Proposed Strategies for Using Non-Coding Sequences, Pseudogenes and Interspersed Elements to Construct Molecular Phylogenies

David Benjamin Weinreb
December 12, 2001

The vast majority of eukaryotic DNA (97% in humans[1]) does not code for proteins or RNAs with clear functions. Such 'junk' DNA can be classified into several categories: introns[2, 3], mini- and microsatellites[4], 3' untranslated regions[5], SINEs[6], LINEs[9] and pseudogenes[10,11]. The sentiment that the majority of junk DNA is nonfunctional has been challenged by studies that demonstrate that long terminal repeats regulate the expression of nearby genes[12]. The *Alu* sequence upstream of the IgE receptor gene interacts with transcription factors to regulate its expression[13] and similarly the mRNA of the *lin-14* gene is regulated by a small RNA encoded in the repeated sequence of the 3'-UTR[17,18]. Furthermore, mutations in minisatellite sequences downstream of the *ras* gene may be associated with many types of cancer [15]. If 'junk DNA' sequences have specific functions, then such sequences are likely to be conserved between closely related genomes, just like gene-coding sequences. Presently, we outline methodologies for extracting phylogenetic relationships from 'junk DNA' using: (1) whole-genome features, (2) conserved non-coding sequences, (3) SINEs and (4) pseudogenes.

Whole-Genome Features

Eukaryotic genomes range in size from 1.2 X 10^7 bp (*Saccharomyces cerevisiae*) to greater than 6 X 10^{11} bp (*Amoeba dubia*)[19] without an apparent correlation with biological complexity (the *C*-value paradox). Striking variability in genome size is apparent in closely related species (for instance, genome size in flowering plants varies 1000-fold)[20]. For this reason, we have fully failed to explain whole-genome features (genome size, intron/exon ratios, numbers of genes, etc) in terms of phylogenetic relationships. Recent studies provide insight into the *C*-value paradox by demonstrating that intra-genome intron distribution is uneven, with larger introns occurring in regions of low recombination rates in both *Drosophila*[21] and humans[22]. We anticipate that such studies may ultimately elucidate the mechanisms through which genome size has evolved. We postulate that annotation of the genomes of many more taxa will be required before we can fully explain phylogenetic variations in such whole genome features.

Sequence Similarity between Non-Coding Sequences

Exceptionally high levels of non-coding sequence homology (~71%) have been identified in a 100 kb stretch in the T cell receptor Cα/Cδ region of the mouse and human genomes, compared to similarity of 73-79% in adjacent gene-coding sequences[25]. Similarly, conserved non-coding sequences (CNSs) have been identified in the intergenic region of both the mouse and human interleukin 4 and interleukin 13 genes; these CNS were subsequently demonstrated to regulate expression of interleukin genes as distant as 120 kb[31]. Hardison (2000) predicts the presence of 270,000 CNSs in the human genome[30] and *PipMaker*[32] has been employed to align and construct percent identity plots for such conserved non-coding regions[33].

Non-coding sequences that are conserved between genomes (especially genomes of distantly related taxa) are more likely to have a specific functionality than sequences that are unique to a single genome. Presently, we propose to use inter-genomic sequence alignments of non-coding regions to: (1) identify CNSs, (2) align and compare CNSs to study phylogenetic relationships (for instance, sequence data from two nuclear introns has been used to construct a phylogeny for mice[28]) and (3) 'search' and 'mine' a genome for non-coding regions that are predicted to have some regulatory function[29] by virtue of them being conserved between genomes. In summary, we advocate the use of a bioinformatic approach to identify large populations of CNSs across multiple genomes and use broad sequence similarities to predict which non-coding sequences may have some functionality.

Inter-Genome Occurrences of SINEs and LINEs

Short interspersed elements (SINEs) and long interspersed elements (LINEs) are integrated into genomes by retroposition. They are generally not 'excisable,' represent irreversible events in genomic evolution and may be valuable phylogenetic markers[26, 27, 35]. Murata *et. al.* (1992) builds a phylogenetic tree for salmonid species based on the presence or absence of each of three families of tRNA-derived SINEs in closely related genomes[26]; similarly, a cladogram based on both the SINEs and LINEs present in cetartiodactyls[34] has confirmed relationships established by more conventional phylogenetic techniques[35]. One of the principal advantages of the SINE/LINE insertion approach is that homoplasy is unlikely because there is a very low probability that a particular SINE will be independently inserted into the same region in two distinct genomes[35]. However, the principal disadvantage is that it requires that we can identify several SINEs/LINEs that were inserted at the appropriate point in evolutionary history to distinguish between the taxa of interest. Presently, we propose that a genome-wide bioinformatic approach can potentially eliminate this limitation: (a) detail genome annotations will allow us to identify all of a genome's SINEs/LINEs, (b) sequencing and annotation of multiple genomes will permit us to both use SINEs to construct phylogenies and estimate the time of insertion of a given interspersed element.

Pseudogenes

Harrison *et. al.* (2001) comment that pseudogenes may be valuable in molecular systematics because they "acquire mutations, insertions and deletions without any apparent evolutionary pressures"[10]. However, while many pseudogenes follow the patterns of mutation expected for non-functional sequences, some experience elevated rates of genetic drift[36,37]. Other pseudogene are remarkably similar to their functional homologs, which reflects either an evolutionarily recent time of pseudogene formation or a functional role for that pseudogene (for instance, pseudogenes are involved in generating immunoglobin heavy chain diversity in chickens)[36].

Here, we suggest the comparative genomic studies based on the following features of the pseudogene populations (as characterized by Harrison *et. al.*) may be of value in phylogenetic reconstructions:

(1) Pseudofolds: Folds and structures can be assigned to pseudogenes based on the structure of the most similar protein[10]. Differences exist in the relative occurrences of predicted globular folds in the *C. elegans* gene products and the hypothetical pseudogene products[10]. Folds more common in the gene than the pseudogene population may represent evolutionarily recent additions to the ensemble of available structural motifs. Evolutionarily trees can be constructed based on *both* the occurrence of folds in proteins and 'pseudoproteins' in different genomes. Comparative genomics of pseudofold

occurrence may help us assess how the ‘preference’ for particular folds or motifs has changed over evolutionary history in distinct lineages.

(2) The Number of Pseudogenes Relative to Genes in a Family: Some gene families (for instance, chemoreceptor and seven-transmembrane receptor genes) have many pseudogenes[10]. Such families may face diminished “evolutionary pressures for their conservation,”[10] which probably reflects a loss of functionality or duplication of functionality by another suites of genes[10]. Comparative genomics should help us answer the questions: (a) At what in evolutionarily history did a families stop facing pressure for its conservation, i.e. when did it lose functionality? (b) How have the evolutionary pressures conserving any particular gene (which may reflect the value of that gene to an organism) vary across phylogenetic lineages?

(3) *Amino Acid Composition*[38]: Older suites of pseudogenes have amino acid compositions[38] more similar to random genomic DNA than younger suites, an observation which can be employed to predict the ‘age’ of a pseudogene[10]. Inter-genomic comparisons of the estimated age of pseudogenes (coupled with an understanding of the variations in nucleotide mutation rates discussed previously) may provide another dimension of genomic information that can be incorporated into phylogenetic analyses.

In conclusion, sequence comparisons of vast non-coding regions have already been used to assemble molecular phylogenies and we propose that detecting non-coding sequences in diverged genomes may help us to predict which introns may have regulatory functions. Recent studies have also used SINEs and LINEs to construct phylogenies of closely related organisms. We contend that bioinformatic techniques should be applied to identify the large population of non-coding sequences, SINEs and pseudogenes in a genome and genome-wide characterization of these elements may be useful in phylogenetic analyses.

TABLE 1: Proposed Strategies for Using Non-Coding DNA to Construct Phylogenetic Analyses

Proposed Approach	Suggested Advantage	Probable Disadvantage
Conserved Non-coding Sequences	May over be 270,000 CNSs in human genome to study; software already available for multiple sequence alignments.	As many non-coding sequences are not conserved, a genome must be searched using bioinformatic techniques to identify these conserved regions.
SINEs/LINEs Insertions	Insertion of SINEs/LINEs represent irreversible events in genome evolution; good markers for phylogenetic reconstructions.	We need extensive bioinformatic surveys and annotations of multiple genomes to find SINEs and LINEs appropriate for the taxa under study.
Pseudogenes	Nucleotide mutations in some pseudogenes may occur without "any apparent evolutionary pressures"[10]. Possible to use pseudofolds to augment more traditional structural genomic comparisons.	Some pseudogenes do not adhere to the pattern of mutation expected for non-functional, random sequences, suggesting they may have some type of functionality.

TABLE 1 (continued):

Proposed Approach	Applied Previously at the Level of Specific Sequences to Construct Phylogenetic Trees?	Applied Previously at a Genome-Wide Level?
Conserved Non-Coding Sequences	Yes (Ref. 28)	No
SINEs/LINEs Insertions	Yes (Refs. 7, 27, 28, 34)	No
Pseudogenes	No	Yes in *C. elegans*. (Harrison *et. al.* 2001)

References

1. Nowak, R (1994) “Mining Treasures From Junk DNA.” *Science* 263: 609-611.

2. Human Genome Project Information, Oak Ridge National Laboratory. http://www.ornl/gov/hgmis/glossary.

3. Introns: DNA sequences that interrupt the protein-coding; transcribed into RNA but not translated into protein.

4. Satellites: Short DNA sequences consisting of short sequences repeated hundreds of times, principally located at the ends of centers of chromosomes. Branch off from the rest of the chromosome, but remain connected by a thin filament[2].

5. 3’-Untranslated regions: end segments of protein coding regions that are transcribed but not translated.

6. Short Interspersed Elements (SINEs): Retroposons that have entered genomes by the integration of a reverse-transcribed copy of RNA. The insertion into a genome tends to be irreversible, unlike DNA transposable elements, because SINEs are generally not precisely excised[7]. Include the ~300-base pair *Alu* sequences[2]. Generally considered to be nonfunctional.

7. Murata, S., Takasaki, N. Saitoh, M. & Okada, N. (1993) “Determination of the phylogenetic relationships among Pacific salmonids by using short interspersed elements (SINEs) as temporal landmarks of evolution.” *Proceedings of the National Academy of Sciences USA*. 90, 6995-6999.

8. Retroposons: sequences that have been entered a genome by integration of a reverse-transcribed copy of RNA[7].

9. Long Interspersed Elements (LINEs): A type of retroposon consisting of repetitive sequences that may approach 7,000 base pairs in length. Like SINEs, they are dismissed as nonfunctional[2].

10. Harrison, P.M., Echols, N. and M.B. Gerstein (2001). “Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome.” *Nucleic Acids Research* Vol 29 (3).

11. Pseudogenes: nonfunctional copies of genes that generally lack introns, are more densely distributed on the arms of chromosomes, and are practically never expressed[2,10]. Two types: processed pseudogenes (resulting from reverse transcription of mRNAs) and

non-processed (resulting from "gene duplication and disablement")[10]. Considered to be the relics of once functional genes that have fallen into disuse and acquired mutations.

12. Sverdlov, E.D. (1998) "Perpetually mobile footprints of ancient infections in human genome." *FEBS LETTERS* 428 (1-2): 1-6.

13. Makalowski, W. (2000) "Genomic scrap yard: how genomes utilize all that junk." *Gene* 259: 61-67.

14. Devlin, B., Krontiris, T., Risch, N. (1993) "Population-Genetics of the HRAS1 minisatellite locus." *American Journal of Human Genetics* 53 (6): 1298-1305.

15. Wyborn, N. King J., Wang, C. Heidemanjoosten, B., Krontiris, T. (1995) "Defining lethal mutations of the HRAS1 minisatellite." *American Journal of Human Genetics.* 57(4): 434-434.

16. Schulz, V.P. and V.A. Zakian. (1994). "The *Saccharomyces* PIF1 DNA helicase inhibits telomere elongation and De-Novo." *Cell* 76(1): 145-155.

17. Feinbaum, R. and V. Ambros. (1999). "The timing of lin-4 RNA accumulation controls the timing of postembryonic developmental events in Caenorhabditis elegans" *Developmental Biology*. 210 (1): 87-95.

18. Lee, R.C., Feinbaum R.L. and V. Ambros (1993). "The C-Elegans heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*." *Cell* 75(5): 843-854.

19. J.M. Comeron (2001) "What controls the length of noncoding DNA?" *Current Opinion in Genetics & Development.* 11:652-659.

20. Cavalier S.T. (1985) *The Evolution of Genome Size*. New York: John Wiley.

21. Carvalho, A.B. and A.G. Clark (1999). "Genetic recombination: Intron size and natural selection."

22. Comeron, J.M. and M. Kreitman (2000). "The Correlation Between Intron Length and Recombination in Drosophila: Dynamic Equilibrium Between Mutational and Selective Forces." *Genetics* 156: 1175-1190.

23. denDennen, J.T., van Neck, J.W., Cremers, P.M., Lubsen, N.H. and Schoenmakers, J.G.G. (1989). "Nucleotide sequence of the rate g crystalline gene region and comparison with an orthologous human region." *Gene* 78 (201-213).

24. Shehee, W.R. (1989). "Nucleotide sequence of the BALB/C mouse β-globin complex." *Journal of Molecular Biology*. 205, 41-62.

25. Koop, B.F. and L. Hood (1994). "Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA." *Nature Genetics*. 7, 48-53.

26. Murata, S., Takasaki, N., Saitoh, M. and N. Okada (1993). "Determination of the phylogenetic relationships among Pacific salmonids by using short interspersed elements (SINEs) as temporal landmarks of evolution." *Proceedings of the National Academy of Sciences USA* 90, 6995-6999.

27. Shimamura *et. al.* (1997). "Molecular evidence from retroposons that whales form a clade within even-toed ungulates." *Nature* 388, 666.

28. DeBry, R.W. and S. Seshadri (2001). "Nuclear intron sequences for phylogenetics of closely related mammals: an example using the phylogeny of *Mus*." *Journal of Mammalogy*, 82(2): 280-288.

29. Miller, W. (2001) "Comparison of genomic DNA sequences: solved and unsolved problems." *Bioinformatics*. 17(5), 391-397.

30. Hardison, R.C. (2000). "Conserved noncoding sequences are reliable guides to regulatory elements." *Trends in Genetics* 16 (9): 369-372.

31. Loots, G.G. (2000). "Identification of a coordinate regulator of interleukins 4, 13, and 15 by cross-species sequence comparisons." *Science*. 288, 136-140.

32. *PipMaker*, http://bio.cse.psu.edu.

33. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W "PipMaker - A Web server for aligning two genomic DNA sequences."Genome Research 10 (4): 577-586.

34. Cetartiodactyls: the monophyletic group including both cetaceans (whale) and artiodactyls (cows, camels and pigs).

35. Nikaido, M., Rooney, A. and N. Okada (1999). "Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: Hippopotamuses are the closest extant relatives of whales." *Proceedings of the National Academy of Sciences*. 96. 10261-10266.

36. Mighell, A.J., Smith, N.R., Robinson, P.A. and A.F. Markham (2000). "Vertebrate Pseudogenes." *FEBS Letters*. 468, 109-114.

37. Grimsley, C., Mather, K.A., and C. Ober (1998). "A pseudogene with increased variation due to balancing selection at neighboring loci." Molecular Biology and Evolution. 15 (12): 1581-1588.

38. The amino acid composition of a pseudogene is used here to indicate the amino acid composition that a hypothetical protein encoded by the gene would have.