The *WORDIFIER* Pattern for Functional and Regulatory Genomics

Lonnie R. Welch, Jens Lichtenberg, and Frank Drews
Bioinformatics Laboratory, Ohio University, Athens, Ohio 45701, {welch|lichtenj}@ohio.edu

According to Christopher Alexander, the father of the 'design patterns' movement, a pattern "describes a problem which occurs over and over again in our environment, and then describes the core of the solution to that problem, in such a way that you can use the solution a million times over, without ever doing it the same way twice [1]." This paper use Alexander's pattern format to present *WORDIFIER*, a design pattern from the domain of functional and regulatory genomics.

With the availability of the genomic sequences of numerous organisms, life scientists are working in conjunction with bioinformaticians to decipher the meanings of the genomes. Projects such as Encyclopedia of Genomic Elements (ENCODE) [2] seek to identify and charatcetrize the functional elements in genomes. The functional elements are often referred to as *words*.

♦ ♦ ♦

Given a genomic sequence (or a set of sequences), an important problem is the enumeration of all subsequences (words) contained in the sequence (or the set of sequences).

The enumeration of all words in a sequence (or a set of sequences) is a fundamental operation in the study of genomic sequences. The enumerated words provide the basic elements that are further characterized and studied. Thus, the bioinformatics open source community has provided solutions to this problem.

To enumerate the words in a sequence, the sequence is decomposed into subsequences and the set of unique subsequences (words) is constructed. The number of instances of each unique word, the word frequency, is counted.

If multiple sequences are analyzed, the process is repeated for each sequence and the union of sets of unique words is constructed. The number of sequences containing each word, the word sequence frequency, is counted.

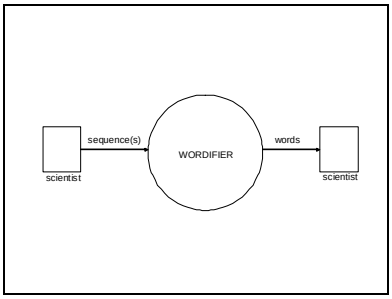

Figure 1. WORDIFIER solution diagram.

♦ ♦ ♦

The WORDIFIER makes use of the SEQUENCE pattern and the WORD pattern. Many operations are performed to characterize the product of WORDIFIER, such as SCORING words, ORDERING word sets, and EXTRACTING subsets of words based on various criteria.

Future work will include the use of POSA and GOF pattern formats to provide a more detailed description of the WORDIFIER pattern. Additionally, other patterns from the domain of regulatory genomics will be codified. A pattern language that unifies the set of patterns will also be developed.

REFERENCES

[1] C. Alexander, S. Ishikawa, and M. Silverstein, *A Pattern Language: Towns, Buildings, Construction*. Oxford University Press, 1977.

[2] E. Birney, Stamatoyannopoulos J. A., Dutta A., Guigo R., Gingeras T. R., Margulies E. H., Weng Z., Snyder M., Dermitzakis E. T., Thurman, R. E., et al., "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project," Nature **447**:799–816, 2007.