A Survey of Intron Research in Genetics

Annie S. Wu[1] and Robert K. Lindsay[2]

[1] Artificial Intelligence Laboratory, University of Michigan, Ann Arbor, MI 48109-2110, aswu@eecs.umich.edu
[2] Mental Health Research Institute, University of Michigan, Ann Arbor, MI 48109-0720, lindsay@umich.edu

Abstract. A brief survey of biological research on non-coding DNA is presented here. There has been growing interest in the effects of non-coding segments in evolutionary algorithms (EAs). To better understand and conduct research on non-coding segments and EAs, it is important to understand the biological background of such work. This paper begins with a review of basic genetics and terminology, describes the different types of non-coding DNA, and then surveys recent intron research.

1 Introduction

There has been growing interest in the effects of non-coding segments in evolutionary algorithms (EAs). Non-coding segments, also called non-coding material or introns in the literature, is a computational model of what is known as non-coding DNA in biological systems. Simply put, non-coding segments refer to the portions of an individual that make no contribution to its fitness value. In genetic programming (GP) systems, non-coding material is a natural by-product of the evolutionary process [19] [26] [29] [30] [28]. In genetic algorithm (GA) systems, studies have included both manually inserted non-coding segments and evolved segments [8] [17] [21] [23] [38] [39] [37]. Both theoretical and empirical studies suggest that non-coding segments may encourage the recombination of and discourage the destruction of existing building blocks in EAs. Evidence indicates that non-coding segments have a stabilizing effect, improving the EA's ability to preserve good building blocks. All of these qualities are desirable in an EA.

Interestingly, there seem to be many parallels between the computational arguments for non-coding segments and the biological hypotheses and explanations for non-coding DNA. To better understand and conduct research on non-coding segments and EAs, it is necessary to understand the biological inspirations of such work. The goal of this paper is to present a brief survey of the research on biological non-coding DNA and introns. This paper begins with a review of basic genetics and terminology, describes the different types of non-coding DNA, and then surveys recent research on non-coding DNA.

2 Basic Genetics

The study of genetics is the study of how living organisms reproduce and evolve. In trying to understand how entire organisms reproduce, biologists have had to

... A A T C G A G G T C C T C G G A ...
... T T A G C T C C A G G A G C C T ...

Fig. 1. Chromosomes consist of complementary strands of DNA nucleotides.

study the cellular and molecular biology of organisms. There are two fundamentally distinct types of cells, eukaryotes and prokaryotes. *Eukaryotes* are cells that have membrane bound organelles, a membrane bound nucleus containing the genetic material of the cell, and introns in the genome. *Prokaryotes* are cells which lack a membrane bound nucleus and membrane bound organelles and store genetic material in a large single molecule of DNA. All prokaryotic organisms are single celled; eukaryotic organisms may be single or multi celled.

Proteins, which are considered the building blocks of life, are the most abundant type of organic molecule in living organisms. A *protein* is made up of one or more polypeptide chains. A *polypeptide chain* is a chain of amino acids. An *amino acid* is an organic molecule consisting of a carbon atom bonded to one hydrogen atom, to a carboxyl group, to an amino group, and to a side group which varies from amino acid to amino acid. There are 20 different amino acids of genetic importance. The order of the amino acids in the polypeptide chains and the folding structure of the polypeptide chains are what give a protein its structural or functional capabilities.

When an organism reproduces, it is imperative that the instructions for building its proteins are reproduced accurately and completely. These instructions are largely maintained by a second type of organic molecule. *Nucleotides* are organic molecules that consist of a five carbon sugar, a phosphate group, and a nitrogenous base. Nucleotides are joined together to form large molecules called *nucleic acids*. The two most common types of nucleic acids are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). DNA is made up of four different nucleotides: adenine, guanine, cytosine, and thymine, abbreviated A, G, C, T. A and T are complementary; G and C are complementary. A molecule of DNA is organized in the form of two complementary chains of nucleotides (see figure 1) wound in a double helix. In eukaryotes, DNA combines with proteins to form *chromosomes*. Chromosomes are found in the nucleus of a cell and the complete set of chromosomes of an organism is called its *genome*. DNA is the genetic material that is propagated from generation to generation, and contains the instructions on how to build the proteins necessary for a particular organism. Though all genetic information is stored in the ordering of the nucleotides in the DNA, DNA is not directly involved in protein synthesis. DNA directs protein synthesis by sending instructions in the form of RNA. RNA is a nucleic acid similar to DNA and is also made up of four types of nucleotides: adenine, guanine, cytosine, and uracil, abbreviated A, G, C, U. In RNA, thymine is replaced by uracil and A and U are complementary. RNA carries out the synthesis of proteins from the DNA instructions.

A *gene* is a segment of DNA that codes for an RNA product. The different

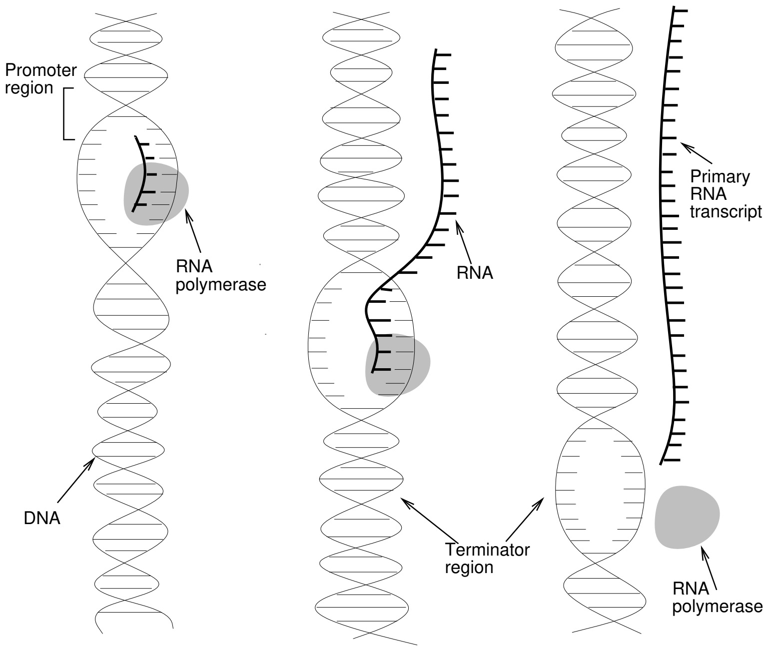

Fig. 2. During transcription, one strand of DNA of a gene is copied into RNA. This figure was adapted from figure 13.3 of [35].

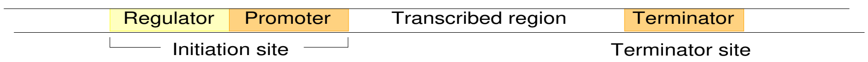

Fig. 3. A gene is bound by initiation and terminator sites.

values of a gene are called *alleles.* The synthesis of proteins from DNA occurs in two steps: transcription and translation. During *transcription*, the DNA of a gene is copied into RNA (see figure 2). Only one strand of the DNA in a chromosome is transcribed. A gene is bounded by its initiation and terminator sites as shown in figure 3. Initiation sites contain zero or more regulator regions and a promoter region. Regulator regions inhibit or allow the expression of a gene. The promoter regions are recognized by an enzyme called RNA polymerase as starting points for transcription. Transcription of the gene continues until the RNA polymerase encounters the terminator site. At this point, transciption ends and the RNA transcript and RNA polymerase are released from the DNA.

There are several types of RNA products. Three of these types — messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA) — have specific functions in the translation step of protein synthesis. During *translation*, mRNA, tRNA, and ribosomes, which are made up of rRNA and proteins, work together to build a protein (see Figure 4). The mRNA contains the ordering of the amino acids, as copied from the DNA, for the protein to be created. This

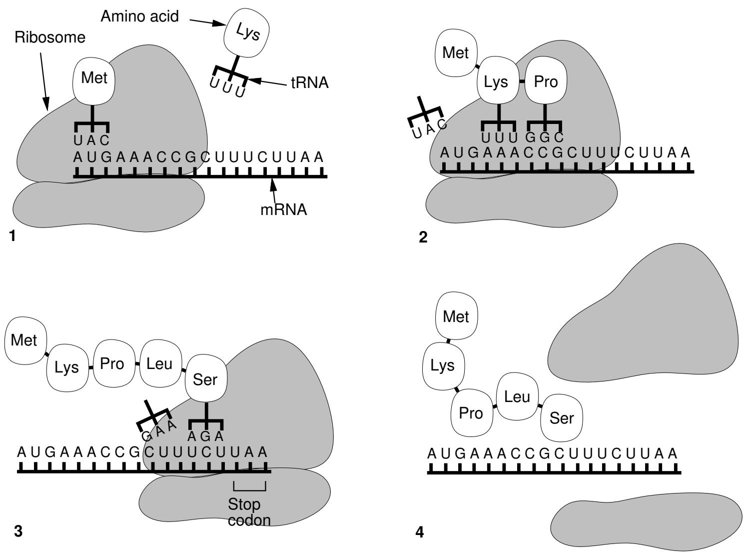

Fig. 4. During translation, three types of RNA work together with ribosomal proteins to build a protein from individual amino acids.

UUU UUC Phenylalanine UUA UUG Leucine	UCU UCC UCA UCG Serine	UAU UAC Tyrosine UAA UAG STOP	UGU UGC Cysteine UGA STOP UGG Tryptophan
CUU CUC CUA CUG Leucine	CCU CCC CCA CCG Proline	CAU CAC Histidine CAA CAG Glutamine	CGU CGC CGA CGG Arginine
AUU AUC Isoleucine AUA AUG Methionine	ACU ACC ACA ACG Threonine	AAU AAC Asparagine AAA AAG Lysine	SGU AGC Serine AGA AGG Arginine
GUU GUC GUA GUG Valine	GCU GCC GCA GCG Alanine	GAU GAC Aspartic acid GAA GAG Glutamic acid	GGU GGC GGA GGG Glycine

Fig. 5. The genetic code. Each codon represents one amino acid or termination sequence.

ordering is stored in the form of *codons*, triplets of nucleotides that represent either an amino acid or a termination signal. Since each codon is three nucleotides long and there are four possible nucleotides for each location, there are a total of $4^3 = 64$ different codons: 61 representing amino acids and three that terminate protein synthesis. Figure 5 shows the entire genetic code. The tRNA "reads" the mRNA three nucleotides at a time and retrieves the correct amino acid from the cytoplasm. The ribosome attaches to the mRNA and moves sequentially down the mRNA chain. As the tRNA retrieve amino acids in the correct order, the

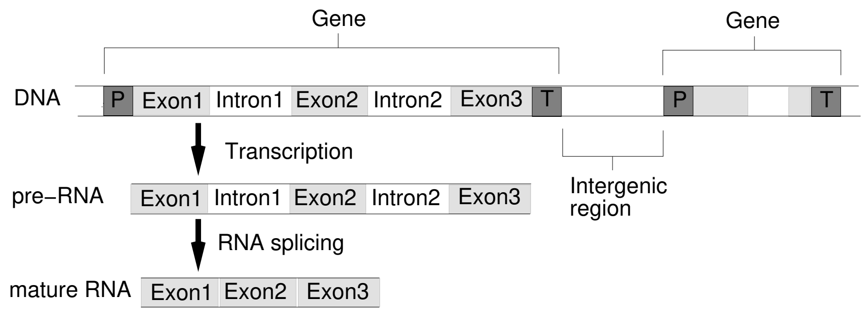

Fig. 6. Non-coding DNA: intragenic regions and introns.

ribosome attaches each new amino acid to the growing polypeptide chain. When the end of the mRNA chain is reached, the ribosome separates from the mRNA and releases a complete polypeptide chain. [4] [22] [35] [36].

3 Non-coding DNA

The term non-coding DNA refers to all DNA that is not involved in the coding of a mature RNA product. Though non-coding DNA is prevalent in biological systems, its origin and function are as yet uncertain. Because a great deal of extra energy is required to sustain and process non-coding DNA, it must not contribute negatively to the genetic process or it would most likely have been eliminated by natural selection long ago. There are three types of non-coding DNA: intergenic regions, intragenic regions, and pseudogenes [27] [22].

Though genes lie linearly along chromosomes, they are not necessarily contiguous. *Intergenic regions* are the regions of DNA in between genes. These regions are not transcribed into RNA. Some portions of intergenic regions are known to regulate the expression of adjacent genes; other portions have no known function. *Intragenic regions*, also called *introns*, are segments of DNA found within genes. Introns are transcribed into RNA along with the rest of the gene but must be removed from the RNA before the mature RNA product is complete. RNA that still contains the intron regions is often called *pre-RNA*. After the introns are spliced out of the pre-RNA, the remaining segments of RNA, the *exons* or *expressed regions*, are joined together to become the mature RNA product. Figure 6 shows an example of intergenic regions, introns, and exons.

The third type of non-coding DNA is the pseudogene. A *pseudogene* is a segment of DNA that is similar to a functional gene, but contains nucleotide changes that prevent its transcription or translation. Pseudogenes are believed to arise from gene duplication or reverse RNA transcription. Reverse RNA transcription refers to the transcription of RNA into DNA. Interestingly, pseudogenes produced from reverse transcription do not contain introns. Since pseudogenes are not expressed, they are not subject to selection pressure from the environment. As a result, pseudogenes accumulate mutations quickly. When a pseudogene mutates enough that its similarity to a functional gene is no longer apparent, it becomes simply non-coding intergenic DNA.

4 Intron Research

The existence of the intron-exon structure has been particularly intriguing. Introns are only found in eukaryotic genomes and make up a large portion of the DNA in eukaryotic genomes. In humans, for example, approximately 30% of the human genome is made up of introns [1]. Only about 3% consists of coding DNA and the rest of the genome consists of other non-coding DNA, repetitive sequences, and regulatory regions. The unusual placement of introns, interrupting the coding regions of genes, and the fact that extra energy is needed to maintain and process these structures that have no apparent function, have made introns an important topic of study since their discovery in the 1970's. Intron research has focused for the most part on three questions: (1) how are they removed from the RNA, (2) what do they do, and (3) what is their origin?

Of these three issues, the first one is probably the most well understood. Introns are removed from RNA by a process called *RNA splicing* which occurs in the nucleus of a cell [22] [31] [32]. There are many different methods of RNA splicing [22] [3] [7]. Most of the splicing processes require the aid of proteins. Proteins recognize specific sequences in the pre-RNA to catalyze the splicing process. The majority of introns in this group follow the GT-AG rule: the intron begins with the dinucleotide GT and ends with the dinucleotide AG. Certain genes also allow for *alternative splicing*, a situation where one gene codes for more than one RNA sequence depending on how many pre-RNA segments are spliced out. Other splicing processes such as that of fungal mitochondria introns involve self-splicing RNA [22] [11]. Though proteins assist in these self-splicing processes, all information necessary for the reaction resides in the intron sequence.

The second question — what do introns do? — continues to be studied. The *exon theory of genes* [2] suggested that exons are the building blocks of proteins, and genes are created from combinations of these building blocks. This theory lead to the *exon shuffling hypothesis* [9] [10] [12] [13] which states that introns increase the rate of recombination of exons and make it easier to move exons around and create new genes. Statistically, "...introns represent hot spots for recombination: by their mere presence and length they increase the rate of recombination, and hence the shuffling of exons, by factors of the order of 10^6 or 10^8 [12, pg. 901]." Evidence suggests that exons may correspond to both structural and functional subunits of proteins [2] [16] [18] including specific examples of the same exon existing in different genes where the same structure or function is required by two different proteins [10]. According to this theory then, the intron-exon structure of eukaryotic genes encourages the formation of new genes from structural and functional subunits of existing genes. This process would certainly be more efficient than building new genes one nucleotide at a time.

If introns are so useful for recombination, why are they found only in eukaryotes and not in prokaryotes? This difference raises the third question: what is the

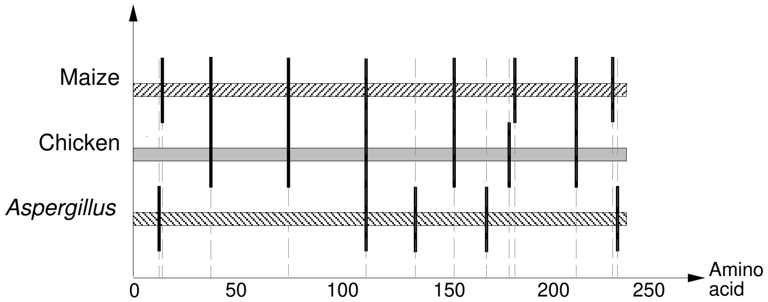

Fig. 7. A comparison of the introns locations in three different TPI genes. Each horizontal shaded bar represents the amino acid sequence created by one TPI gene. The bold vertical lines show the approximate locations of the introns in the RNA templates that coded for the amino acid sequences shown.

origin of introns? There are two main schools of thought. The "introns-early" theory asserts that the ancestral organisms of both eukaryotes and prokaryotes possessed introns and that prokaryotes lost introns in the evolutionary process. The "introns-late" theory asserts that ancestral organisms did not possess introns and that eukaryotes gained introns in the evolutionary process.

The introns-early theory suggests that the last common ancestor of prokaryotes and eukaryotes had introns in its genome [5] [6]. To accommodate their short reproductive and life cycles, prokaryotes subsequently lost the introns from their genomes due to selection for increased efficiency in gene expression and for a reduction in genome size. The price paid for this increased efficiency was a decreased potential for future evolution due to the loss of the introns' assistance in exon shuffling. Eukaryotes, on the other hand, continued to evolve with the assistance of introns and have been able to develop much more complex and diverse organisms. Accordingly, we currently find much less complexity and variation in prokaryotes than in eukaryotes.

Research on the gene for the protein triosephosphate isomerase (TPI) has pushed the known existence of the intron structure back before the divergence of plants and animals [25] [15] [14]. TPI is an extremely old protein whose gene sequence is relatively conserved across all organisms. Studies on the introns of this gene found five introns in *Aspergillis*, six introns in chickens and humans, and eight introns in maize. Five of the introns from the chicken and maize genes were found at identical locations in the corresponding genes; one intron occurred at similar locations on the two genes, differing by only three amino acid positions; and the maize gene had two additional introns. The similarity between *Aspergillis* and maize was less apparent, but still substantial. One intron was found at the same location in both genes. Two others were found at similar locations, and two introns in the *Aspergillis* gene occurred at completely novel locations compared to the chicken and maize genes. Figure 7 shows the approximate locations of the introns in the amino acid sequences of these three TPI genes. "The striking agreement of five of the intron positions in TPI between maize and vertebrates suggests that all of these introns were in place before the division of plants and animals [15, pg. 151]." Random insertion of introns into these genes would be hard pressed to achieve such a high rate of similarity. Though these findings do not prove the existence of introns in the last common ancestor of eukaryotes and prokaryotes, they do support an early origin for introns and suggest an evolutionary tendency towards the loss of introns rather than random insertion of introns in eukaryotic genomes.

In addition to the similar positions of introns in the same gene of different organisms, there are a number of statistical measurements and estimations of introns and exons that discourage the belief of random insertion of introns into genes. Among these measurements are the distribution of the lengths of introns and exons [22] [14] and the positions of introns with respect to the codons [24]. In addition, from known exon sizes and intron positions, it has been possible to predict the positions of introns that have been lost from one species but may still exist in another [14] [16].

The introns-late theory suggests that introns developed in the eukaryotic evolutionary process. [3] [33] [7]. Since prokaryotes have traditionally been considered more primitive than eukaryotes, the even-more-primitive genome of the common ancestor of prokaryotes and eukaryotes is often assumed to resemble the tightly organized prokaryotic genome. The introns-late theory contends that introns were inserted into eukaryotic genomes some time after the division of the prokaryotic and eukaryotic lines of evolution. Proponents of a late insertional origin of introns argue that the data supporting the exon theory of genes is intermittent and thus not solid enough to favor an early origin of introns [34]. There is a growing interest in the different classes of introns and the appearance and distribution of these classes in the genomes of organisms. A study of the different classes of introns showed that the relationship between the classes is related to the phylogenetic organization of the organisms in which they appear [3]. This suggests that introns arose and evolved in eukaryotic genomes. It has been speculated that introns could have arisen from gene duplication, transposable elements, or self insertion [22] [33].

5 Summary

EAs have successfully incorporated many ideas from biological systems into computational search algorithms, including that of non-coding material. This paper reviews the basics of genetics and surveys recent research on biological non-coding DNA. Though the function of introns is not completely understood and the benefits of non-coding segments are not yet certain, a number of parallels exist between biological hypotheses on introns and computational hypotheses on non-coding segments.

First of all, both introns and non-coding segments are thought to separate building blocks of what is being evolved. Introns (and intergenic regions) separate a segment of DNA into exons which are thought to code for functional or structural subunits of proteins. Building or modifying proteins from such subunits is expected to be easier and faster than building proteins one nucleotide at a time. The discovery and exchange of building blocks or partial solutions is one of the unique aspects of evolutionary search algorithms. According to the *building block hypothesis* [20] such algorithms are expected to search for multiple building blocks in parallel and recursively combine these building blocks to form a complete solution.

Secondly, both introns and non-coding segments are thought to increase the rate of recombination during evolution. Combined with the first point above, the extra material in a genome or individual is expected to increase the chance of crossover combining existing building blocks and decrease the chance of crossover destroying any useful material. Specifically, the exon shuffling hypothesis theorizes that introns increase the recombination rate of exons and assist in the creation of new genes from exon building blocks. The exact same argument may be made for the building blocks of an EA system.

Third, the ability to dynamically evolve the placement of introns and non-coding segments appears to be important. Biological organisms with the same gene have been found to have similar but not identical collections of introns. There is also the issue of why prokaryotes do not have introns. A number of computational systems have investigated the evolution of non-coding segments [17] [30] [38] allowing the EA to determine both the placement and arrangement of information on an individual.

Acknowledgements

This research was supported by NASA grant NGT-51057. The authors would like to thank John Holland for many interesting discussions relating to this work.

References

1. G. I. Bell and T. G. Marr, editors. *Computers and DNA*. Addison-Wesley, 1988.
2. C. C. F. Blake. Do genes-in-pieces imply proteins-in-pieces? *Nature*, 273:267, 1978.
3. T. Cavalier-Smith. Intron phylogeny: a new hypothesis. *Trends in Genetics*, 7(5):145–148, May 1991.
4. H. Curtis. *Biology*. Worth Publishers, 1983.
5. W. F. Doolittle. Genes in pieces: were they ever together? *Nature*, 272:581, 1978.
6. W. F. Doolittle. What introns have to tell us: Hierarchy in genome evolution. *Cold Spring Harbor Symposia on Quantitative Biology*, 52:907–913, 1987.
7. A. Flavell. Introns continue to amaze. *Nature*, 316:574–575, August 1985.
8. S. Forrest and M. Mitchell. Relative building-block fitness and the building-block hypothesis. In *FOGA*, 1992.
9. W. Gilbert. Why genes in pieces? *Nature*, 271:501, February 1978.
10. W. Gilbert. Genes-in-pieces revisited. *Science*, 228:823–824, May 1985.
11. W. Gilbert. The RNA world. *Nature*, 319:618, February 1986.
12. W. Gilbert. The exon theory of genes. *Cold Spring Harbor Symposia on Quantitative Biology*, 52:901–905, 1987.

13. W. Gilbert. Gene structure and evolutionary theory. In *New perspectives on evolution*, pages 155–163. Wiley-Liss, 1991.
14. W. Gilbert and M. Glynias. On the ancient nature of introns. *Gene*, 135, 1993.
15. W. Gilbert, M. Marchionni, and G. McKnight. On the antiquity of introns. *Cell*, 46:151–153, July 1986.
16. M. Go. Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature*, 291:90–92, May 1981.
17. D. E. Goldberg, B. Korb, and K. Deb. Messy genetic algorithms: Motivation, analysis, and first results. *Complex Systems*, 3:493–530, 1989.
18. D. L. Hartl. New perspectives on the molecular evolution of genes and genomes. In *New perspectives on evolution*, pages 123–137. Wiley-Liss, 1991.
19. T. Haynes. Duplication of coding segments in genetic programming. In *13th AAAI*, 1996.
20. J. H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.
21. J. R. Levenick. Inserting introns improves genetic algorithm success rate: Taking a cue from biology. In *ICGA-4*, pages 123–127, 1991.
22. B. Lewin. *Genes 5*. John Wiley & Sons, 1994.
23. R. K. Lindsay and A. S. Wu. Testing the robustness of the genetic algorithm on the floating building block representation. In *13th AAAI*, 1996.
24. M. Long, C. Rosenberg, and W. Gilbert. Intron phase correlations and the evolution of the intron/exon structure of genes, 1995. Under review.
25. M. Marchionni and W. Gilbert. The triosephosphate isomerase gene from maize: introns antedate the plant-animal divergence. *Cell*, 46:133–141, July 1986.
26. N. F. McPhee and J. D. Miller. Accurate replication in genetic programming. In *ICGA-6*, 1995.
27. M. Nei. *Molecular Evolutionary Genetics*. Columbia University Press, 1987.
28. P. Nordin and W. Banzhaf. Complexity compression and evolution. In *ICGA-6*, 1995.
29. P. Nordin and W. Banzhaf. Evolving turing-complete programs for a register machine with self modifying code. In *ICGA-6*, 1995.
30. P. Nordin, F. Francone, and W. Banzhaf. Explicitly defined introns and destructive crossover in genetic programming. Workshop on GP, ML, 1995.
31. B. Patrusky. The intron story. *MOSAIC*, 23(3):22–33, Fall 1992.
32. M. Robertson. The post-RNA world. *Nature*, 335:16–18, September 1988.
33. J. H. Rogers. How were introns inserted into nuclear genes? *Trends in Genetics*, 5(7):213–216, July 1989.
34. A. Stoltzfus, D. F. Spencer, M. Zuker, J. M. Logsdon, Jr., and W. F. Doolittle. Testing the exon theory of genes: the evidence from protein structure. *Science*, 265:202–207, July 1994.
35. R. A. Wallace, G. Sanders, and R. Ferl. *Biology: The Science of Life*. Harper College, 3rd edition, 1991.
36. J. D. Watson. *Molecular Biology of the Gene*. W. A. Benjamin, 2nd edition, 1970.
37. A. S. Wu. *Non-coding DNA and floating building blocks for the genetic algorithm*. PhD thesis, University of Michigan, 1995.
38. A. S. Wu and R. K. Lindsay. A comparison of the fixed and floating building block representation in the genetic algorithm, 1995. Submitted to *Evol. Comp.*
39. A. S. Wu and R. K. Lindsay. Empirical studies of the genetic algorithm with non-coding segments. *Evolutionary Computation*, 3(2), 1995.