# New Motif Discovery

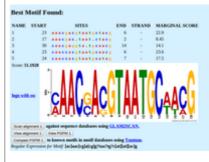六月 19, 2013, 3:26 p.m. by Rosalind Team                    Topics: Bioinformatics Tools, Proteomics

## Know Your Meme

In "Finding a Protein Motif", we used the ProSite database to help us find motifs in proteins. Yet Prosite is helpful only for finding motifs that have already been identified in other sequences; to discover novel candidate motifs, we will need to enlist other tools.

Programs that identify a new motif represent it as a matrix containing the probability of each possible symbol at each position in the pattern. They may also represent it as a regular expression that can be used in another applications, such as ProSite.

Motifs are usually displayed in the form of a sequence logo (see Figure 1) One of the most widely used programs for discovering motifs in a collection of homologous DNA or protein sequences is MEME. MEME takes as input a collection of DNA/protein sequences and outputs motifs exceeding a user-specified statistical confidence threshold. MEME is one of a collection of tools for analyzing motifs belonging to a larger MEME Suite. The MEME Suite can be found here.



Figure 1. A sample GLAM2 gapped motif.

MEME cannot discover motifs that exhibit indels, because it does not take gaps into consideration. This limitation is overcome by another program from the MEME Suite called GLAM2. Discovering motifs containing gaps is intrinsically more difficult than discovering ungapped motifs, because there are vastly more possible gapped motifs than ungapped ones.

Each motif in MEME is assigned an expected value (E), which is the likelihood of it occurring by chance. This denotes the statistical significance of the motif. GLAM2 assigns scores instead of e-values: the score favours alignment of similar residues, and disfavours insertions and deletions.

## Problem

The novel-motif finding tool MEME can be found here.

   **Given:** A set of protein strings in FASTA format that share some motif with minimum length 20.

   **Return:** Regular expression for the best-scoring motif.

## Sample Dataset

```
>Rosalind_7142
PFTADSMDTSNMAQCRVEDLWWCWIPVHKNPHSFLKTWSPAAGHRGWQFDHNFFVYMMGQ
FYMTKYNHGYAPARRKRFMCQTFFILTFMHFCFRRAHSMVEWCPLTTVSQFDCTPCAIFE
WGFMMEFPCFRKQMHHQSYPPQNGLMNFNMTISWYQMKRQHICHMWAEVGILPVPMPFNM
SYQIWEKGMSMGCENNQKDNEVMIMCWTSDIKKDGPEIWWMYNLPHYLTATRIGLRLALY
>Rosalind_4494
VPHRVNREGFPVLDNTFHEQEHWWKEMHVYLDALCHCPEYLDGEKVYFNLYKQQISCERY
PIDHPSQEIGFGGKQHFTRTEFHTFKADWTWFWCEPTMQAQEIKIFDEQGTSKLRYWADF
QRMCEVPSGGCVGFEDSQYYENQWQREEYQCGRIKSFNKQYEHDLWWCWIPVHKKPHSFL
KTWSPAAGHRGWQFDHNFFSTKCSCIMSNCCQPPQQCGQYLTSVCWCCPEYEYVTKREEM
>Rosalind_3636
ETCYVSQLAYCRGPLLMNDGGYGPLLMNDGGYTISWYQAEEAFPLRWIFMMFWIDGHSCF
NKESPMLVTQHALRGNFWDMDTCFMPNTLNQLPVRIVEFAKELIKKEFCMNWICAPDPMA
```

```
GNSQFIHCKNCFHNCFRQVGMDLWWCWIPVHKNPHSFLKTWSPAAGHRGWQFDHNFFQMM
GHQDWGTQTFSCMHWVGWMGWVDCNYDARAHPEFYTIREYADITWYSDTSSNFRGRIGQN
```

## Sample Output

```
DLWWCWIPVHK[NK]PHSFLKTWSPAAGHRGWQFDHNFF
```