Creation aspects of conserved non-coding sequences

Mátyás Cserháti

Because of their widespread distribution, conserved non-coding sequences have important implications for the creation/evolution debate. Such sequences are indeed highly conserved, which means they resist mutational change. Thus, they are design elements in the genetic makeup of organisms that may help to differentiate between taxa. Moreover, many evolutionists now believe regulatory sequences are the central motor for molecular evolution, and that evolution of these regulatory regions is what mainly alters protein expression. But this does not explain molecule-to-man evolution, which requires a continuous supply of new genetic information. It does however provide an explanation for the origin of variability within the created kinds in the biblical creation model. In this model, the information content of genes is conserved, while certain regulatory changes bring about changes in gene expression.

Non-coding genetic sequences have received much attention from evolutionists in recent times since they are supposedly an important component in a 'viable mechanism' for evolutionary change. However, the most significant characteristic of these sequences is that they are conserved. This term occurs very often in the scientific literature of molecular evolution. For example, a textbook on bioinformatics notes:

> 'As biological sequence data has accumulated, it has become apparent that nature is conservative. A new biochemistry isn't created for each new species, and new functionality isn't created by the sudden appearance of whole new genes.'[1]

Because of their conserved nature, these sequences are referred to us conserved non-coding sequences, or CNSs*.

Basic characteristics of conserved non-coding sequences (CNSs)

The computer program *rVista* identifies CNSs as non-coding sequences, longer than 100 base pairs (bp), which share a sequence identity of over 70%.[2] A study by Dermitzakis *et al.* has uncovered highly conserved CNSs across a number of mammal genomes, which suggests they play an important function. The CNSs were also shown to be widespread within the human genome—65,600 of them, twice the number of human genes.[3] Moreover, CNSs cover large tracts of regions of non-coding sequence. In the case of mammals, there are no orthologous* gene pairs published to date which are devoid of such sequences. According to other estimates, 26–56% of all mammalian non-coding space is conserved.[2] This means that in mammals a large part of the genome is functional, even though only a small percentage of it is made up of genes (roughly 1.2% in humans). This evidence goes directly against the genocentric evolutionary paradigm, which says that only genes are important for molecular evolution. Current research is constantly showing that the function of CNSs is associated with the regulation of the genes, thereby assigning functions to parts of the genome previously assumed to be just junk DNA. Kaplinsky *et al.* and Dermitzakis *et al.* have also shown that maize-rice CNSs, as well as a number of CNSs in 12 mammal species, could be used successfully as PCR* primer binding sites in all 10,000 grass species belonging to the family *Poaceae*.[3,4] This is notable since PCR primers require almost perfectly matching sequences for binding, and means that these CNSs are very similar.

A study by Bejerano *et al.* showed that in genomic comparisons between man and rodents 481 segments of DNA found within introns, untranslated regions (UTRs*), exons, and intergenic sequences (termed 'ultraconservative sequences') of over 200 bp long (but with sequences even longer than 700 or even 1,000 bp) are all 100% conserved. According to evolutionary standards, using the slowest neutral substitution rate over a 1Mb stretch of DNA, the chances of even 1 such sequence appearing within 2.9 billion bp is less than 10^{-22}. Compared with the distribution of single nucleotide polymorphisms (SNPs*) in the human genome, far fewer validated SNPs were registered than expected, which led to their conclusion that the mutation rate of these sequences was 20 times slower than the average rate in the genome.[5] This is quite extraordinary, and raises the question of what is the cause of such stasis in mutations. In this study, a portion of these ultra-conserved sequences tended to cluster in areas in and around genes (e.g. within introns and promoters) associated with transcription factors, RNA binding and the regulation of splicing (since many of these sequences were abundant in the RNA recognition motif RRM). Many of these sequences were found near genes involved in different developmental tasks, implying that such genes play a fundamental role in the morphogenetic make-up of the organism, and therefore tolerate close to zero change. Since these sequences are highly conserved (even more so than house-keeping genes), their existence across several vertebrate species shows they could be considered as intrabaraminic common design elements.

* Defined in the glossary at the end of this article.

Indeed, the widespread conserved nature of certain types of DNA sequence can be fully appreciated when we consider they are thought to have been conserved during evolution, i.e. before plants and animals supposedly diverged. But according to the creation model, these conserved sequences are simply elements with a common design and function within different genomes.

CNSs in molecular evolutionary studies

CNSs have been studied extensively in supposedly close evolutionary relatives, since they are assumed to be similar and therefore easy to find. In addition, genes from distant species have also been studied.[6–10] According to evolutionary belief, one must be very careful when comparing species. Those which have sufficiently diverged should be chosen so that functional non-protein-coding sequences become apparent.[11] Closely related species will supposedly not have had enough time to diverge, which will result in many false positive sequences.

Conversely, species distant on the phylogenetic tree will have diverged too much, and therefore conserved sequences will not be found in great numbers, such as seen when comparing humans and flies. Therefore, researchers try to select species in between in order to find an adequate number of CNSs which are sufficiently conserved. For example, geneticist S. Brenner suggested that the genome of *Fugu rubripes* (Japanese pufferfish) be sequenced because it is 'sufficiently' distant from man. The result was the discovery of more than 1,000 genes 'homologous' between man and *Fugu*.[7] The explanation is that these genes must be very important since they tolerate very little change in their sequences, and are thus labelled as 'highly conserved'.

However, a different interpretation of these results is simply that CNSs are found in a wide variety of organisms because they are common design elements. Moreover, studies made between related species may reveal particular elements specific to distinct groups, which might be useful in delineating baraminic barriers.

Transcription factor binding sites and regulatory networks

Basic characteristics of transcription factor binding sites

During gene regulation, certain types of effector molecules bind to the DNA in front (upstream*) of a given gene (figure 1). These effector molecules, called transcription factors (TFs), also known as *trans* elements, are mainly regulatory proteins that come into physical contact with the DNA sequence at certain sites called transcription-factor binding sites (TFBS, also called *cis* elements). The expression or silencing of the genes they regulate is a result of the net positive (activating) and negative (inhibiting) effects of these regulatory proteins, which is based on many molecular interactions (hydrophobic reactions, hydrogen bonding, polarity, size, shape and packing of molecule groups on the protein-DNA surface).[6] DNA-binding domains have also been found to be conserved within DNA-binding proteins that bind to TFBSs.[12]

Coactivators and corepressors are also involved in gene regulation, but exert their effects indirectly by binding to TFs. The net effect of TFs is to activate the basal transcription inititation complex, which results in the synthesis of mRNA* and therefore activation of the gene.[13] In eukaryotes, TF-TFBS complexes tend to group together in each others' vicinity (usually around the area responsible for transcription) in order to bring about their regulatory effect through contact with each other and the DNA molecule. Prokaryotic TFs do not form large transcriptional complexes because they have a different promoter-sequence makeup.[6] It is even today a great mystery to evolutionists as to how gene regulation could have evolved to bridge the differences between the two types of cells.

A study by Loots *et al.* has shown that TFBSs are located mainly within regions of DNA that are highly conserved.[14] Many types of these non-coding TFBS sequences are involved in the regulation of individual genes and are found within areas such as promoter region, 5' and 3' UTR regions, as well as introns and exons.[15,16] The conserved nature of TFBSs is so important that a number of computer algorithms[14,17–20] make use of this characteristic to find such sequences; one of the most notable programs is FootPrinter*.[21,22] The process of finding CNSs that are shared by a number of species is called 'phylogenetic footprinting'.[23] The basic assumption of FootPrinter is that TFBSs can be found in sequences with similar function if they have undergone little mutational change, i.e. as a result of stabilizing selection,[5,7]

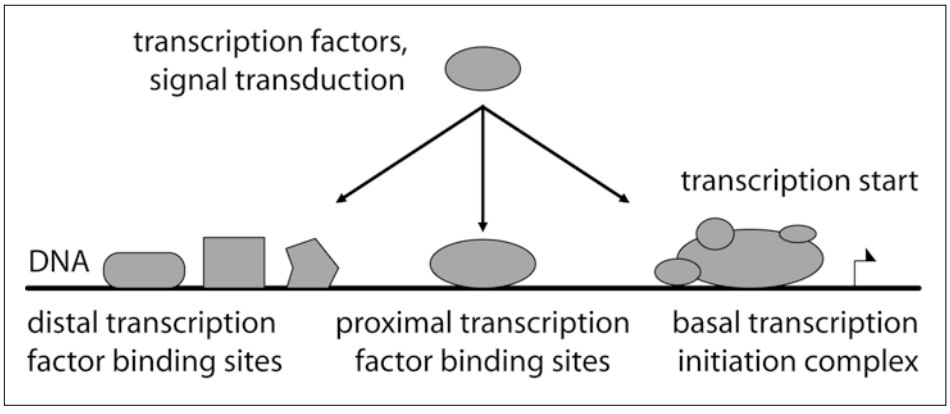

Figure 1. Gene regulation at the DNA level.

compared to surrounding sequences which have been free to mutate. DNA segments under selection (that is, those with similar function) will have higher sequence similarity than non-selected regions.[19,24]

This concept supports the creation model, since creationists have always stressed that selection simply preserves information that was already present to begin with. However, one must be careful when interpreting such data, since many of these studies have been performed between species which could belong to different baramins. TFBSs may therefore represent common design elements or perhaps elements common to an apobaramin*. For example, it has been shown that roughly the same genes are used by all 150,000 species of flies to create their appearance.[25] Variability within a monobaramin therefore arises from the combination of such informational units represented by TFBSs.

A study by Moses *et al.* of 7 TFBSs from four *Saccharomyces* (yeast) species showed that the mutation rate of background sequences varies significantly in different gene groups.[26] This is quite an anomaly for evolutionists, who predicted that mutation rates would not differ substantially between groups of DNA sequences.[5,7,27]

CNSs also have medical implications; e.g. Gumucio *et al.* have found *cis* elements responsible for the regulation of human ε and γ globin genes, which had been missed because of their conserved nature.[28]

It is important to note that individual TFs can bind to the same or to different TFBSs. So the conserved nature of TFBSs reflects the way they bind to regulatory proteins. The more interactions a base pair has with a regulatory protein, the more destabilizing its effect is if substituted, although this rule has many exceptions.[26,29] The A/T and C/G content of various TFBSs also affects their ability to bind to individual proteins. For example, in dimers such as leucine zippers, the two ends of the TFBS are conserved, since these parts come into contact with the transcription-factor protein. In contrast, other TBFSs have a conserved 'core'. For example, in the case of the yeast transcription factor Gal4p, which is also a leucine zipper, only the ends of the individual dimer subunits come into contact with the major groove of the DNA molecule (depicted in figure 2 in a protein-DNA complex). It is these parts of the Gal4p's TFBS (that is the three bases at each end of the sequence **CCG**GAGGACAGTCCTC**CGG**) that are conserved. This corresponds to a lower substitution rate, i.e. substitutions are not tolerated at these contact points between the protein and the TFBS.

Regulatory networks

Multiple TFBSs involved in the regulation of a given gene do so by aggregating to form a genetic regulatory network. Genetic regulatory networks consist of multiple regulatory modules, which themselves are also made up of multiple TFBSs. These networks are quite complex, and are commonly described in the literature. DNA chip data illustrate this complex genetic relationship.[30,31]

Genetic networks process signals that come from outside the cell, resulting in different types of cellular responses. TFBSs are capable of communicating not only with each other but also with a wide spectrum of other genes spatially and temporally outside the scope of their own regulatory modules.[32] Most TFs regulate a small number of genes, but there are some general TFs that can regulate large numbers of genes.[8]

TFs are encoded by their own genes, which make up large parts of the genome.[33] In fact, a gigantic task of genome research today is to enumerate and annotate all the genetic regulatory elements within a given organism.

TFs play an important role in development, where fine-tuned regulation must take place spatially as well as temporally. The affinity of some TFs is much higher only when in complex with other TFs. Certain core TFBS elements in the promoters of some genes are indispensable to gene regulation and subsequent transcription. This is amply proven by promoter-deletion experiments, where larger and larger parts of the promoter are sequentially deleted until the gene cannot be transcribed anymore.

While it is true that TFs are able to bind their cognate TFBSs on their own, they must bind other TFs and act in unison in order to induce gene expression. TFBSs therefore make up irreducibly complex systems, integrating the effect of each individual TFBS in order for all of the regulatory machinery to work.[34,35] It must also be stressed that since TFBSs work in networks, genes taking part in gene regulations are themselves the most enriched with CNSs. This shows their important role in regulation: on average they have 9 CNSs, compared to an average of 2.1 and 2.4 CNSs for genes which encode structural proteins or enzymes, respectively.[36] Also, for example, 15% of the genes on the 1.9 megabases of *Arabidopsis* chromosome 4, with predicted or known functions, are now understood to

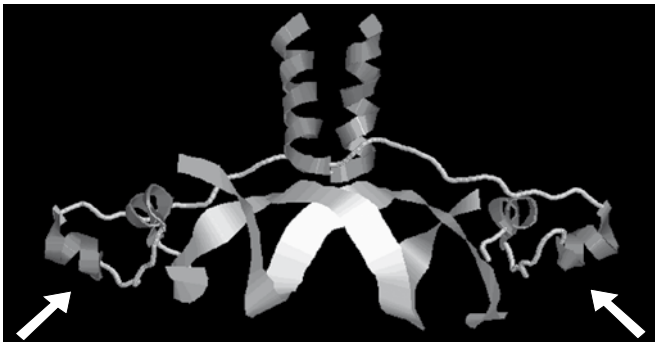

Figure 2. Protein Database image of a Gal4p protein-DNA complex. (PDB id is 1D66; from the Protein Database (www.rcsb.org/pdb/)). The Gal4p protein is made up of a pair of dimers each consisting of two α-helices which fit into the major groove of the DNA molecule. The parts of the Gal4p transcription factor subunits that come into contact with the two ends of the TFBS are indicated by arrows. The TFBS for the Gal4p protein is **CCG**GAGGACAGTCCTC**CGG**; the three bases at each end of the sequence are conserved, and it is these which make contact with the protein.

be involved in transcription, which is close to the number in other eukaryotes.[37] This means that those genes which are the most downstream* in the regulatory network (which resembles a feed-forward loop) encode proteins which are responsible for effecting cellular processes. The most typical examples are those belonging to developmental pathways, enlisting signals, transducers, transcriptional regulators and targets. Therefore, the promoters of many such terminal genes contain a single regulatory module. These are regulated mainly by activator factors encoded by genes further upstream.[38,39]

Indeed, a large part of the genome is responsible for its own regulation—extracting information from itself to regulate itself.[40] The widespread redundancy of regulation factors within gene networks helps to stabilize the effects of mutations which might otherwise disrupt the network's functioning, and therefore makes it robust to perturbations from environmental changes. This goes to show that genomes were designed to intrinsically resist change. It also raises a key question that if genes were somehow to arise from the primordial soup of chemical evolution, then how could they be possibly regulated without fine-tuned regulatory network complexes existing in the first place? Without inherent regulatory networks present to express genes at the proper time and space, all genes would be unnecessarily active all the time, which the cell would not be able to cope with.

Baraminic characteristics of CNSs vs supposed evolutionary changes

CNSs denote differences in baramins

The divergence of supposedly functionless non-coding DNA is the key to finding CNSs, themselves being unchanged by mutations.[11] Evolutionists therefore have to appeal to rapid, but short-lived, mutational processes (such as genome rearrangements, substitutions, insertions, and deletions) to account for the sudden appearance of genes and CNSs.[5,7,41] However, in a study comparing *Drosophila melanogaster* and *D. virilis* the authors found that indels were 20 times less frequent than point substitutions. Moreover, there have only been a few documented cases in *Drosophila* of point mutations which affect gene regulation, all of which lead to genetic diseases. It was also thought that the typical redundancy of regulatory elements was used as a buffer against the mutational effects of point mutations, which destroy and delete genetic information instead of creating it.[42,43]

What is also interesting is that intra-species comparison helps to identify CNSs.[5,7] But since certain sequence differences are visible within single species, this raises the question of whether they really do point to evolutionary divergence and descent. A study of the frequency of CNS distribution in promoters of 130 orthologous genes of different lengths in maize, sorghum, barley, wheat and rice raises the same question. Guo and Moose showed in two studies that wheat and barley had approximately the same percentage of sequence identity as maize and sorghum (92.3% and 89.4%, respectively), and this was larger than that between barley and maize, or wheat and maize (85.7% and 85.8%; see table 1). The evolutionary explanation here was that sorghum and maize are members of the Panicoideae subfamily, while barley and wheat are members of the Pooideae subfamily of grasses. Their orthologous gene and CNS content would therefore be more similar to each other, and would also follow phylogenetic relationships.[44] This, however, is evidence consistent with creation, since it would only be logical that, considering the wide range of species within the plant kingdom, organisms in the same baramin would have more similar genes and CNSs. Genetic similarity would be expected between all types of plants because of their shared basic biochemistry (such as the Szent-Györgyi cycle), and therefore also common regulation and genetic elements.

Conservation of CNSs could be interpreted as genetic elements being within the same genetic context but with characteristics typical only to some monobaramins. Organisms within the same monobaramin would be more similar in their biochemical, genetic, physiological and morphological makeup.[45,46] Differences between monobaramins are also reflected in their signal transduction pathways, and in the specific TFBS content of species which are dependent on these specific signal transduction pathways. It is also interesting to note that a number of gene-finding programs have to be calibrated for individual types of organisms because of the species-specific variance in genetic makeup.[47–49]

Species within a baramin can be distinguished from other species based on sequence homology. For example,

Table 1. CNS content for 5 cereal crop genes. (From Guo and Moose[44]).

	Maize			Rice			Sorghum			Barley		
Species	No. of Genes	Coding Identity (%)	Percent CNS	No. of Genes	Coding Identity (%)	Percent CNS	No. of Genes	Coding Identity (%)	Percent CNS	No. of Genes	Coding Identity (%)	Percent CNS
Rice	15	82.7	5.7									
Sorghum	7	89.4	32.2	7	81.0	9.4						
Barley	8	85.7	2.9	8	88.6	4.1	2	92.4	3.2			
Wheat	6	85.8	0.4	6	85.5	0.5	1	81.1	0.0	7	92.3	34.9

Sixty-seven genes representing 18 orthologous gene sets that contained sequences from at least three cereal species (see Methods) were used in multiple-species sequence comparisons with VISTA (>70% identity, 20-bp window size).

when comparing the intronic and intergenic sequences of the two fruit flies *D. melanogaster* and *D. virilis*, one finds that for more than 100 kb of sequence about 26% and 22% of intergenic and intronic sequences of these two species is conserved, respectively.[40] Another intrabaraminic comparison shows that around 20% of the intergenic regions of 142 orthologous genes of the two nematodes, *C. elegans* and *C. briggsae* are homologous.[24] According to preliminary estimates, this is also similar to that found for other eukaryotes.

In a study done by Jareborg *et al.*, 88.5 Mb of non-coding CNSs (excluding those they deemed as orthologous) were compared between human and mouse DNA. Only 0.14% were found to be more than 60% homologous.[50] This is more than 100 times greater than the difference found between the *Drosophila* species, and denotes the sharp differences between baramins. In the case of the intrabaraminic studies, some of the sequences conserved between the two species are a result of the constraint on functional elements, as well as random homology. Conservation occurs because the two species supposedly did not have enough time to diverge, and would therefore have some random genetic sequences in common. This, however, is quite strange, since the two *Drosophila* species which are supposed to have diverged 40 Ma ago[40] have remained basically the same, whereas man and mouse supposedly diverged at about the same time as the two fly species. One could therefore pose the question as to what kinds of mutational effects would keep the genetic material of two closely related fruit fly species highly conserved during such a long time, yet cause large differences between man and mouse. This is problematic for evolution.

The tautology of the regulatory evolution paradox

Because of the high conservation of regulatory proteins, it has been suggested that the evolution of gene regulation proceeds mainly via substitutions, indels and tandem repeat variants of TFBSs, as well as through changes in the regulatory factors themselves.[38,51] However, a detailed understanding of the evolution of TFBSs is still missing because the promoter structures have not undergone bioinformatic analysis, or even more laboratory studies, which is the only true mean of verifying that a sequence is indeed a TFBS.[38]

Since the 1960s and 1970s, several authors (e.g. Britten, Davidson, Ohno, Wilson) have also proposed that morphological changes take place mostly through alterations in gene expression instead of evolution of the coding sequences themselves.[38,52–54] For example, they suggest that changes in the regulation of homeodomain*-containing proteins affects both the morphology of flowers and the development of embryos. It has even been shown that the fraction of conserved genetic sequences is about the same in non-coding and coding sequences.[40,55] This would mean that since coding sequences are much smaller than regulatory sequences, coding sequences would be more resistant to change than intergenic regions which are involved in gene regulation. This makes a case against molecular evolution, since the coding regions of the genes themselves would be responsible for phenotypic expression via transcription of the genes.

For example, a study of three *Drosophila* species shows that although the *Ubx* gene is the same in all three species, the trichome patterns on the posterior femur of the second leg are different.[25] This evidence indicates that (although there *may* be other genes involved, albeit tangentially) no net genetic evolution has occurred in the coding regions. This suggests that certain morphological changes are possible, within limits—for example, no morphological novelties or new cellular functions are produced. Even with limited genetic material, if it is recombined in a number of ways it can bring about new species that are able to adapt to individual ecological niches.

Another study showed that regulatory changes, instead of changes in the transcription unit in the *teosinte-branched1* genes, were responsible for the morphological changes that transformed teosinte into a species of maize. The non-transcribed region, and a transcribed unit for this gene responsible for the lengths of branches and the morphology of their tassels or ears in maize species, were looked at. There was no amino acid sequence difference between maize and teosinte, whereas the measurement of mRNA levels in both species showed that regulatory changes had occured. Significant HKA tests*, which indicate the level of polymorphism compared to neutrally selected genes, showed variance in the non-transcribed region. In addition, neighbour-joining trees showed a significant selective effect for the non-transcribed region when compared with the transcription unit. The study also showed that the selection of maize occurred within hundreds of years—very much consistent with a biblical time scale, where agriculture was adopted earlier than hunting—crops being domesticated within the last 10,000 years according to evolutionary time-scale.[56]

Evolutionists suggest that changes in the TFBS content of promoters following gene duplication would also be responsible for changes in gene expression. That is, the more a pair of genes have diverged from each other, the larger the differences in TFBS content and expression patterns. However, a detailed study by Zhang *et al.* of 202 pairs of yeast genes showed there was only a weak correlation between TFBS content and expression, and showed that the 10 most highly co-expressed gene pairs do not have even half of their TFBSs in common. They believe other factors, such as motif-motif interactions*, *trans* factors, and chromatin structures might be responsible for differences in expression.[57]

Another study by Castillo-Davis *et al.* shows that after studying paralogous* and orthologous genes in the genomes

of *C. elegans* and *C. briggsae*, the rate of change in both the regulatory region and the coding region is faster in paralogous genes than in orthologous genes. For example, fractions of shared regulatory motifs were lower in sets of duplicated genes. It would be only logical that the difference would be greater among orthologs than paralogs. The explanation given is that paralogous genes predate speciation and are therefore older than the orthologs in the other species, i.e. they have had time to diverge. However, the authors found that more than 93% of the duplicated genes post-dated speciation.[39]

One must take into account that not only can the expression profiles of similar genes differ greatly, but genes which are very different may also have very similar expression profiles. This would suggest a varying rate of evolution for different promoters and promoter elements based on, for example, the functions they perform, and introduces vague notions such as mutational hot spots and cold spots.[24,38,58] According to Bergman and Kreitman, these observations have led to a paradoxical evolutionary model where a stabilizing selection acts on the phenotype of genes selected, but at the same time allows for flux in the composition of the underlying *cis*-regulatory sequences.[11,40]

In such cases, the creation model simply says that duplicated genes belong to the same family because of their similar genetic makeup. This would mean that in most cases duplicated genes would have similar TFBS content, and therefore their expression profiles would be expected to be about the same. However, since evolutionary genetics needs to explain the genetic transformation of molecules into highly developed organisms, inexplicably rapid rates of evolution are invoked to 'bridge the gap' between promoters which are believed by evolutionists to have diverged from one another, but which, by now, have acquired different expression. In this case, evolution is clearly used in a tautological manner: the very supposition that evolution makes is used to explain how rapidly conserved sequences might have evolved to acquire diverging profiles through time.

Conclusion

In this paper we have covered two major paradigm shifts within molecular biology, based on the analysis of a large number of genetic sequences, which clearly point to creation and not evolution. Firstly, comparative genomics reveals that not only coding sequences but also non-coding sequences are highly conserved—the result that conserved non-coding sequences, or CNSs, are widely found among different kinds of genes has made them useful for sequence analysis and in medicine. This conserved nature indicates that non-coding sequences are resistant to mutation, which destroys rather than creates new genetic information. From a creationist viewpoint, non-coding sequences can be seen as distinctly created genetic entities with specific roles in regulation.

The distribution of CNSs is extensive, and sheds additional light on the function of non-coding sequences. It also further weakens the evolutionary 'junk DNA hypothesis' and the 'genocentric' notion that only genes are fundamental in the expression of phenotype. Instead, in a wide range of organisms, many CNSs are functional and an integral part of the genome. Examples of such sequences are TFBSs, which serve regulatory functions upstream of genes, and also make up highly complex regulatory networks—they are an example of irreducibly complex system. Many such CNSs and TFBSs are common genetic elements in a wide number of organisms, all belonging to different baramins. This is clear evidence for design by a common designer. Moreover, differences in such genetic elements may help delineate the boundaries between baramins.

The second paradigm shift discussed here begins with the idea that the coding sequence of a gene often remains mostly intact, and that morphological and biochemical changes can take place as a result of mutation in the genes's promoter region. As we have seen with teosinte and maize, such changes can be quite large, even leading to variations within a created kind, and can also give rise to biogeographical variation and adaption to different ecological niches. As created entities, genes and their promoters cannot simply mutate to form functionally new promoters. Individual elements can sometimes be reshuffled between promoters, but this does not amount to the creation of new genetic information. As in the case of paralogous genes, evolutionists often apply tautological explanations to justify the 'rapid rates of evolution' needed to bridge the gap between genes which have supposedly diverged. According to philosophical principles, such as Occam's razor, the idea of separately created genetic entities is more plausible than some elusive mechanism needed to salvage the idea of molecular evolution.

Acknowledgments

I would like to thank Dr János Pauk of the Protestant Creation Research Fellowship in Hungary who gave me the seminal idea about the widespread conservative nature of DNA sequences on which I based this paper. I would also like to thank Bendegúz Levente Szűk of the Hungarian Christian Science Group for reading over the manuscript, as well as commenting on this work.

Glossary

Apobaramin: A group consisting of the entirety of one or more baramins or kinds, such as canines, birds, humans, etc.

CNS: conserved non-coding sequence: certain DNA sequences of variable lengths, which do not encode proteins, but are still functional. Their main roles lie in gene regulation, therefore their sequences are conserved.

Downstream/upstream: a term denoting sequences either behind (downstream—in the 3' direction), or in front of (upstream—in the 5' direction) a given DNA sequence.

FootPrinter: a computer program used to find conserved sequences across a number of sequences from different species provided as input.

Homeodomain: a stretch of 60 amino acids corresponding to the homeobox part of genes (which are almost identical in all genes they are found in) which are fundamental in determining which groups of genes are expressed during development.

HKA test: a type of statistical analysis developed by Hudson *et al.*,[59] where polymorphism/variability within a given species and DNA sequence divergence between species is proportional to the neutral mutation rates.

Motif-motif interactions: special synergistic regulatory effects which occur between different transcription factor binding sites within a promoter sequence. They occur through the mediatory effect of transcription factors.

mRNA interference: a type of genetic regulation, where mRNA binds to a gene with roughly the same sequence, thereby interfering with its expression. Many genes expressing mRNA taking part in such interference were thought to be 'pseudogenes', which were once genes but which lost their function, and therefore mutate away, but still resemble the gene they interfere with.

Ortholog: genes of similar function found between species which are relatives of one another.

Paralog: a diverged duplicate of an ortholog gene.

PCR: polymerase chain reaction: a common experimental procedure in molecular biology where a DNA sequence of interest is amplified in large quantities in order to study it. This is done by annealing short oligomers about 15–25 bases long called primers to the edges of the sequence of interest, which are complementary to both strands of the DNA. Amplification is achieved by synthesizing the DNA between the two primers with a DNA synthesizing enzyme.

SNP: single nucleotide polymorphism: positions within the genome which show sequence variation. The alleles of genes differ because of the difference in their sequences, which give rise to variations within the given gene.

UTR: untranslated region: those parts of the gene sequence which are translated into mRNA but are not translated into proteins. 5' UTRs lie in front of the ATG start codon starting the protein sequence, while 3' UTRs lie after the terminal polyA sequence, which denotes the end of the protein sequence.

References

1. Gibas, C. and Jambeck, P., *Developing Bioinformatics Computer Skills*, O'Reilly & Associates, Sebastopol, CA, 2001.
2. Inada, D.C. *et al.*, Conserved Noncoding Sequences in the Grasses, *Genome Research* **13**: 2030–2041, 2003.
3. Dermitzakis, E.T. *et al.*, Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs), *Science* **302**(7): 1033–1035, 2003.
4. Kaplinsky, N.J. *et al.*, Utility and distribution of conserved noncoding sequences in the grasses, *PNAS* **99**(9):6147–6151, 2002.
5. Bejerano, G. *et al.*, Ultraconserved elements in the human genome, *Science* **304**(28):1321–1325, 2004, supporting online material: <www.sciencemag.org/cgi/content/full/1098119/DC1>, 9 Sep. 2004.
6. Mirny, L.A. and Gelfand, M.S., Structural analysis of conserved base pairs in protein-DNA complexes, *Nucleic Acids Research* **30**(7):1704–1711, 2002.
7. Boffelli, D. *et al.*, Comparative genomics at the vertabrate level, *Nature Reviews Genetics* **5**:456–465, 2004.
8. Barabási, A. and Oltvai, Z.N., Network biology: understanding the cell's functional organization, *Nature Reviews* **5**:101–113, 2004.
9. Tompa, M. *et al.*, Identifying functional elements by comparative dna sequence analysis, *Genome Research* **11**:1143–1144, 2001.
10. Dubchak, I. *et al.*, Active conservation of noncoding sequences revealed by three-way species comparisons, *Genome Research* **10**:1304–1306, 2000.
11. Cliften, P.F. *et al.*, Surveying *Saccharomyces* genomes to identify functional elements by comparative dna sequence analysis, *Genome Research* **11**:1175–1186, 2001.
12. Ahmad, S. *et al.*, Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information, *Bioinformatics* **20**(4):477–486, 2004.
13. Weaver, R.F. and Hedrick, P.W., *Genetics*, Wm. C. Brown Publishers, Dubuque, IA, 1997.
14. Loots, G.G. *et al.*, rVista for comparative sequence-based discovery of functional transcription factor binding sites, *Genome Reseach* **12**:832–839, 2002.
15. Levy, S. *et al.*, Enrichment of regulatory signals in conserved non-coding genomic sequences, *Bioinformatics* **17**:871–877, 2001.
16. Majewski J. and Ott J., Distribution and characterization of regulatory elements in the human genome, *Genome Research* **12**:1827–1836, 2002.
17. Wassermann, W.W. and Sandelin, A., Applied bioinformatics for the identification of regulatory elements, *Nature Reviews Genetics* **5**:276–287, 2004.
18. McGuire, A.M. *et al.*, Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes, *Genome Research* **10**:744–757, 2000.
19. Kechris, K.J., Detecting DNA regulatory motifs by incorporating positional trends in information content, *Genome Biology* **5**:1–21, 2004.
20. Stojanovic, N., Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions, *Nucleic Acids Research* **27**(19):3899–3910, 1999.
21. Blanchette, M. and Tompa, M., FootPrinter: a program designed for phylogenetic footprinting, *Nucleic Acids Research* **31**(13):3840–3842, 2003.
22. Blanchette, M. and Tompa, M., Discovery of regulatory elements by a computational method for phylogenetic footprinting, *Genome Research* **12**:739–748, 2002.
23. Hardison, R.C. *et al.*, Long human–mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome, *Genome Research* **7**:959–966, 1997.
24. Webb, C.T., Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*, *Nucleic Acids Research* **30**(5):1233–1239, 2002.
25. Moses, A.M. *et al.*, Position specific variation in the rate of evolution in transcription factor binding sites, *BMC Evolutionary Biology* **3**(19), 2003.

26. Stern, D.L., A role of *Ultrabithorax* in morphological differences between *Drosophila* species, *Nature* **396**(3):463–465, 1998.
27. Dermitzakis, E.T., Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover, *Mol. Biol. Evol.* **19**(7):1114–1121, 2002.
28. Gumucio, D.L. *et al.*, Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human γ and ε globin genes, *Molecular and Cellular Biology* **12**(11):4919–4929, 1992.
29. Storbeck, C.J., Definition of regulatory sequence elements in the promoter region and the first intron of the myotonic dystrophy protein kinase gene, *The J. of Biological Chemistry* **273**(15):9139–9147, 1998.
30. Sayyed-Ahmad, A., Tuncay, K. and Ortoleva, P.J., Transcriptional regulatory network refinement and quantification through kinetic modeling, gene expression microarray data and information theory, *BMC Bioinformatics* **8**:20, 2007.
31. Gao, F., Foat, B.C. and Bussemaker, H.J., Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data, *BMC Bioinformatics* **5**:31, 2004.
32. Park, P.J. *et al.*, Comparing expression profiles of genes with similar promoter regions, *Bioinformatics* **18**(12):1576–1584, 2002.
33. Lijavetzky, D. *et al.*, Genome-wide comparative phylogenetic analysis of the rice and Arabidopsis Dof gene families, *BMC Evolutionary Biology* **3**(17), 2003.
34. Behe, M.J., *Darwin's Black Box*, Free Press, New York, 1996.
35. Woodmorappe, J., Irreducible complexity: some candid admissions by evolutionists, *J. Creation* **17**(2):56–59, 2003.
36. Arnone M.I. and Davidson, E.H., The hardwiring of development: organization and function of genomic regulatory systems, *Development* **124**:1851–1864, 1997.
37. Singh, K.B., Transcriptional regulation in plants: the importance of combinatorial control, *Plant Physiology* **118**:1111–1120, 1998.
38. Milo, R. *et al.*, Network motifs: simple building blocks of complex networks, *Science* **298**: 824–826, 2002.
39. Kirchhamer, C.V. *et al.*, Modular cis–regulatory organization of developmentally expressed genes: Two genes transcribed territorially in the sea urchin embryo, and additional examples, *Proc. Natl. Acad. Sci. USA* **93**:9322–9328, 1996.
40. Wray, G.A. *et al.*, The evolution of transcriptional regulation in eukaryotes, *Mol. Biol. Evol.* **20**(9):1377–1419, 2003.
41. Castillo-Davis C.I. *et al.*, *cis*-Regulatory and protein evolution in orthologous and duplicate genes, *Genome Research* **14**:1530–1536, 2004.
42. Bergman, C.M. and Kreitman, M., Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences, *Genome Research* **11**:1335–1345, 2001.
43. Ludwig, M.Z. and Kreitman, M., Evolutionary dynamics of the enhancer region of *even-skipped* in *Drosophila*, *Mol. Biol. Evol.* **12**(6):1002–1011, 1995.
44. Guo, H., and Moose, S.P., Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution, *The Plant Cell* **15**:1143–1158, 2003.
45. Gish D.T., *Creation Scientists Answer Their Critics*, Institute for Creation Research, El Cajon, CA, 1993.
46. ReMine, W.J., *The Biotic Message: Evolution Versus Message Theory*, St. Paul Science, St. Paul, MN, 1993.
47. Aubourg, S., and Rouze, P., Genome annotation, *Plant Physiol. Biochem.* **39**:1–13, 2001.
48. Rouze, P. *et al.*, Genome annotation: which tools do we have for it? *Current Opinion in Plant Biology* **2**:90–95, 1999.
49. Foissac, S. *et al.*, EUGENE'HOM: a generic similarity-based gene finder using multiple homolgous sequences, *Nucleic Acids Research* **31**(13):3742–3745, 2003.
50. Jareborg, N. *et al.*, Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs, *Genome Research* **9**:815–824, 1999.
51. Doebley, J. and Lukens, L., Transcriptional regulators and the evolution of plant form, *The Plant Cell* **10**:1075–1082, 1998.
52. Ludwig, M.Z. *et al.*, Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change, *Development* **125**:949–958, 1998.
53. Ohno S., *Evolution by Gene Duplication*, Springer-Verlag, Heidelberg, Germany, 1970.
54. Taverna, D.M. and Goldstein, R.A., Why are proteins so robust to site mutations? *J. Molecular Biology* **315**:479–484, 2002.
55. Frazer, K.A., *et al.*, Evolutionarily conserved sequences on human chromosome 21, *Genome Research* **11**:1651–1659, 2003.
56. Rong-Lin, W. *et al.*, The limits of selection during maize domestication, *Nature* **398**:236–238, 1999.
57. Zhang Z. *et al.*, How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? *Trends in Genetics* **20**(9):403–407, 2004.
58. Clark, A.G., The search for meaning in noncoding DNA, *Genome Research* **11**:1319–1320, 2001.
59. Hudson, R.R., Kreitman, M. and Aguadé, M., A test of neutral molecular evolution based on nucleotide data, *Genetics* **116**:153–159, 1987.

Mátyás Cserháti has an M.Sc. in biology from the Eötvös-Loránd University of Natural Sciences. He is an active member of the Protestant Creation Research Fellowship in Hungary.