Toward a unified framework for motif discovery methods

Lonnie Welch, Sophie S. Schbath, Finn Drablos, Frank Drews, Mark M. Hoebeke, Geir Kjetil Sandve, Rami Al-Ouran, Xiaoyu Liang, Kyle Kurz

Toward a Unified Framework for Motif Discovery Methods

Lonnie Welch, Sophie Schbath, Finn Drablos, Frank Drews, Mark Hoebeke,
Geir Kjetil Sandve, Jens Lichtenberg, Rami Al-Ouran, Xiaoyu Liang, Kyle Kurz

High throughput sequencing technologies are producing a plethora of sequenced genomes which need to be annotated. In addition to annotating open reading frames in genomes, scientists are identifying the regulatory components of genomes. An important bioinformatics method that is employed in this context addresses the problem of *motif discovery*, which seeks interesting, functional subsequences from genomic data.

Bioinformatics researchers have developed a diverse set of methods to solve the motif discovery problem. Existing techniques can be divided into 2 major classes: alignment-based approaches, such as MEME, and enumerative approaches, such as YMF. While there are many different methodologies, no optimal technique has emerged. Thus, there is a need for continued research on methods to solve the motif discovery problem.

Some researchers have developed ensemble approaches, which combine outputs of multiple motif discovery tools. While ensembles can improve results, they do not provide a unified model, and they provide little flexibility for how tool features are combined. Thus, a common framework is needed.

The authors present *OpenMotif*, a framework and open source repository (http://code.google.com/p/open-motif/) that provides a unifying model for motif discovery. The source code of OpenMotif is available under GNU General Public License v3. The authors have integrated statistical models, a scalable word enumeration algorithm, and a module discovery method into OpenMotif. This has resulted in a software architecture and a Galaxy-based workflow management interface. OpenMotif has been applied to several test cases, including genomic data of *E. coli, A. thaliana,* and *H. sapiens.*