

# Word Seeker: Discovering Genome-wide Patterns

Lonnie Welch<sup>1,2,3</sup>, Eric Petri<sup>1</sup>, Dazhang Gu<sup>1</sup>, Klaus Ecker<sup>1</sup>

<sup>1</sup>School of Electrical Engineering and Computer Science

<sup>2</sup>Biomedical Engineering Program

<sup>3</sup>Molecular and Cellular Biology Program

Ohio University, Athens, OH 45701

The purposes of most genomic information are unknown. This limits our ability to understand and address problems that have genetic causes. Does the 'junk' portion of genomes have biological meaning? If so, what is the meaning? What are the biological words, phrases, grammar, etc.? The answers will lead to a more complete understanding of the purpose of the genome and the functions of undiscovered genomic elements. This knowledge will help to cure problems that are due genetic causes. We have implemented a Word Seeker tool as illustrated in two data flow diagrams below. Using suffix tree and Teiresias algorithms, the tool discovered elements in Arabidopsis (a model plant genome) that occur with unexpected frequencies in the 'junk' portion of the genome, and they are found to be statistically overrepresented. Such elements may form biological words, phrases, and grammar which have biological functions. As one biologist put it, there is no *junk* DNA.

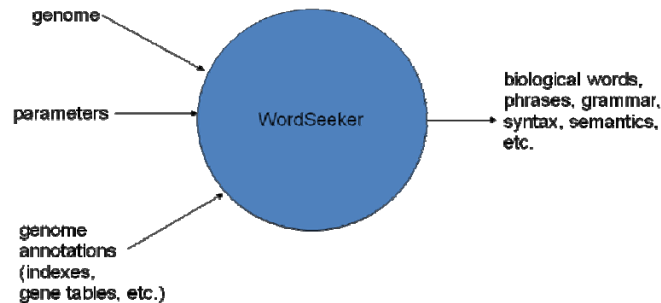
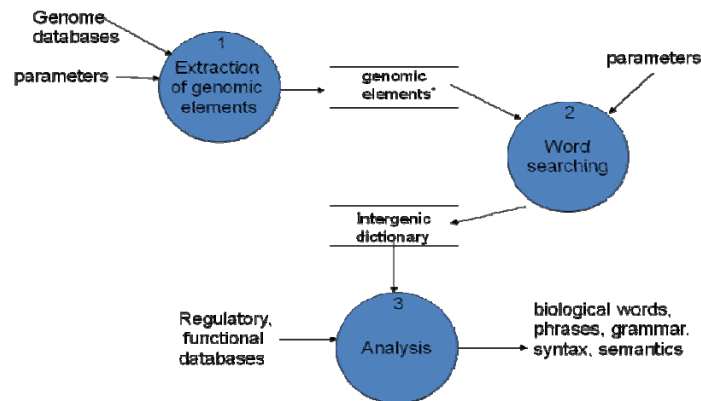


Figure 1 Level 1 data flow diagram



\*Introns, exons, UTRs, intergenic regions, and promoters

Figure 2 Level 2 data flow diagram