 Davidson, E.H. (2001) *Genomic Regulatory Systems: Development and Evolution*, Academic Press
6 Wittkopp, P.J. *et al*. (2004) Evolutionary changes in *cis* and *trans* gene regulation. *Nature* 430, 85–88
7 Knight, J.C. (2005) Regulatory polymorphisms underlying complex disease traits. *J. Mol. Med.* 83, 97–109
8 Carey, M. and Smale, S.T. (2000) *Transcriptional Regulation in Eukaryotes: Concepts, Strategies and Techniques*, Cold Spring Harbor Laboratory Press
9 Lee, T.I. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799–804
10 Chin, C.S. *et al*. (2005) Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence. *Genome Res.* 15, 205–213
11 Keightley, P.D. *et al*. (2005) Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* 3, e42
12 Carter, A.J. and Wagner, G.P. (2002) Evolution of functionally conserved enhancers can be accelerated in large populations: a population-genetic model. *Proc Biol Sci* 269, 953–960
13 Keightley, P.D. and Johnson, T. (2004) MCALIGN: stochastic alignment of noncoding DNA sequences based on an evolutionary model of sequence evolution. *Genome Res.* 14, 442–450

0168-9525/$ - see front matter
doi:10.1016/j.tig.2005.08.001

Cnidarians and ancestral genetic complexity in the animal kingdom

David J. Miller[1], Eldon E. Ball[2] and Ulrich Technau[3]

[1]Comparative Genomics Centre, Molecular Sciences Building 21, James Cook University, Townsville, Queensland 4811, Australia
[2]Centre for the Molecular Genetics of Development and Molecular Genetics and Evolution Group, Research School of Biological Sciences, Australian National University, P.O Box 475, Canberra, ACT2601, Australia
[3]Sars International Centre for Marine Molecular Biology, Thormøhlensgt. 55, 5008 Bergen, Norway

Eleven of the twelve recognized wingless (Wnt) subfamilies are represented in the sea anemone *Nematostella vectensis*, indicating that this developmentally important gene family was already fully diversified in the common ancestor of 'higher' animals. In deuterostomes, although duplications have occurred, no novel subfamilies of Wnts have evolved. By contrast, the protostomes *Drosophila* and *Caenorhabditis* have lost half of the ancestral Wnts. This pattern – loss of genes from an ancestrally complex state – might be more important in animal evolution than previously recognized.

Introduction

One of the most deep-rooted assumptions in animal biology is that the evolution of vertebrate characteristics, such as a sophisticated humoral immune system, the neural crest and a highly complex nervous system, was enabled by new sets of genes. This notion appears legitimate when mammals are compared with the model ecdysozoans *Drosophila* and *Caenorhabditis* but, as we learn more about the genetic makeup of additional organisms, the list of 'vertebrate-specific' genes seems to be shrinking by the day. The broadening of comparative genomics to include animals such as the sea anemone *Nematostella vectensis*, the coral *Acropora millepora* (both members of the cnidarian Class Anthozoa) and the ragworm *Platynereis dumerilii* (Annelida, Polychaeta) requires some radical rethinking of traditional assumptions about the origins of many vertebrate genes. There have been intriguing hints that some 'vertebrate-specific' genes might predate the origin of the Bilateria (see Glossary) [1–4], and this point is elegantly made in a recent paper on wingless (Wnt) gene diversity in *Nematostella* [5], which broadens, and pushes back in time, the conclusions previously reached for the same gene family by Prud'homme *et al.* [6].

Glossary

Cnidaria: a basal phylum, traditionally characterized as having two body layers, radial symmetry and being at the tissue grade of morphological organisation. The defining characteristic of the phylum is the presence of a nematocyst, or stinging cell. There are two basic morphologies; the sessile polyp and the swimming medusa or jellyfish. The phylum contains four classes, the basal Anthozoa, to which the sea anemone *Nematostella* and the coral *Acropora* belong, the Cubozoa or 'sea wasps', the Scyphozoa, or 'true' jellyfish, and the Hydrozoa, which includes the familiar freshwater *Hydra*.
Bilateria: a monophyletic group of metazoan animals characterized by bilateral symmetry. This group, which could also be termed the 'higher Metazoa' excludes the Cnidaria, Ctenophora, Porifera (sponges) and Placozoa.
Oral–aboral axis: the single obvious body axis of the two 'radiate' phyla (Cnidaria and Ctenophora), marked at one end by the mouth or oral pore.
Deuterostomes: those bilaterians in which the anus opens near the former site of the blastopore.
Protostomes: those bilaterians in which the mouth opens near the former site of the blastopore.

Corresponding author: Miller, D.J. (david.miller@jcu.edu.au).
Available online 11 August 2005