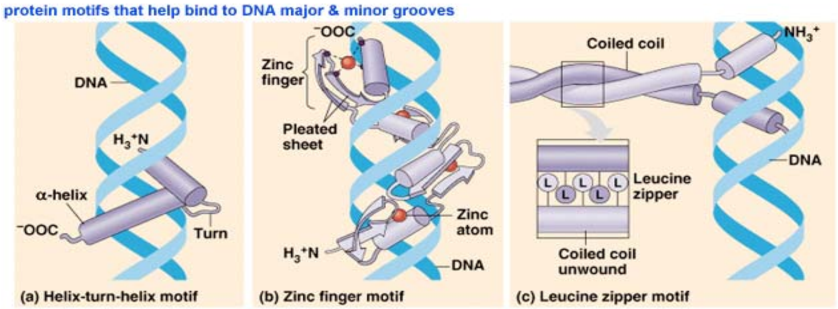
TOOLS FOR MOTIF AND PATTERN SEARCHING

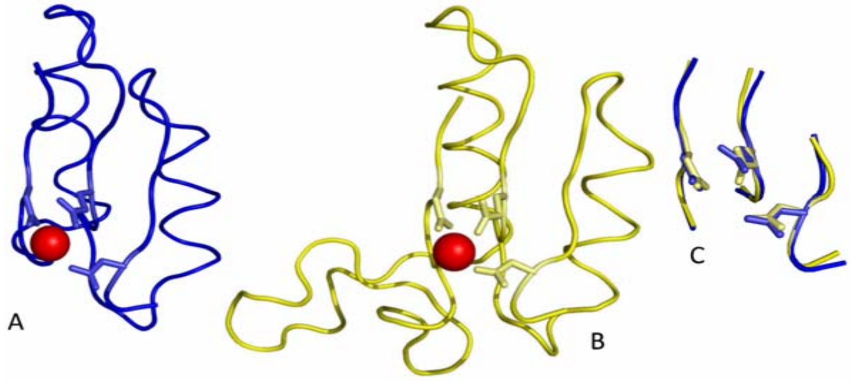
Prat Thiru

OUTLINE

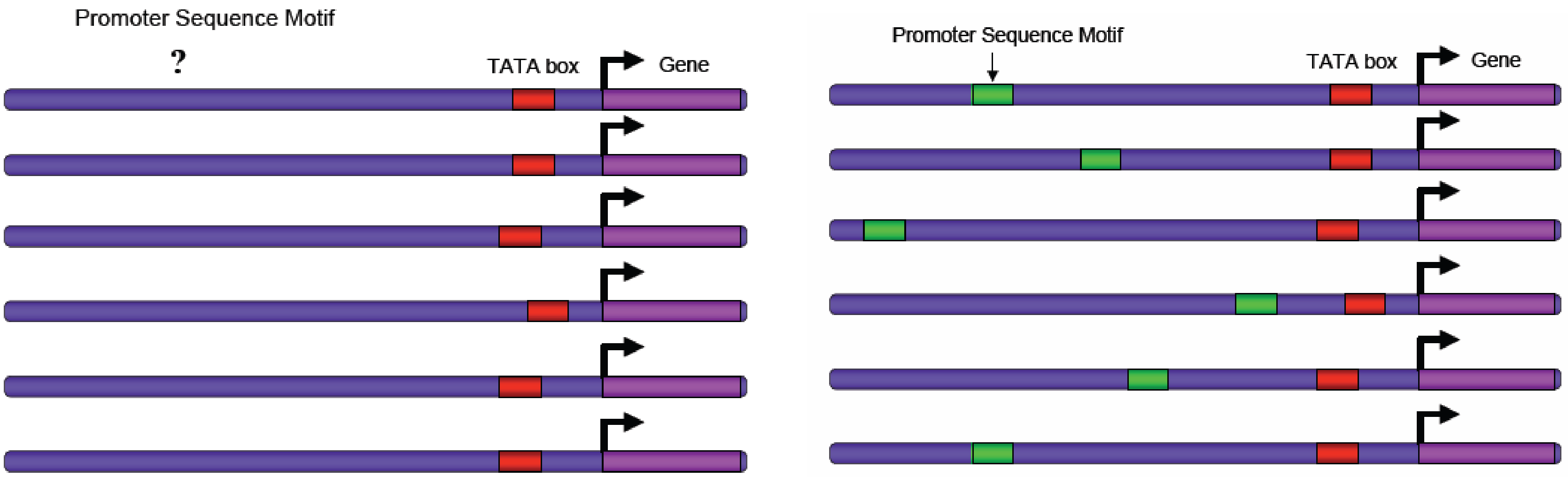
- What are motifs?
- Algorithms Used and Programs Available
- Workflow and Strategies
- MEME/MAST Demo (online and command line)

Protein Motifs

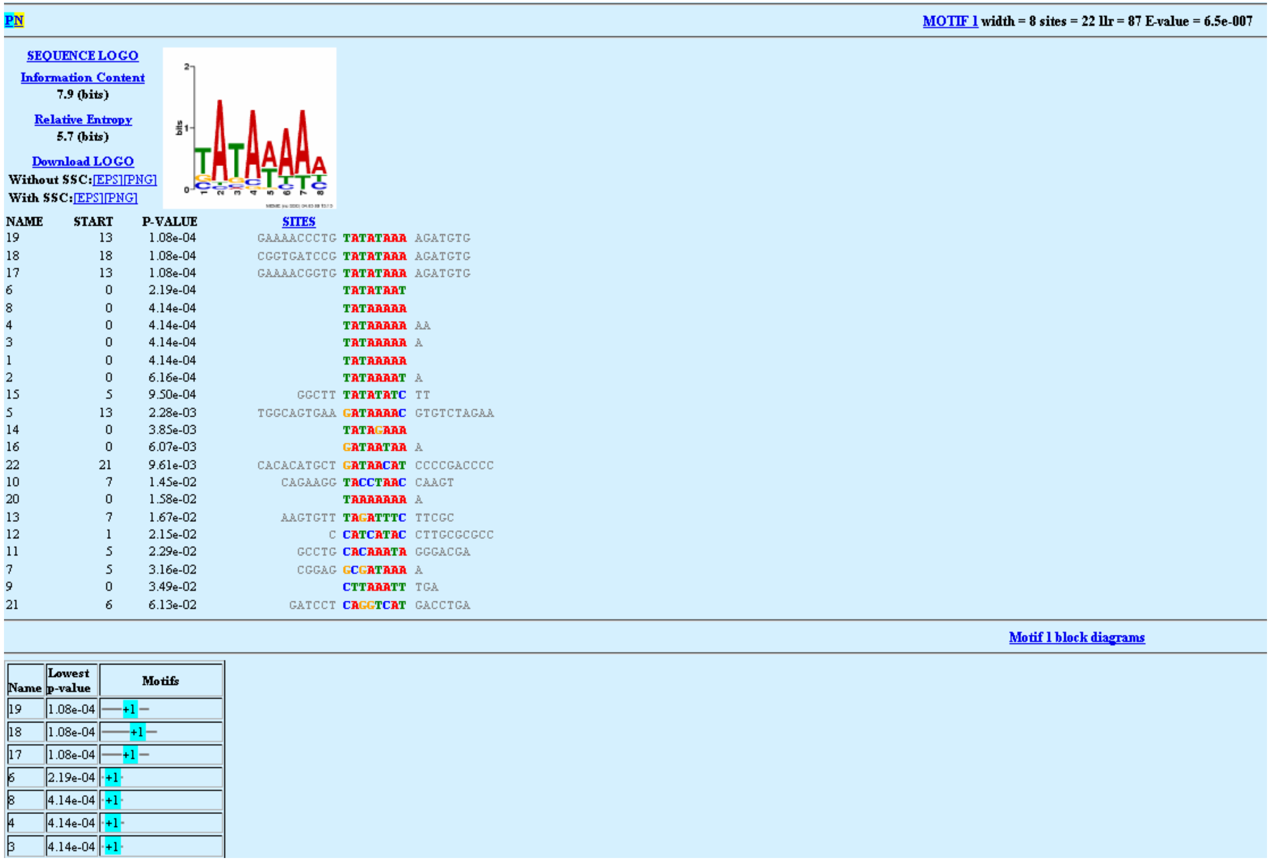




DNA Motifs


Promoter Sequence Motif
?
TATA box
Gene
Promoter Sequence Motif
TATA box
Gene

MEME Output



PN

MOTIF 1 width = 8 sites = 22 llr = 87 E-value = 6.5e-007

SEQUENCE LOGO
Information Content
7.9 (bits)
Relative Entropy
5.7 (bits)
Download LOGO
Without SSC:[EPS][PNG]
With SSC:[EPS][PNG]

```
NAME    START  P-VALUE            SITES
19         13  1.08e-04  GAAAACCCTG TATATAAA AGATGTG
18         18  1.08e-04  CGGTGATCCG TATATAAA AGATGTG
17         13  1.08e-04  GAAAACGGTG TATATAAA AGATGTG
6           0  2.19e-04             TATATAAT
8           0  4.14e-04             TATAAAAA
4           0  4.14e-04             TATAAAAA AA
3           0  4.14e-04             TATAAAAA A
1           0  4.14e-04             TATAAAAA
2           0  6.16e-04             TATAAAAT A
15          5  9.50e-04       GGCTT TATATATC TT
5          13  2.28e-03  TGGCAGTGAA GATAAAAC GTGTCTAGAA
14          0  3.85e-03             TATAGAAA
16          0  6.07e-03             GATAATAA A
22         21  9.61e-03  CACACATGCT GATAACAT CCCCGACCCC
10          7  1.45e-02     CAGAAGG TACCTAAC CAAGT
20          0  1.58e-02             TAAAAAAA A
13          7  1.67e-02     AAGTGTT TAGATTTC TTCGC
12          1  2.15e-02           C CATCATAC CTTGCGCGCC
11          5  2.29e-02       GCCTG CACAAATA GGGACGA
7           5  3.16e-02       CGGAG GCGATAAA A
9           0  3.49e-02             CTTAAATT TGA
21          6  6.13e-02      GATCCT CAGGTCAT GACCTGA
```

Motif 1 block diagrams

Name	Lowest p-value	Motifs
19	1.08e-04	+1
18	1.08e-04	+1
17	1.08e-04	+1
6	2.19e-04	+1
8	4.14e-04	+1
4	4.14e-04	+1
3	4.14e-04	+1

Definitions

- Motif: Conserved regions of protein or DNA sequences
- Pattern: Qualitative description of a motif

  eg. regular expression C[AT]AAT[CG]X
- Profile: Quantitative description of a motif

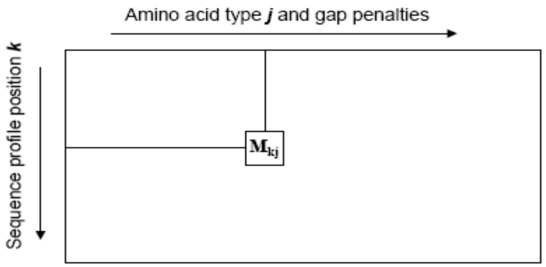
  eg. position weight matrix

Patterns

- Regular Expression Symbols
  - [ ] – OR eg. [GA] means G or A
  - { } – NOT eg. {P,V} means not P or V
  - ( ) – repeats eg. A(3) means AAA
  - X or N or “.” – any
- Complex patterns representation difficult
- Loose frequency information

  eg. [AT] *vs* 20%A 80%T

Profiles


Amino acid type *j* and gap penalties
Sequence profile position *k*
M_{kj}

$$M_{kj} = \log\left(\frac{p_{kj}}{p_j}\right)$$

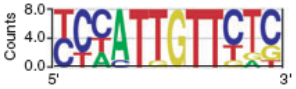
- p_{kj} = probability of amino acid j at position k in the profile
- p_j = "background" probability of amino acid j in sequence

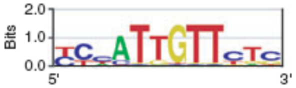
A	.02	.95	.01	.59	.51	.40	.44	.39
C	.09	.02	.03	.13	.20	.11	.05	.10
G	.10	.01	.00	.15	.13	.12	.08	.09
T	.79	.03	.96	.13	.16	.37	.43	.42
motif	T	A	T	A	A	[AT]	[AT]	[AT]

Sequence Logos

HEM13	CCCATTGTTCTC
HEM13	TTTCTGGTTCTC
HEM13	TCAATTGTTTAG
ANB1	CTCATTGTTGTC
ANB1	TCCATTGTTCTC
ANB1	CCTATTGTTCTC
ANB1	TCCATTGTTCGT
ROX1	CCAATTGTTTTG
	YCHATTGTTCTC

A	002700000010
C	464100000505
G	000001800112
T	422087088261

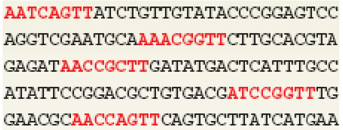




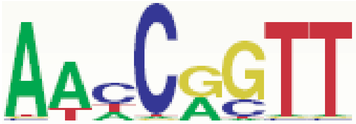
Algorithms

- Enumeration
- Probabilistic Optimization
- Deterministic Optimization

1. Identify motifs
2. Build a consensus



```
AATCAGTTATCTGTTGTATACCCGGAGTCC
AGGTCGAATGCAAAACGGTTCTTGCACGTA
GAGATAACCGCTTGATATGACTCATTTGCC
ATATTCCGGACGCTGTGACGATCCGGTTTG
GAACGCAACCAGTTCAGTGCTTATCATGAA
```



Enumeration

- Exhaustive search: word counting method, count all n-mers and look for overrepresentation
- Less likely to get stuck in a *local* optimum
- Computationally expensive
    - ➢YMF http://wingless.cs.washington.edu/YMF/YMFWeb/YMFInput.pl
    - ➢Weeder http://159.149.109.9/weederaddons/locator.html

Probabilistic Optimization

- Uses a Gibbs sampling approach
- One *n*-mer from each sequence is randomly picked to determine initial model. In subsequent iterations, one sequence, *i*, is removed and the model is recalculated. Pick a new location of motif in sequence *i* iterate until convergence
- Assumes most sequences will have the motif

➢ AlignAce
http://atlas.med.harvard.edu/cgi-bin/alignace.pl

➢ Gibbs Motif Sampler

http://bayesweb.wadsworth.org/gibbs/gibbs.html

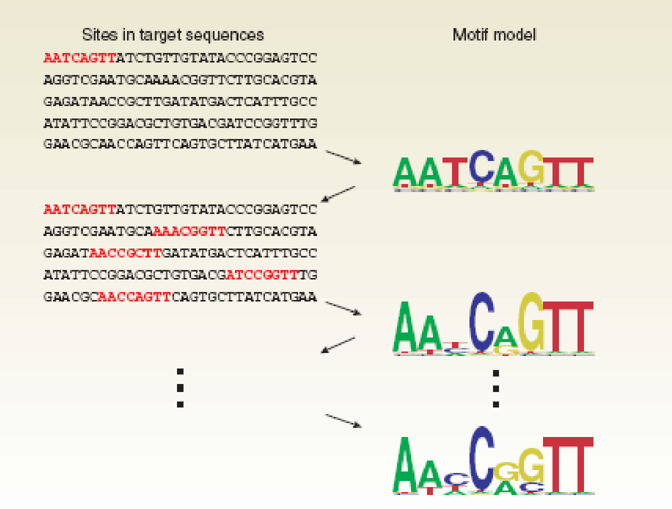
Deterministic Optimization

- Based on expectation maximization (EM)
- EM: iteratively estimates the likelihood given the data that is present
    - I. **Expectation step:** Use current parameters (and observations) to reconstruct hidden structure
    - II. **Maximization step:** Use that hidden structure (and observations) to re-estimate parameters

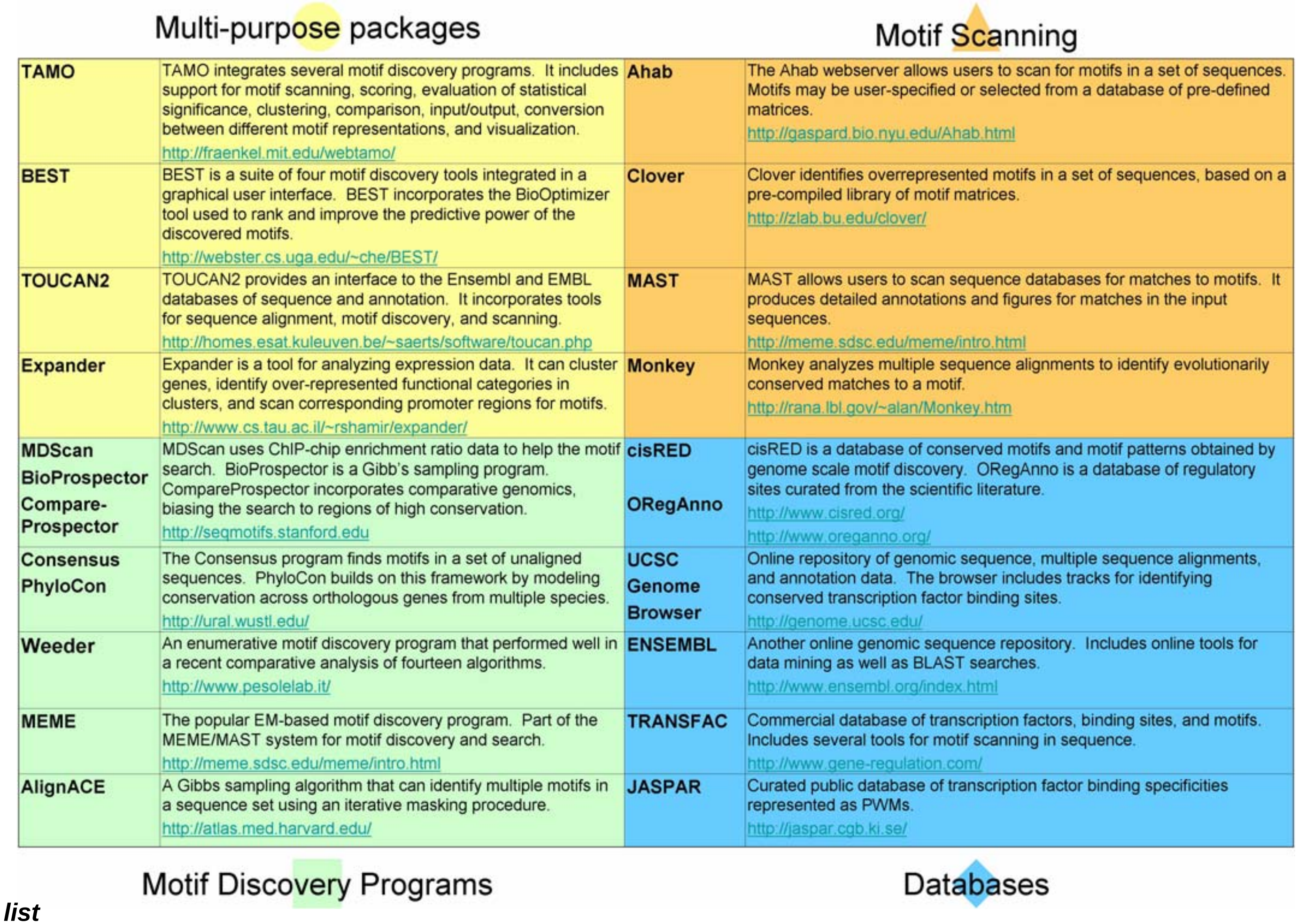
➢ MEME
http://meme.sdsc.edu

Multiple EM for Motif Elicitation MEME

- Starting from a single site, EM alternates between assigning sites and updating motif model
- Performs a single iteration for each *n*-mer in target sequences, selects the best motif from this site and then iterates only that one to convergence
- Search space increases significantly with increasing number of sequence and/or sequence lengths


Sites in target sequences
Motif model
AATCAGTTATCTGTTGTATACCCGGAGTCC
AGGTCGAATGCAAAACGGTTCTTGCACGTA
GAGATAACCGCTTGATATGACTCATTTGCC
ATATTCCGGACGCTGTGACGATCCGGTTTG
GAACGCAACCAGTTCAGTGCTTATCATGAA
AATCAGTT
AATCAGTTATCTGTTGTATACCCGGAGTCC
AGGTCGAATGCAAAACGGTTCTTGCACGTA
GAGATAACCGCTTGATATGACTCATTTGCC
ATATTCCGGACGCTGTGACGATCCGGTTTG
GAACGCAACCAGTTCAGTGCTTATCATGAA

Programs Available*



Multi-purpose packages

Motif Scanning

TAMO	TAMO integrates several motif discovery programs. It includes support for motif scanning, scoring, evaluation of statistical significance, clustering, comparison, input/output, conversion between different motif representations, and visualization. http://fraenkel.mit.edu/webtamo/	Ahab	The Ahab webserver allows users to scan for motifs in a set of sequences. Motifs may be user-specified or selected from a database of pre-defined matrices. http://gaspard.bio.nyu.edu/Ahab.html
BEST	BEST is a suite of four motif discovery tools integrated in a graphical user interface. BEST incorporates the BioOptimizer tool used to rank and improve the predictive power of the discovered motifs. http://webster.cs.uga.edu/~che/BEST/	Clover	Clover identifies overrepresented motifs in a set of sequences, based on a pre-compiled library of motif matrices. http://zlab.bu.edu/clover/
TOUCAN2	TOUCAN2 provides an interface to the Ensembl and EMBL databases of sequence and annotation. It incorporates tools for sequence alignment, motif discovery, and scanning. http://homes.esat.kuleuven.be/~saerts/software/toucan.php	MAST	MAST allows users to scan sequence databases for matches to motifs. It produces detailed annotations and figures for matches in the input sequences. http://meme.sdsc.edu/meme/intro.html
Expander	Expander is a tool for analyzing expression data. It can cluster genes, identify over-represented functional categories in clusters, and scan corresponding promoter regions for motifs. http://www.cs.tau.ac.il/~rshamir/expander/	Monkey	Monkey analyzes multiple sequence alignments to identify evolutionarily conserved matches to a motif. http://rana.lbl.gov/~alan/Monkey.htm
MDScan BioProspector Compare-Prospector	MDScan uses ChIP-chip enrichment ratio data to help the motif search. BioProspector is a Gibb's sampling program. CompareProspector incorporates comparative genomics, biasing the search to regions of high conservation. http://seqmotifs.stanford.edu	cisRED ORegAnno	cisRED is a database of conserved motifs and motif patterns obtained by genome scale motif discovery. ORegAnno is a database of regulatory sites curated from the scientific literature. http://www.cisred.org/ http://www.oreganno.org/
Consensus PhyloCon	The Consensus program finds motifs in a set of unaligned sequences. PhyloCon builds on this framework by modeling conservation across orthologous genes from multiple species. http://ural.wustl.edu/	UCSC Genome Browser	Online repository of genomic sequence, multiple sequence alignments, and annotation data. The browser includes tracks for identifying conserved transcription factor binding sites. http://genome.ucsc.edu/
Weeder	An enumerative motif discovery program that performed well in a recent comparative analysis of fourteen algorithms. http://www.pesolelab.it/	ENSEMBL	Another online genomic sequence repository. Includes online tools for data mining as well as BLAST searches. http://www.ensembl.org/index.html
MEME	The popular EM-based motif discovery program. Part of the MEME/MAST system for motif discovery and search. http://meme.sdsc.edu/meme/intro.html	TRANSFAC	Commercial database of transcription factors, binding sites, and motifs. Includes several tools for motif scanning in sequence. http://www.gene-regulation.com/
AlignACE	A Gibbs sampling algorithm that can identify multiple motifs in a sequence set using an iterative masking procedure. http://atlas.med.harvard.edu/	JASPAR	Curated public database of transcription factor binding specificities represented as PWMs. http://jaspar.cgb.ki.se/

Motif Discovery Programs

Databases

**incomplete list*

Fraenkel, E., et al. ***Practical Strategies for Discovering Regulatory DNA Sequence Motifs*** **PLoS Computational Biology 2:201-210 (2006)**

Programs Available: EMBOSS
Motif Searching

http://iona.wi.mit.edu/bio/tools/emboss/

- wordcount: Counts words of a specified size in a DNA sequence
- prophecy: Creates matrices/profiles from multiple alignments
- profit: Scan a sequence or database with a matrix or profile

Programs Available: EMBOSS

Pattern Searching

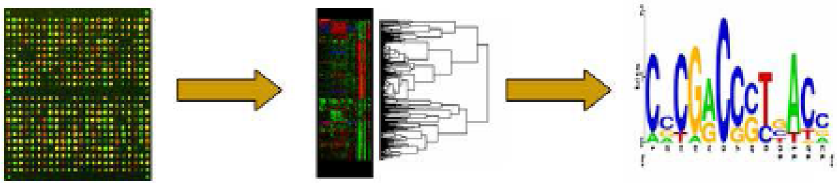
http://iona.wi.mit.edu/bio/tools/emboss/

- fuzznuc: Nucleic acid pattern search
- fuzzpro: Protein pattern search

Programs Available: Other

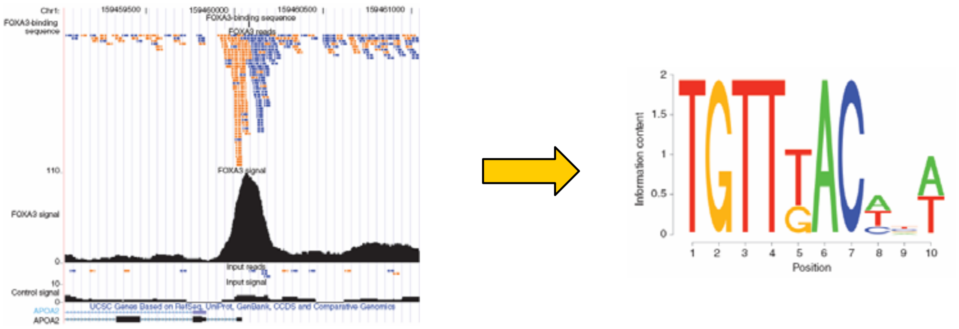
- Allegro (Expression)

http://acgt.cs.tau.ac.il/allegro/

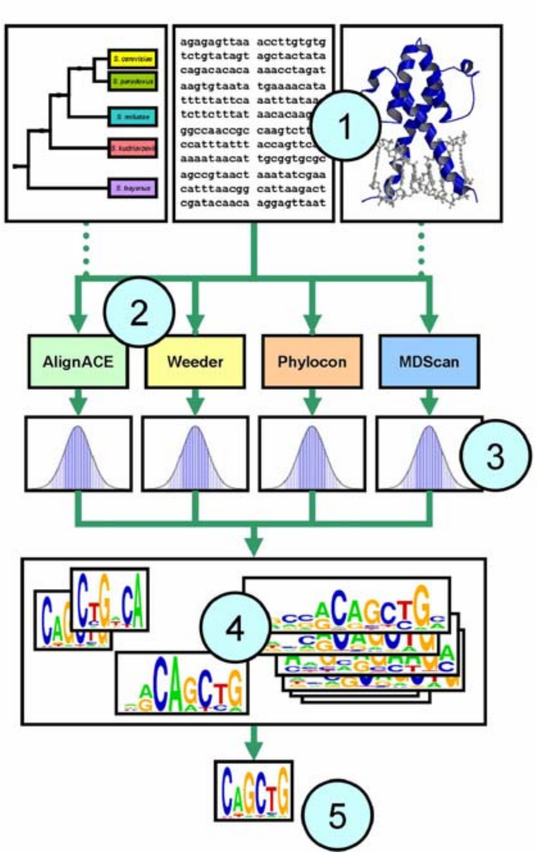


- CisGenome (ChIP-Seq)

http://www.biostat.jhsph.edu/~hji/cisgenome



Workflow and Strategies



1 **Assemble input data.** Results may be improved by restricting the input to high-confidence sequences. Some algorithms achieve improved performance by using phylogenetic conservation information from orthologous sequences or information about protein DNA-binding domains.

2 **Choose several motif discovery programs for the analysis.** For recommended programs see Figure 3.

3 **Test the statistical significance of the resulting motifs.** Use control calculations to estimate the empirical distribution of scores produced by each program on random data.

4 **Clustering and post-processing the motifs.** Motif discovery analyses often produce many similar motifs, which may be combined using clustering. Phylogenetic conservation information may be used to filter out statistically significant, but non-conserved motifs that are more likely to correspond to spurious sequence patterns.

5 **Interpretation of motifs.** Algorithms exist for linking motifs to transcription factors and for combining motif discovery with expression data.

Fraenkel, E., et al. ***Practical Strategies for Discovering Regulatory DNA Sequence Motifs*** **PLoS Computational Biology 2:201-210 (2006)**

Further Reading

- **Practical Strategies for Discovering Regulatory DNA Sequence Motifs**

http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.0020036

- **How does DNA sequence motif discovery work?**

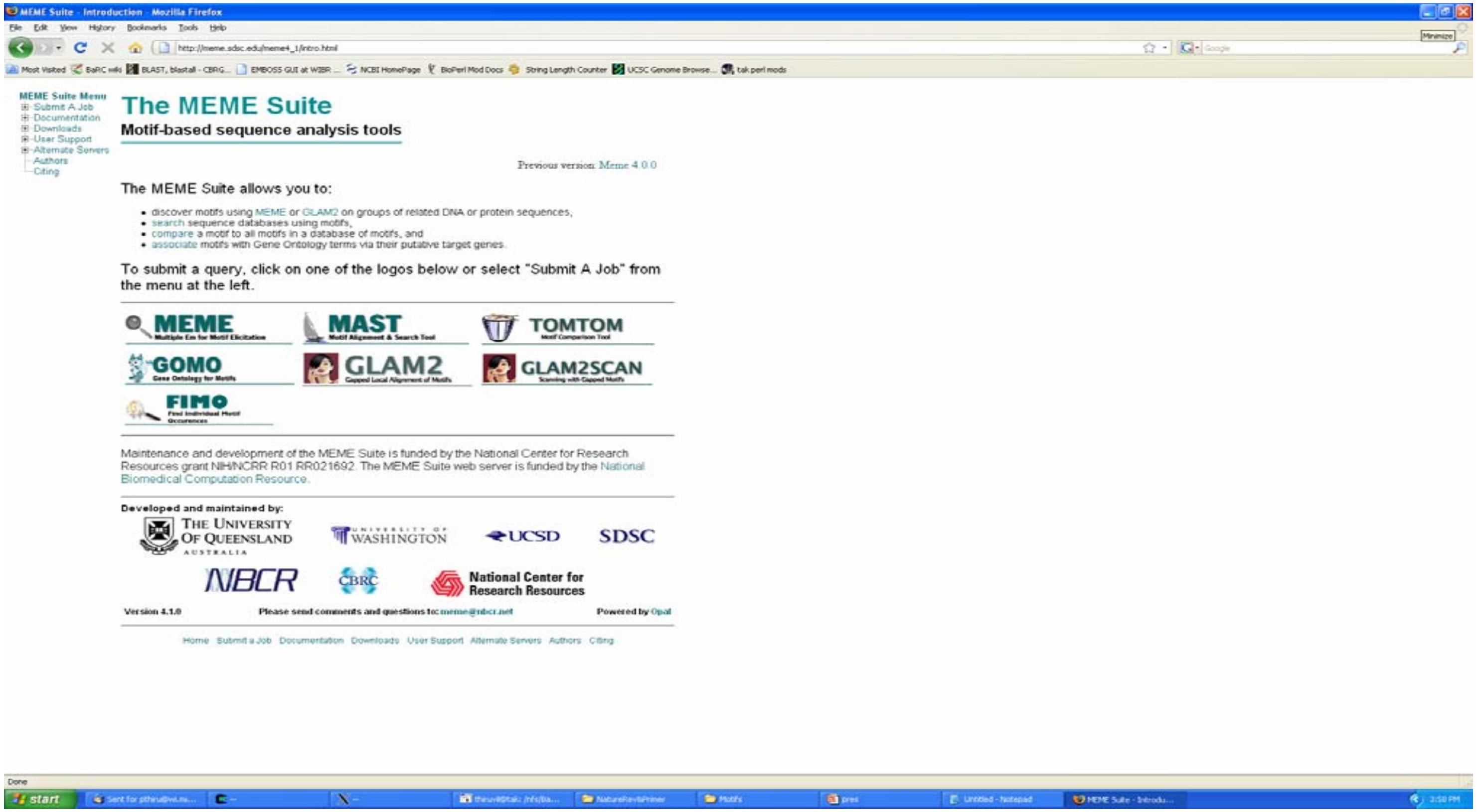
http://www.nature.com/nbt/journal/v24/n8/full/nbt0806-959.html

- **MEME**

Bailey, T. L. et al. Nucl. Acids Res. 2006 34:W369-W373; doi:10.1093/nar/gkl198

MEME/MAST Demo

http://meme.sdsc.edu


MEME Suite - Introduction - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://meme.sdsc.edu/meme4_1/intro.html
Most Visited BaRC wiki BLAST, blastall - CBRG... EMBOSS GUI at WIBR ... NCBI HomePage BioPerl Mod Docs String Length Counter UCSC Genome Browse... tak perl mods
MEME Suite Menu
Submit A Job
Documentation
Downloads
User Support
Alternate Servers
Authors
Citing
Done
start

The MEME Suite

Motif-based sequence analysis tools

Previous version: Meme 4.0.0

The MEME Suite allows you to:

- discover motifs using MEME or GLAM2 on groups of related DNA or protein sequences,
- search sequence databases using motifs,
- compare a motif to all motifs in a database of motifs, and
- associate motifs with Gene Ontology terms via their putative target genes.

To submit a query, click on one of the logos below or select "Submit A Job" from the menu at the left.

MEME
Multiple Em for Motif Elicitation
MAST
Motif Alignment & Search Tool
TOMTOM
Motif Comparison Tool
GOMO
Gene Ontology for Motifs
GLAM2
Gapped Local Alignment of Motifs
GLAM2SCAN
Scanning with Gapped Motifs
FIMO
Find Individual Motif Occurences

Maintenance and development of the MEME Suite is funded by the National Center for Research Resources grant NIH/NCRR R01 RR021692. The MEME Suite web server is funded by the National Biomedical Computation Resource.

Developed and maintained by:

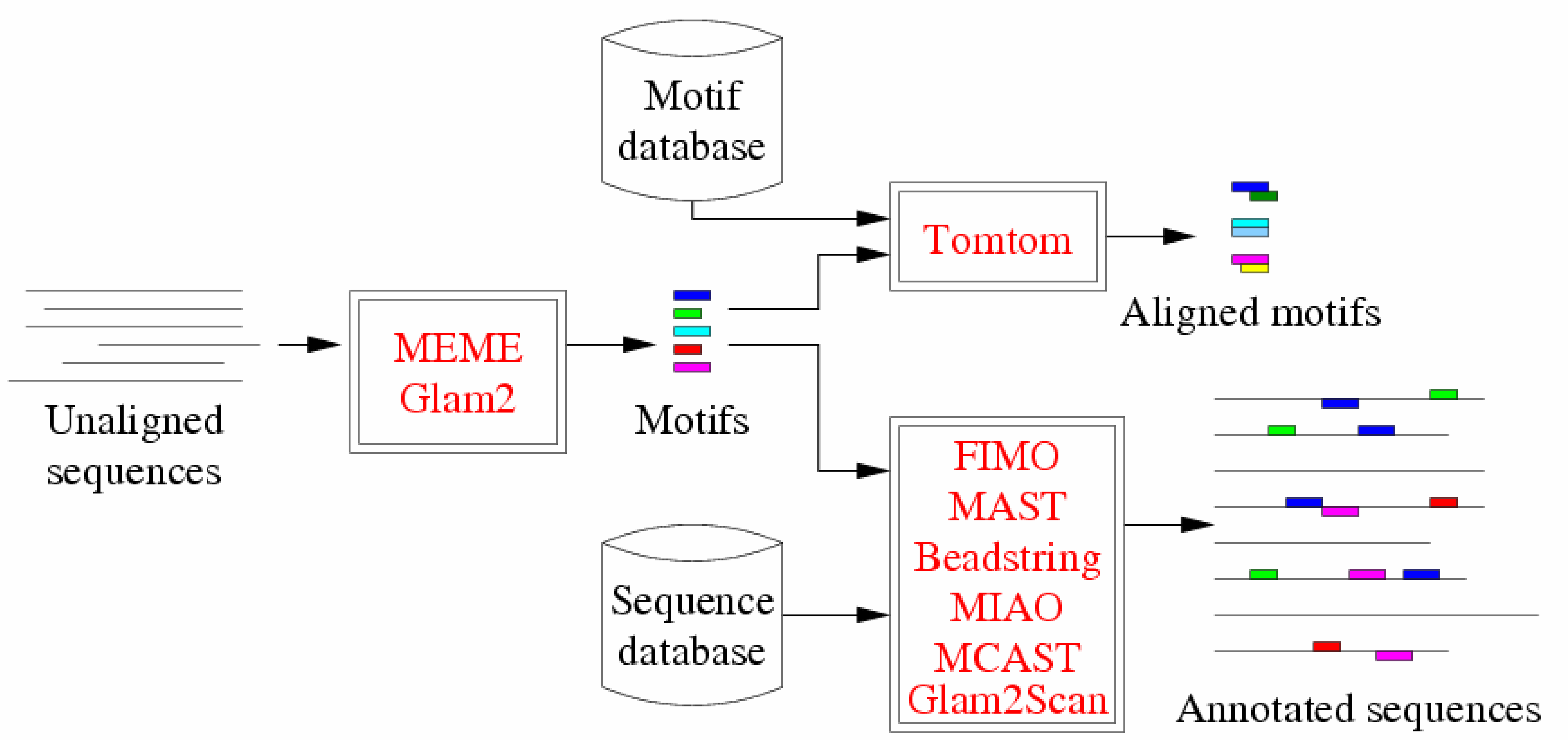
The University of Queensland Australia
University of Washington
UCSD
SDSC
NBCR
CBRC
National Center for Research Resources

Version 4.1.0 Please send comments and questions to: meme@nbcr.net Powered by Opal

Home Submit a Job Documentation Downloads User Support Alternate Servers Authors Citing

MEME/MAST Demo

MEME Suite



Motif database

Tomtom

Aligned motifs

Unaligned sequences

MEME
Glam2

Motifs

FIMO
MAST
Beadstring
MIAO
MCAST
Glam2Scan

Sequence database

Annotated sequences

MEME/MAST Demo

Command Line (on tak)

Usage: MEME (Find ungapped motifs in unaligned sequences)

eg. `meme sample.fa -dna -maxw 10 -nmotifs 5 -mod zoops -pal -maxsize 1000000 -o sample_meme`

meme <dataset> [optional arguments]

- <dataset> file containing sequences in FASTA format
- [-text] output in text format (default is HTML)
- [-dna] sequences use DNA alphabet
- [-protein] sequences use protein alphabet
- [-mod oops|zoops|anr] distribution of motifs
- [-nmotifs <nmotifs>] maximum number of motifs to find
- [-evt <ev>] stop if motif E-value greater than <evt>
- [-minw <minw>] minimum motif width
- [-maxw <maxw>] maximum motif width

For complete list of options enter “meme” at the command prompt

MEME/MAST Demo

Command Line (on tak)

Usage: MAST (Searches a sequence database for occurrences of known motifs)

eg. mast motifs.txt -d data.fa

mast <mfile> <database> [optional arguments ...]

- **<mfile>   file containing motifs to use; may be a MEME output file or similar file**
- **[-d <database> | -stdin] search sequences in <database> with motifs**

For complete list of options enter “mast” at the command prompt