**Sequence analysis**

**TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs**

D. Benjamin Gordon¹, Lena Nekludova¹, Scott McCallum¹ and Ernest Fraenkel¹,²,*

¹Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA and ²MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Cambridge, MA 02139, USA

Received on February 18, 2005; revised on April 14, 2005; accepted on April 30, 2005
Advance Access publication May 19, 2005

**ABSTRACT**

**Summary:** TAMO (Tools for Analysis of MOtifs) is an object-oriented computational framework for interpreting transcriptional regulation using DNA-sequence motifs. To simplify the application of multiple motif discovery programs to genome-wide data, TAMO provides a sophisticated motif object with interfaces to several popular programs. In addition, TAMO provides modules for integrating motif analysis with diverse data sources including genomic sequences, microarrays and various databases. Finally, TAMO includes tools for sequence analysis, algorithms for scoring, comparing and clustering motifs, and several useful statistical tests. Recently, we have applied these tools to analyze tens of thousands of motifs derived from hundreds of microarray experiments.

**Availability:** TAMO is a Python/C++ package and requires Python 2.3 or higher. Source code and documentation are available at http://web.wi.mit.edu/fraenkel/TAMO/

**Contact:** efraenkel@wi.mit.edu

**1 INTRODUCTION**

Motif discovery from genome-wide data is improved by using multiple programs in concert (Harbison et al., 2004; Tompa et al., 2005), but the management of such calculations is difficult. The purpose of the TAMO package is to provide tools that facilitate combining motif information from different motif-finding programs and interpreting motif data using biological databases. In contrast to other packages, TAMO provides an integrated motif analysis framework in a high-level language; however, it is built on fast C++ routines.

The TAMO package includes (1) command-line programs for motif discovery, scoring and analysis, and (2) source code for use by programmers to develop new analysis tools using the python language. The package provides both a simple interface and a wealth of supporting object definitions and lower-level functions for operating on motifs, microarrays, sequences and other bioinformatic data sources. These routines are supported by functions for manipulating genome sequence data and by a library of statistical tests. Finally, TAMO includes a utility that automatically downloads data from sources such as SGD and UCSC (Cherry et al., 1997; Karolchik et al., 2003).

**2 EXAMPLE USAGE**

Table 1 illustrates how the TAMO framework can be used to unify diverse algorithms and data sources. In this example, we use three programs to search for motifs among the promoters of a set of genes in yeast. We then identify and display the best-scoring motif, and test whether promoters containing it are significantly associated with any categories in the GO database (Ashburner et al., 2000).

In the example, AlignACE (Hughes et al., 2000) is run 10 times with different random number seeds and MDscan (Liu et al., 2002) is run repeatedly with different motif widths. Finally, MEME (Bailey et al., 2000) is used to find the best-scoring motif. The output is then analyzed using the GO database (Ashburner et al., 2000).

**Table 1. Sample python TAMO code**

```python
from TAMO import MotifTools
from TAMO.seq import Fasta
from TAMO.MotifMetrics import ProbeSet
from TAMO.AlignAce import AlignAce
from TAMO.MD.AlignAce import MDscan
from TAMO.MD.Meme import Meme
from TAMO.DataSources import GO

DATA = 'datafile.fasta'
IDS = Fasta.ids(DATA)
ALLMOTIFS = []
for i in range(10):
    seed = int(random.random() * 1e9)
    A = AlignAce(DATA, seed=seed)
    ALLMOTIFS.extend(A.motifs)
for width in range(6,19):
    M = MDscan(DATA, width)
    ALLMOTIFS.extend(M.motifs)
M = Meme(DATA)
ALLMOTIFS.extend(M.motifs)
PROMOTERS = ProbeSet('yeast_promoters.fasta')
for m in ALLMOTIFS:
    m.church = PROMOTERS.church(m, IDS)
MotifTools.sortby(ALLMOTIFS, 'church')
bestmotif = ALLMOTIFS[0]
print bestmotif
bestmotif.print_textlogo()
bestrmotif.giflogo('best_motif.gif')
matching_ORFs = PROMOTERS.matching_ids(bestmotif)
for P, category in GO.orfs2cats(matching_ORFs):
    print P, category
```

**Imports**

**Input**

**Motif Discovery**

**Scoring**

**Output**

**Annotation**
and Elkan, 1995) is applied to the data. Next, the sample code computes the ‘group specificity score’ of AlignACE (called ‘church’ by TAMO) for the motifs found by all three programs, identifies the motif with the best score and displays it in various ways. To help interpret the biological meaning of the motif, the sample code identifies promoters containing sub-sequences that match the motif and searches the corresponding list of ORFs for statistically over-represented GO categories.

3 PACKAGE FEATURE OVERVIEW

Motifs and motif discovery. TAMO is developed around a unified motif representation of a position-specific scoring matrix (PSSM) that can be assembled from many sources. The package also contains interfaces to publicly available motif discovery programs as well as its own internal motif discovery programs.

External data sources. Several TAMO modules provide access to public repositories of genomic information. For example, there are modules to provide access to SGD feature maps and functions for translating between different types of feature identifiers (e.g. gene name to Swiss-Prot ID, etc.) The GO module uses gene annotations from GO-slim and has facilities for finding functional categories that are statistically over-represented within a set of genes. TAMO also provides interfaces to the human Gene Atlas (Su et al., 2004), to yeast transcription rates (Holstege et al., 1998) and to other genome-wide data. Finally, TAMO provides fast, random-access interfaces to human and yeast (Saccharomyces cerevisiae) genome sequences.

Motif scoring, comparison and clustering. TAMO includes metrics for evaluating motif quality including the ‘group specificity score’ (Hughes et al., 2000), the enrichment score (Harbison et al., 2004), the ROC AUC metric (Clarke and Granek, 2003) and several others. Functions are provided for finding the optimal alignment of two motifs and for quantitatively reporting their similarity (or divergence) with several choices of distance metrics. TAMO also includes implementations of the k-medoids and UPGMA algorithms for clustering motifs.

Sequence and microarray data. TAMO has fast routines for reading, writing, manipulating and using motifs to scan large collections of sequences. A general-purpose ‘dataset’ object stores collections of microarray experiments and provides methods to quickly extract sets of genes or experiments that satisfy user-supplied P-value or ratio thresholds.

Statistics. TAMO includes a set of useful statistical routines for computing P-values for normal, binomial, Poisson and hyper-geometric distributions. The Shapiro–Wilk normality test and the Wilcoxon–Mann–Whitney rank sum test are also provided.

ACKNOWLEDGEMENTS

E.F. is a Whitehead Fellow and was funded in part by Pfizer. D.B.G. was supported by NIH/NIGMS NRSA award GM068278.

REFERENCES


