Title: "Open Source Implementation of Batch-Extraction for Coding and Non-coding Sequences"

Authors: Jens Lichtenberg, Lonnie Welch

Affiliations: School of EECS, Ohio University, Athens, Ohio 45701, USA, lichtenj@ohio.edu

URL for project: http://opensource.msseeker.org

URL for code: http://extractor.msseeker.org

Open Source License being used: The GNU General Public License (GPL)

Abstract

Regulatory genomics is focused on the discovery of general characteristics explaining the co-regulation of genes. Most applications used to discover co-regulation, return lists of putative co-regulated genes and their respective expression values. In order to determine shared regulatory elements it is necessary to analyze the non-coding and in some cases coding sequences of these genes.

While it is possible to extract promoter sequences, exons, introns, intergenic regions and untranslated regions for single genes or entire genomes through tools like the UCSC Genome Browser or directly through databases like Ensembl, it is difficult to extract these elements in an automated fashion for a batch of gene symbols.

Based on the available Ensembl API a Perl based open source solution to extract non-coding as well as coding sequences is presented. The tool reads in a list of gene symbols and extracts exons, introns, 5' and 3' untranslated regions, coding sequences and promoters (of variable length) for each gene in the list. The sequences are stored in fasta files separated based on the segment characteristic of the extracted sequence (e.g. 1 fasta file for all exons of all supplied gene symbols).

Due to limitations of the annotations within Ensembl it is sometimes not possible to extract the entire set of annotated sequences. In such cases the user is notified.

This application is integrated into an existing enumerative motif discovery framework (WordSeeker) in order to extract the desired sequences for a gene list automatically and supply it to the subsequent motif discovery phase. Due to its open source nature it is also possible to integrate this tool into other existing motif discovery frameworks (enumerative and alignment-based), sequence analysis frameworks, or as a post-processing stage in microarray, ChIP-chip or proteomics experiments.

To enhance the visibility of the tool, it will be published to the CPAN and BioPerl repositories in the near future.

An Open Source Framework for Bioinformatics Word Enumeration and Scoring

Kyle Kurz, Jens Lichtenberg, Lee Nau, Dr. Frank Drews, Dr. Lonnie Welch
Ohio University, welch@ohio.edu

Main Project Page: http://bio-s1.cs.ohiou.edu/~wordseek
Download Page: http://bio-s1.cs.ohiou.edu/~wordseek/download

The software package presented here provides an open source framework for word enumeration (and subsequent scoring of those words) within biological sequence data. Using this framework, developers may choose to implement any of a number of algorithms to perform the enumeration, with no change to the execution logic of the framework.

Two major problems are addressed with our framework, one in the field of biology and one in computer science. Biological research creates enormous amounts of genomic data for each experiment performed. From there, the biologists must usually select a subset of the "words" (short DNA nucleotide subsequences) for further analysis. Bioinformatics provides computerized tools to aid biologists in the pruning process by providing information about statistically interesting words, under the assumption that over and under-represented words should provide some real biological function.

Through the creation of a modular framework, different algorithms can be used to accommodate the requirements of a specific job. Our framework allows a highly scalable software package, as it does not have the limitations of being tied to a single enumeration or scoring algorithm, and the best algorithm for a dataset can be selected at runtime. The implementation of multiple algorithms allows the user to focus on the job-specific optimizations such as high speed or low memory footprint. This flexibility is often not possible with single algorithm tools. By building a minimal set of requirements for functionality based on our WordSeeker [1] tool and analysis of other enumerative tools such as Weeder[2] and YMF[3], we have been able to provide a highly abstract system with interchangeable modules for the various stages, allowing the framework to grow and mature as new algorithms and methods are developed.

Using object-oriented software design and class abstraction, our framework is not only extensible, but easily modifiable. The base classes form a frame around two major phases, the enumeration of words and the scoring of those words. Built in C++, the framework utilizes virtual functions to provide consistent interfacing between components, regardless of the underlying algorithms. Well documented abstract class definitions provide a description of the minimal set of functions a developer must implement, as well as allowing the extension of algorithm specific functionality through standard interfaces. Future work will provide additional post-processing functionality through similar abstract interfaces.

In summary, our framework streamlines the development process for new techniques and modules in a general bioinformatics toolkit and facilitates the research process by allowing the selection of various implementations based on the biologists' needs for a given dataset.

1. J. Lichtenberg, M. Alam, T. Bitterman, F. Drews, K. Ecker, L. Elnitski, S. Evans, E. Grotewold, D. Gu, E. Jacox, K. Kurz, S. S. Lee, X. Liang, P. M. Majmudar, P. Morris, C. Nelson, E. Stockinger, J. D. Welch, S. Wyatt, A. Yilmaz, and L. R. Welch, "Construction of Genomic Regulatory Encyclopedias: Strategies and Case Studies," *Proceedings of the Ohio Collaborative Conference on Bioinformatics*, IEEE Computer Society press, June 2009

2. G. Pavesi, P. Mereghetti, G. Pesole, *Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes*. Nucleic Acids Research 2004 Jul 1;32(Web Server issue): W199-W203.

3. S. Sinha and M. Tompa. *YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation.* Nucleic Acids Research 2003 Vol. 31, No. 12 3586-3588