MINING DNA NETWORK EXPRESSIONS WITH EVOLUTIONARY MULTI-OBJECTIVE PROGRAMMING

uOttawa
L'Université canadienne
Canada's university

Manuel Belmadani, M. Computer Science (specialization in Bioinformatics) Candidate
mbelm006@uottawa.ca

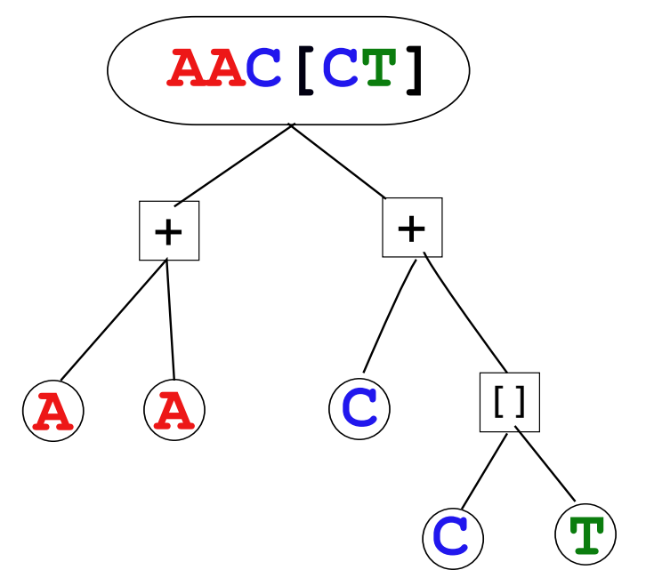
Abstract

High-throughput sequencing experiments, such as ***ChIP-seq***, identify protein interactions within large samples of DNA. Such experiments may identify and output thousands of peak sequences where an interaction of interest is regulated by that region of DNA. Computational biologists often require a **de novo** analysis of these peaks to mine short overrepresented patterns, frequently found amongst the sequences. This task is referred to as **motif discovery**. Existing tools struggle at motif discovery if the number of input sequences increases. The time and complexity costs of calculating pattern occurrence often requires compromises on runtime, quality, or length of the output predictions.

The proposed solution **(MotifGP)** is a motif discovery tool driven by a **Strongly Typed Genetic Programming (STGP)** algorithm equipped with **multi-objective optimization**. STGP is a stochastic search process based off evolutionary algorithms which evolves candidate solutions as typed checked programs. Multi-objective optimization allows the algorithm to evolve according to multiple metrics used to build predictions that are high in both support and specificificity. The software produces a collection of non-dominated multi-objective solutions, which can then be aligned to reference DNA motif databases (such as JASPAR, TRANSFAC, and Uniprobe) in order to validate/identify predictions.

Strongly Typed Genetic Programming

Genetic Programming is a search heuristic where a population of programs created under a template grammar are evolved over generations of crossovers and mutations. MotifGP uses a grammar to produce Network Expressions that can be matched against sequences.



Strongly Typed Network Expression grammar

Root type
$< String > Individual \rightarrow < String >$
Primitive
$< String > Concatenation \rightarrow < String >< String >$
$< String > CharacterClass \rightarrow < Char >< Char >< Char >< Char >$
#Terminals
$< String > \rightarrow < Char >$
$< Char > \rightarrow A|C|G|T$

Multi-objective optimization

In population based evolutionary algorithms, new individuals are created over generations and evaluated by a ***fitness function***. This scores the individuals in how well they perform as solutions. In **multi-objective** optimization, the individuals have more than one fitness values. Fitnesses with multiple values are compared and ranked by a multi-objective optimization algorithm **[3]**.

MotifGP matches candidate solutions (network expressions) against a set of input and control sequences. We use the match count on each set to compute two metrics of *discrimination* and *coverage*:

$$Discrimination = \frac{|P|}{|P|+|C|}$$

$$Coverage = \frac{|P|}{k}$$

For a given pattern, let **P** be the set of matched input sequences, **C** be the set of matched control sequences, and **k** be the total number of input sequences.

Discrimination favors solutions that are not highly represented in the control sequences, while *Coverage* maintains solutions that are found in multiple input sequences.

Benchmarking

For each dataset, **MotifGP** was executed 10 times with a different random initialization. The **E-value** of the top DREME motif alignment is used as the benchmark value for MotifGP.

Dataset	<	DREME	10^{-1}	10^{-2}	10^{-3}	**Peaks**	**Size**	**Time (s)**
cMyc	**5**	8.59888E-005	3	2	0	3422	425K	309
CTCF	**5**	3.50248E-007	3	2	0	39609	4.8M	7691
E2f1	**10**	0.0385603	0	0	0	20699	2.6M	2769
Esrrb	**10**	0.000574497	0	0	0	21647	2.7M	4054
Klf4	**7**	3.28351E-008	3	0	0	10875	1.4M	1504
Nanog	**10**	0.000967177	0	0	0	10343	1.3M	1987
nMyc	**10**	0.00121468	0	0	0	3761	467K	362
Oct4	0	7.16982E-005	2	5	1	7182	891K	1418
Smad1	**2**	0.00107949	6	0	0	1126	16K	71
Sox2	**10**	0.00539771	0	0	0	4526	561K	427
STAT3	**9**	0.0323896	0	0	0	2546	316K	164
Tcfcp2l1	**6**	0.00307791	1	0	0	26910	3.3M	5004
Zfx	**10**	0.230923	0	0	0	10338	1.3M	1258

Over the 10 individual runs of each dataset, we counted the number of runs where a prediction outperformed the benchmark. We also looked at the runs that are within up to 3 orders of magnitude of the benchmark E-Value. This experiment shows the algorithm performed favorably in term of stability in **11 out of 13** datasets, including 6 where MotifGP suceeded in all 10 runs.

Methodology

To determine if MotifGP can outperform DREME, we compared both software against **13 datasets from mouse Embryonic Stem Cells** ChIP-Seq experiments **[1].**

The recorded runtimes for each DREME run were used as time limits for MotifGP.

Each dataset may contain one or more factor motif. Desirable predictions should report long overlap and the lowest E-values once aligned to known motifs.

Conclusions and Future Work

Combining **STGP** and **multi-objective optimization** directs the solution search towards motifs that align with high confidence and fully overlaps known motifs from databases.

Other existing ***ChIP-seq*** experiments can be revisited using MotifGP as it may be able to uncover subtle patterns that other tools could not.

Future development of MotifGP will investigate different multi-objective fitness function variants and explore other evolutionary methods.

References

[1] Chen, X. et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133, 1106–1117 (2008).
[2] Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics 27, 1653–1659 (2011).
[3] Fortin, F.-A. & Parizeau, M. Revisiting the NSGA-II crowding-distance computation. in Proceedings of the 15th annual conference on Genetic and evolutionary computation 623–630 (ACM, 2013).
[4] Fortin, F.-A., De Rainville, F.-M., Gardner, M.-A. G., Parizeau, M. & Gagné, C. DEAP: Evolutionary Algorithms Made Easy. J. Mach. Learn. Res. 13, 2171–2175 (2012)
[5] Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. Genome Biol. 8, R24 (2007)..

DISTRIBUTED
EVOLUTIONARY
ALGORITHMS IN
PYTHON

python™

Tomtom
Motif Comparison Tool

Terminology

Motif: A recurring feature in a set of elements, such as frequent patterns in DNA strings.

Transcription Factor (TF): Proteins that bind to DNA at specific sites.

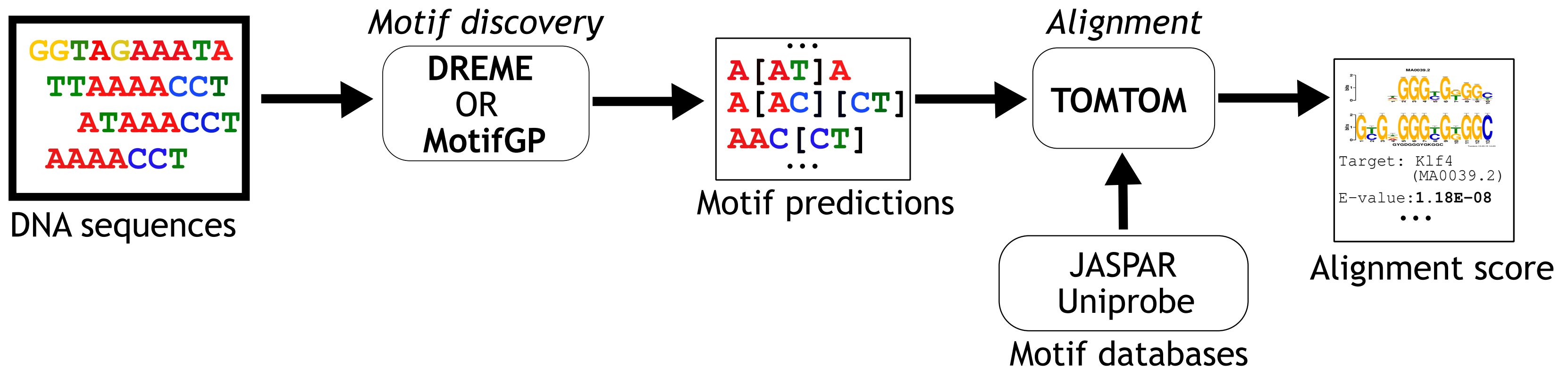
Network Expressions: Subset of the regular expression grammar. Defines patterns in which each position has a set of possible characters.

DNA Motif Discovery: Mining DNA sequences for motifs. Can be used to identify TF binding sites.

Background and Motivation

A state-of-the-art motif discovery tool, **DREME**, predicts DNA motifs by mining and refining k-mer (words) from an input set of sequences **[2]**. The predictions are compared to known motif databases with **TOMTOM [5]**. Reported alignments are listed and scored by E-value.

De novo motif discovery pipeline

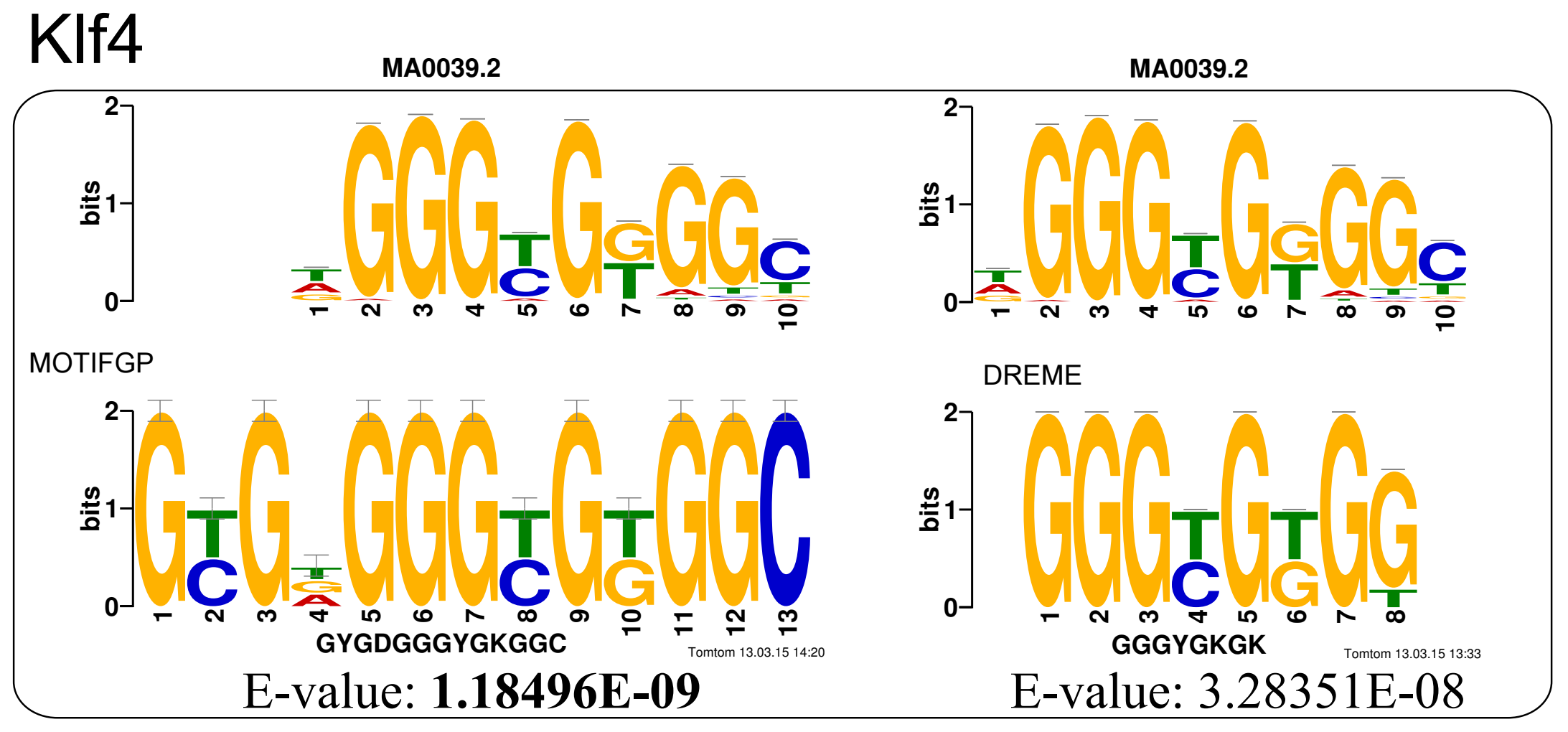


While DREME predictions have significant E-values, the pattern length is limited to 8 characters. Some known motifs are more than twice the size DREME can predict. This is a restriction set to keep low runtimes under real ChIP-seq datasets. We propose our software, **MotifGP,** as an alternative to this limitation. **Our goal is to explore the potential of multi-objective evolutionary computing in finding significant motif predictions on large datasets, under runtimes comparable to DREME.**

Experimental Results

A sample of top ranking motifs identified by MotifGP are compared to DREME's prediction on the same datasets. Each top logo is the target motif from the database, aligned by TOMTOM. The bottom logos are the predictions by MotifGP (at the left), and DREME, (at the right).

Klf4

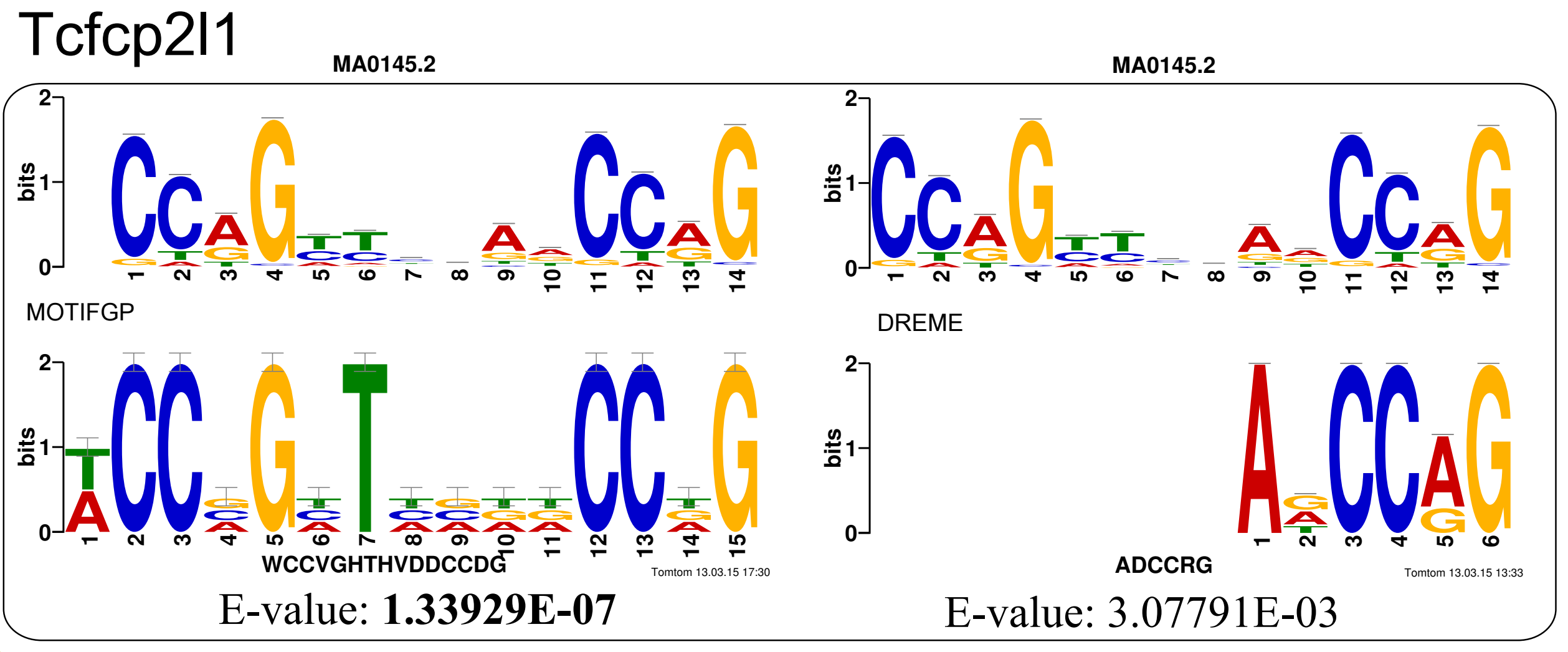


E-value: **1.18496E-09**

E-value: 3.28351E-08

As shown by the ***Klf4*** dataset prediction, MotifGP has the ability to overlap an entire target motif. It even finds additional nucleotides attached to the left end of the predicted motif, found in the input sequences.

Tcfcp2l1

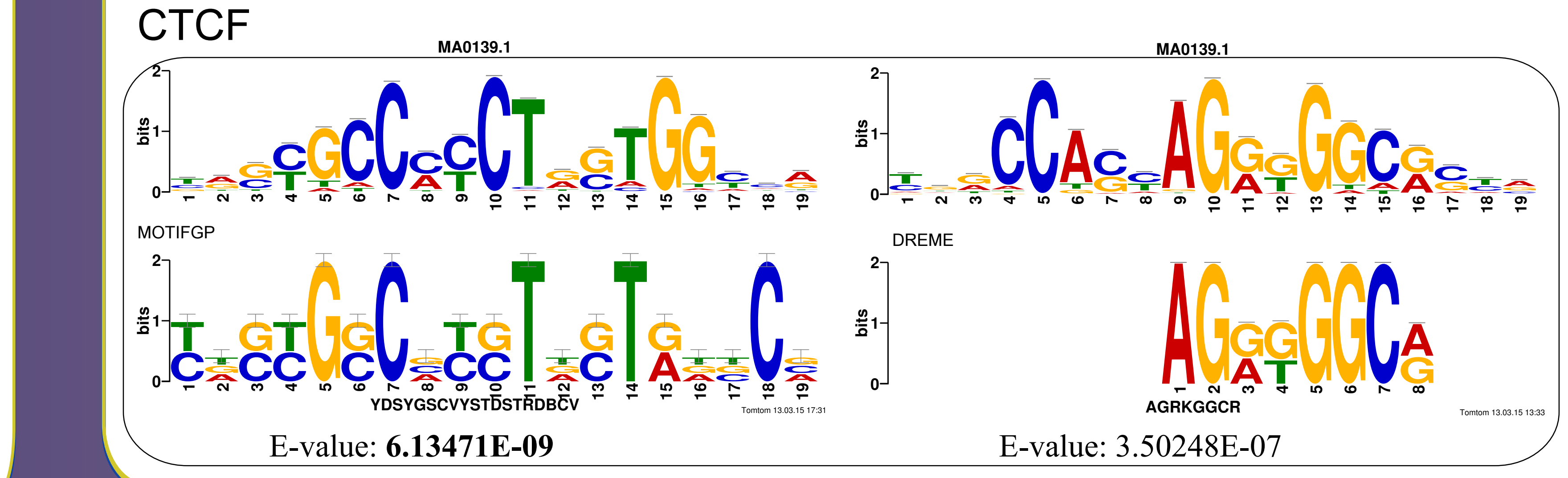


E-value: **1.33929E-07**

E-value: 3.07791E-03

The ***Tcfcp2l1*** motif has two fragments joined by a weak gap (position 7 to 8). Evolutionary computing allows such fragments of a motif to evolve independently and eventually crossover, allowing short gaps to naturally appear in produced predictions.

CTCF



E-value: **6.13471E-09**

E-value: 3.50248E-07

MotifGP's motif fully overlaps the 19 bases of ***CTCF***. Each tool's prediction targeted a different strand (orientation), showing a complementation between each methods.