CoSREM: A graph mining algorithm for the discovery of combinatorial splicing regulatory elements

Eman Badr and Lenwood Heath
Departments of Computer Science, GBCB, and PPWS
ebadr@vt.edu, heath@vt.edu

Computational Biology and Bioinformatics Group

Abstract

Motivation: Alternative splicing (AS) is a post transcriptional regulatory mechanism for gene expression regulation. Splicing decisions are affected by the combinatorial behavior of different splicing factors that bind to multiple binding sites in exons and introns, which are known as splicing regulatory elements (SREs). Here we propose CoSREM (Combinatorial SRE Miner), a graph mining algorithm to discover combinatorial SREs in human exons. Our model does not assume a fixed length of SREs and incorporates experimental evidence as well to increase the accuracy. CoSREM is able to identify sets of SREs and is not limited to SRE pairs as are current approaches.

Results: We identified 37 SRE sets that include both enhancer and silencer elements. We show that our results intersect with previous results, including some that are experimental. We also show that the SRE set GGGAGG and GAGGAC identified by CoSREM may play a role in exon skipping events in several tumor samples. We also applied CoSREM on different tissues from human to identify tissue specific regulatory elements.

Dataset

We utilize LEIsc (Log of the Enrichment Index, scaled) scores from Ke et al. (2011). We also utilized all unique coding exons for known human genes available from the ENCODE project (Karolchik et al., 2004). It includes 205,163 exons from 29,179 genes. Data was acquired from the RefSeq Genes track.The December, 2013, human genome assembly (GRCh38/hg38) was used. A third data source is the RNA-seq data set from the human body map project (Flick et al., 2013) to study the SRE effects on tissue specificity. The data includes 16 different human tissues. We focused on brain, heart, liver, and muscle tissues.

Alternative splicing

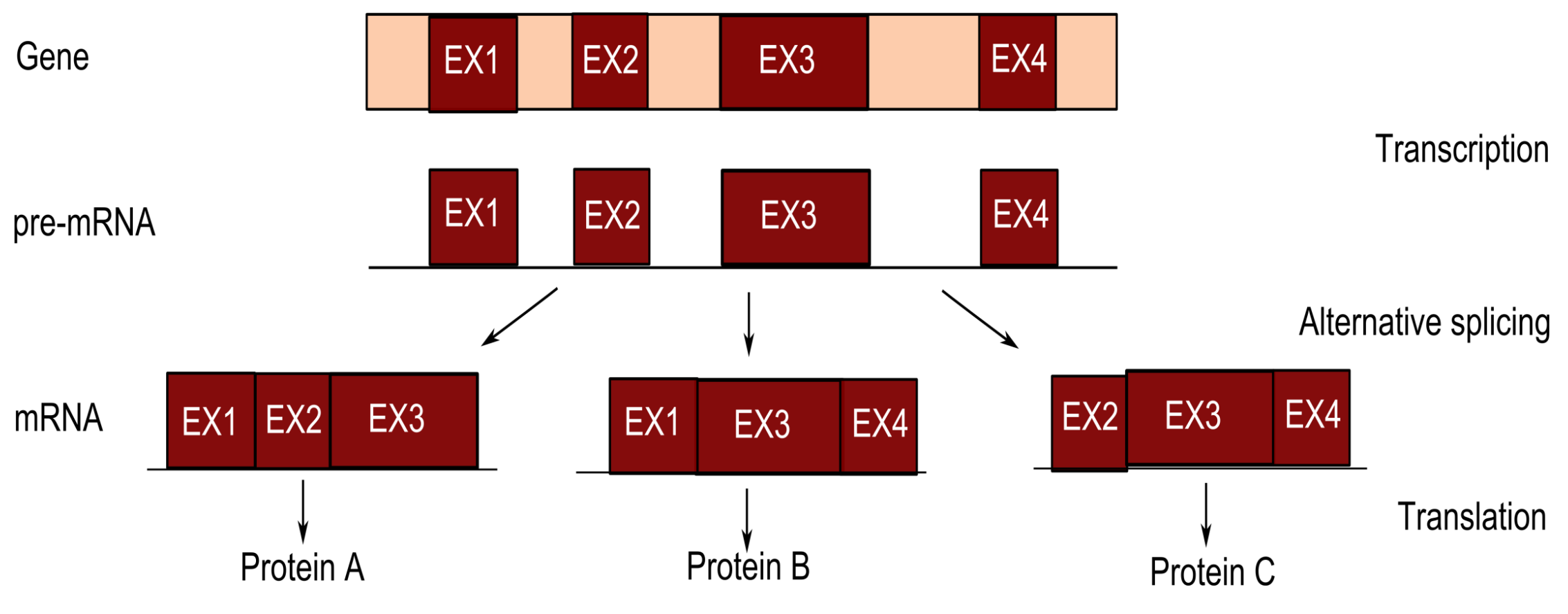


Fig 1. Alternative splicing

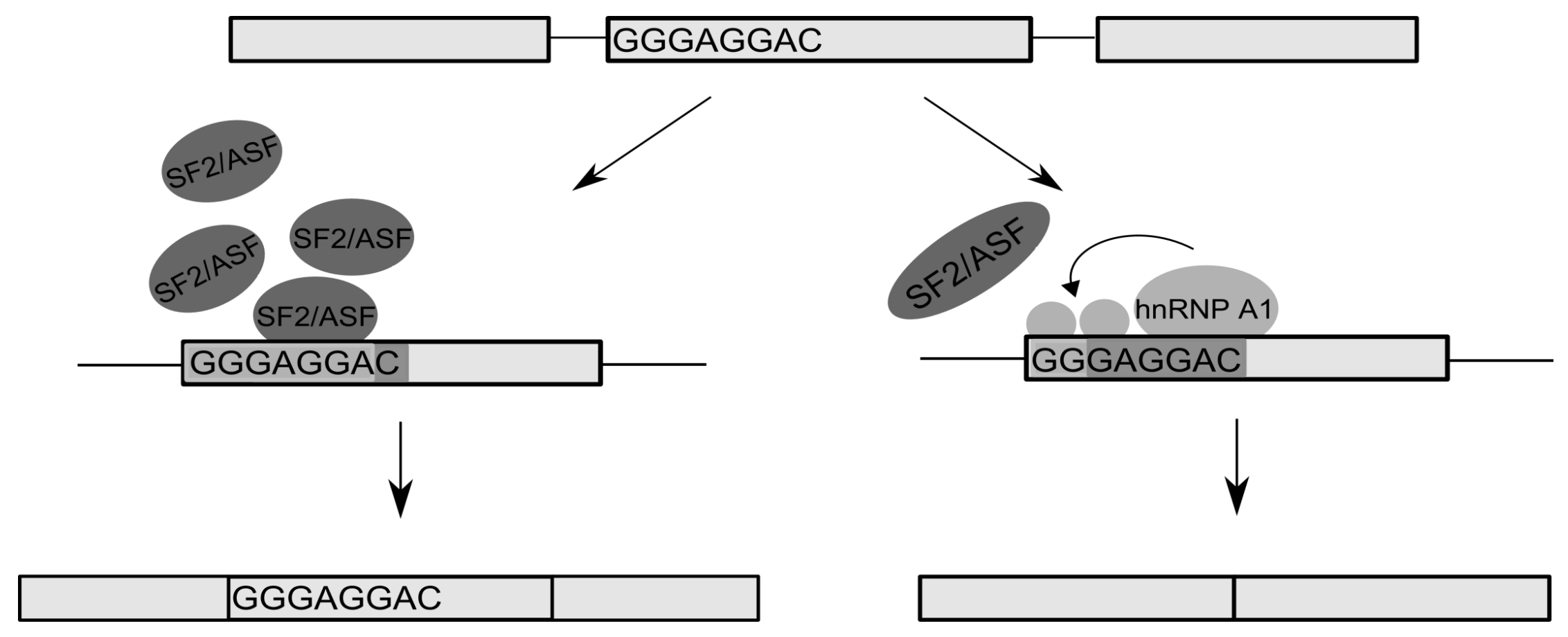
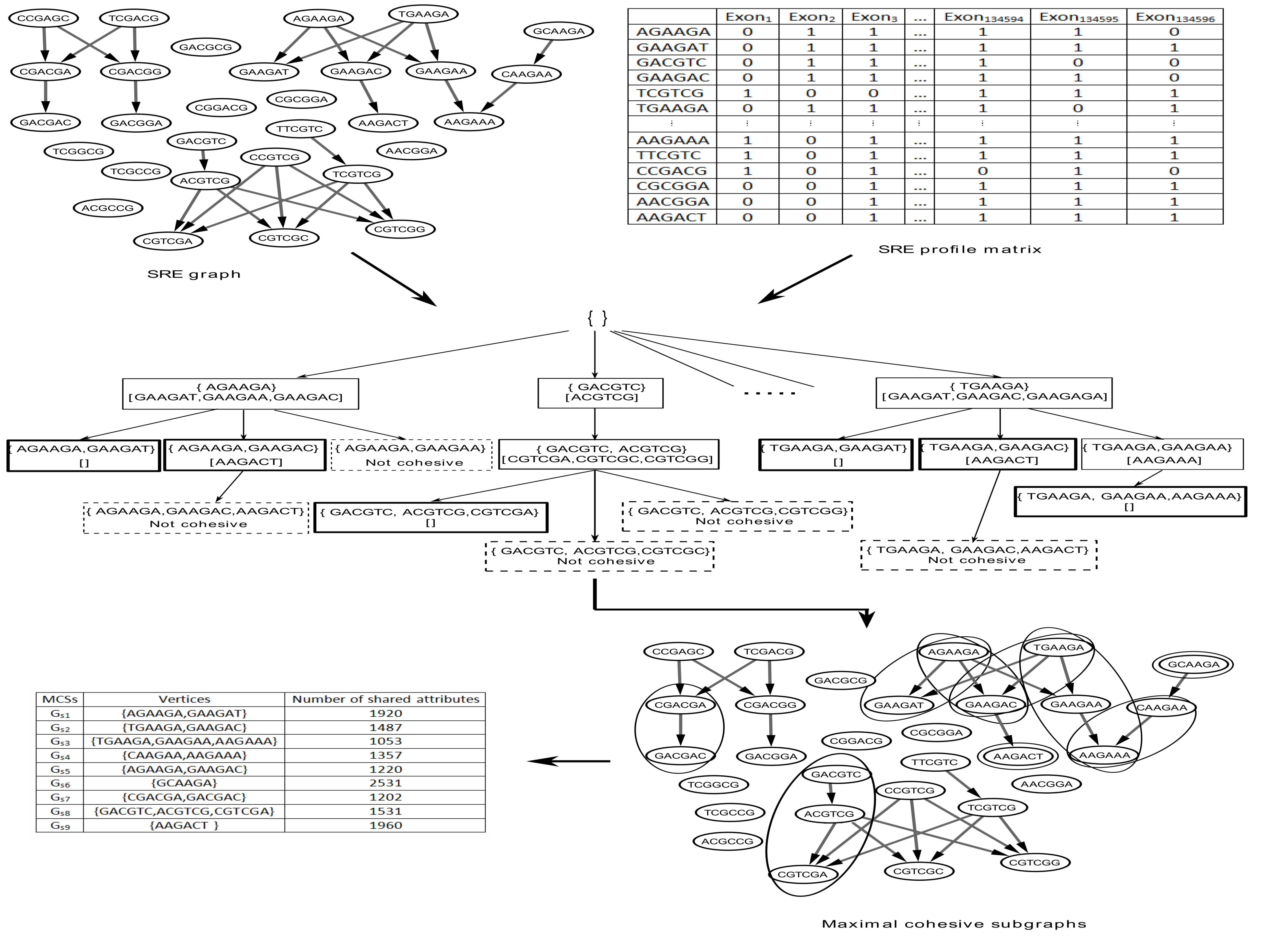


Fig 2. antagonistic behavior between anhncers and silencers

Level1: GenMCS algorithm


SRE graph
SRE profile matrix
{ }
{ AGAAGA} [GAAGAT,GAAGAA,GAAGAC]
{ GACGTC} [ACGTCG]
.....
{ TGAAGA} [GAAGAT,GAAGAC,GAAGAGA]
{ AGAAGA,GAAGAT} []
{ AGAAGA,GAAGAC} [AAGACT]
{ AGAAGA,GAAGAA} Not cohesive
{ GACGTC, ACGTCG} [CGTCGA,CGTCGC,CGTCGG]
{ TGAAGA,GAAGAT} []
{ TGAAGA,GAAGAC} [AAGACT]
{ TGAAGA,GAAGAA} [AAGAAA]
{ AGAAGA,GAAGAC,AAGACT} Not cohesive
{ GACGTC, ACGTCG,CGTCGA} []
{ GACGTC, ACGTCG,CGTCGG} Not cohesive
{ TGAAGA, GAAGAA,AAGAAA} []
{ GACGTC, ACGTCG,CGTCGC} Not cohesive
{ TGAAGA, GAAGAC,AAGACT} Not cohesive
Maximal cohesive subgraphs

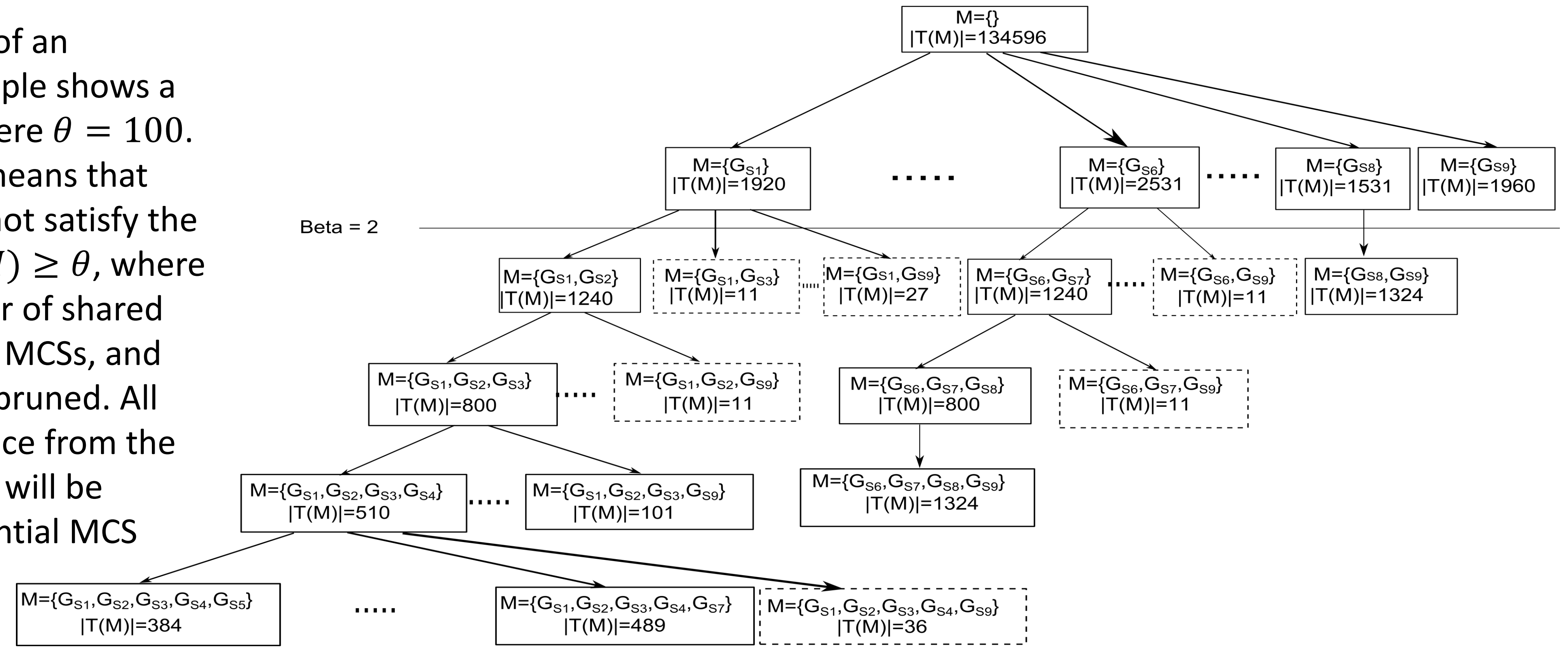
	Exon$_1$	Exon$_2$	Exon$_3$	...	Exon$_{134594}$	Exon$_{134595}$	Exon$_{134596}$
AGAAGA	0	1	1	...	1	1	0
GAAGAT	0	1	1	...	1	1	1
GACGTC	0	1	1	...	1	0	0
GAAGAC	0	1	1	...	1	1	0
TCGTCG	1	0	0	...	1	1	1
TGAAGA	0	1	1	...	1	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
AAGAAA	1	0	1	...	1	1	1
TTCGTC	1	0	1	...	1	1	1
CCGACG	1	0	1	...	0	1	0
CGCGGA	0	0	1	...	1	1	1
AACGGA	0	0	1	...	1	1	1
AAGACT	0	0	1	...	1	1	1

MCSs	Vertices	Number of shared attributes
G_{s1}	{AGAAGA,GAAGAT}	1920
G_{s2}	{TGAAGA,GAAGAC}	1487
G_{s3}	{TGAAGA,GAAGAA,AAGAAA}	1053
G_{s4}	{CAAGAA,AAGAAA}	1357
G_{s5}	{AGAAGA,GAAGAC}	1220
G_{s6}	{GCAAGA}	2531
G_{s7}	{CGACGA,GACGAC}	1202
G_{s8}	{GACGTC,ACGTCG,CGTCGA}	1531
G_{s9}	{AAGACT }	1960

Fig. 3. An example of mining cohesive subgraphs. The graph at the top left corner represents the SRE graph G$_{USE}$. We choose R = 30 which means the SRE graph contains the top 30 6-mers in rank. The matrix on the right is the SRE profile matrix P$_{ESE}$.Setting α = 1000 means that the connected vertices should co-occur in at least 1000 exons to be considered a cohesive subgraph. The tree in the middle shows how GenMCS proceeds. The bold boxes represent cohesive subgraphs. The dotted boxes represent subgraphs that are not cohesive and the remaining branch will be pruned. The output is 9 subgraphs as illustrated in the bottom graph.

Level2: BuildMCSTree algorithm

Fig. 4. An example of an *MCStree*. The example shows a part of the tree where $\theta = 100$. The dotted boxes means that this MCS set does not satisfy the user threshold $T(M) \geq \theta$, where $T(M)$ is the number of shared exons between the MCSs, and this branch will be pruned. All vertices with distance from the root $\geq \beta$ threshold will be considered as potential MCS collection.



Analysis and Results

Combinatorial SREs	Number of Exons	Putative splicing factors
CAAGGA;TGAGGA	105	SRp55
GATGCC; TGCCTA	127	SRp55
GCGGGAG;GGGAGG	169	SF2/ASF
TGAGGA;GGTGAG	199	SRp55
GAGGAC;GGGAGG	233	SF2/ASF
GAAGGC;AGGCAG	373	SF2/ASF

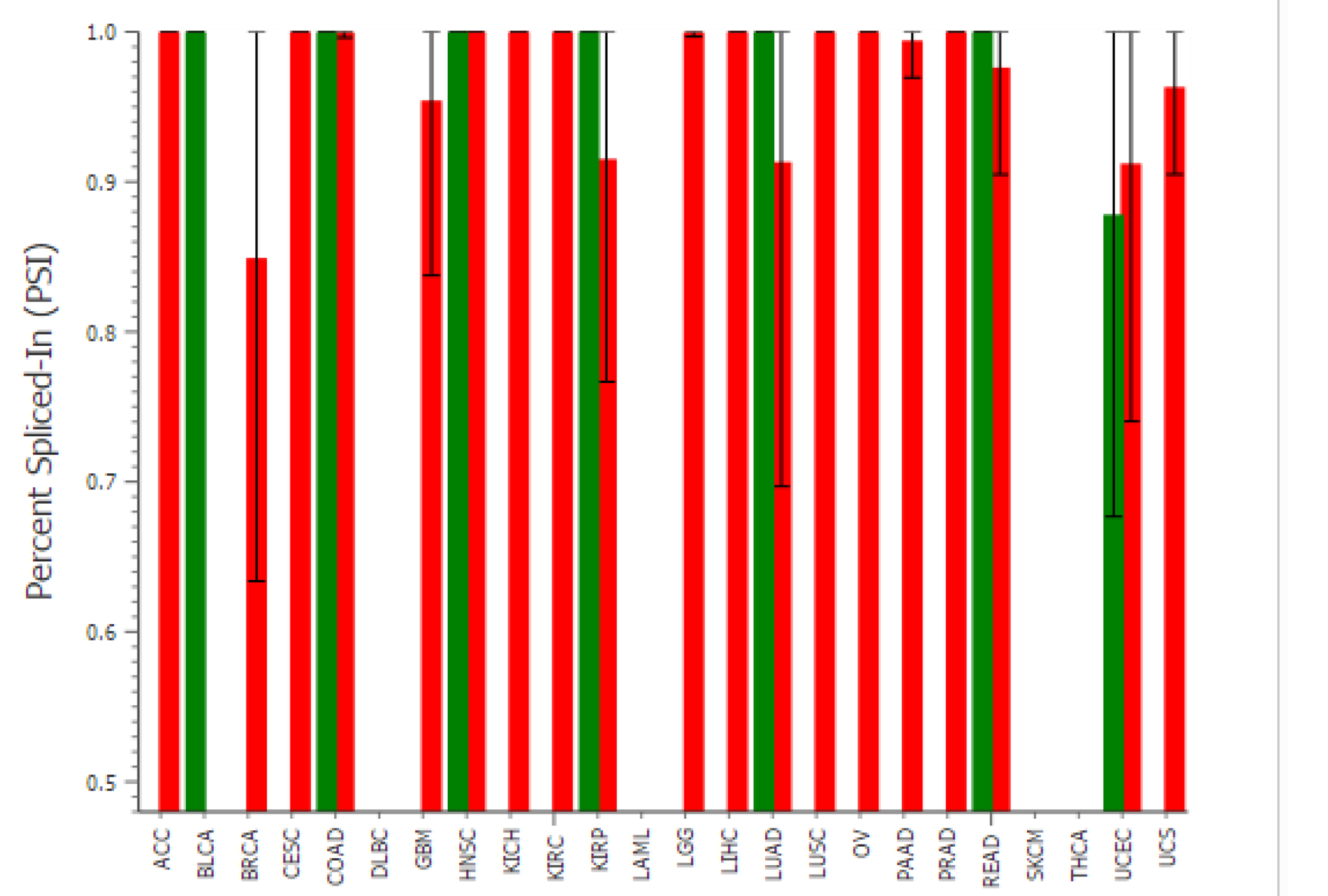


Fig5. A bar plot to illustrate the difference in the PSI (Percent SplicedIn) values between normal and tumor samples for exon 17 in PRKCG gene. The red bars represent the PSI of tumor samples while the black bars represent the normal samples. This figure is generated using TCGA Spliceseq (Ryan et al., 2012).

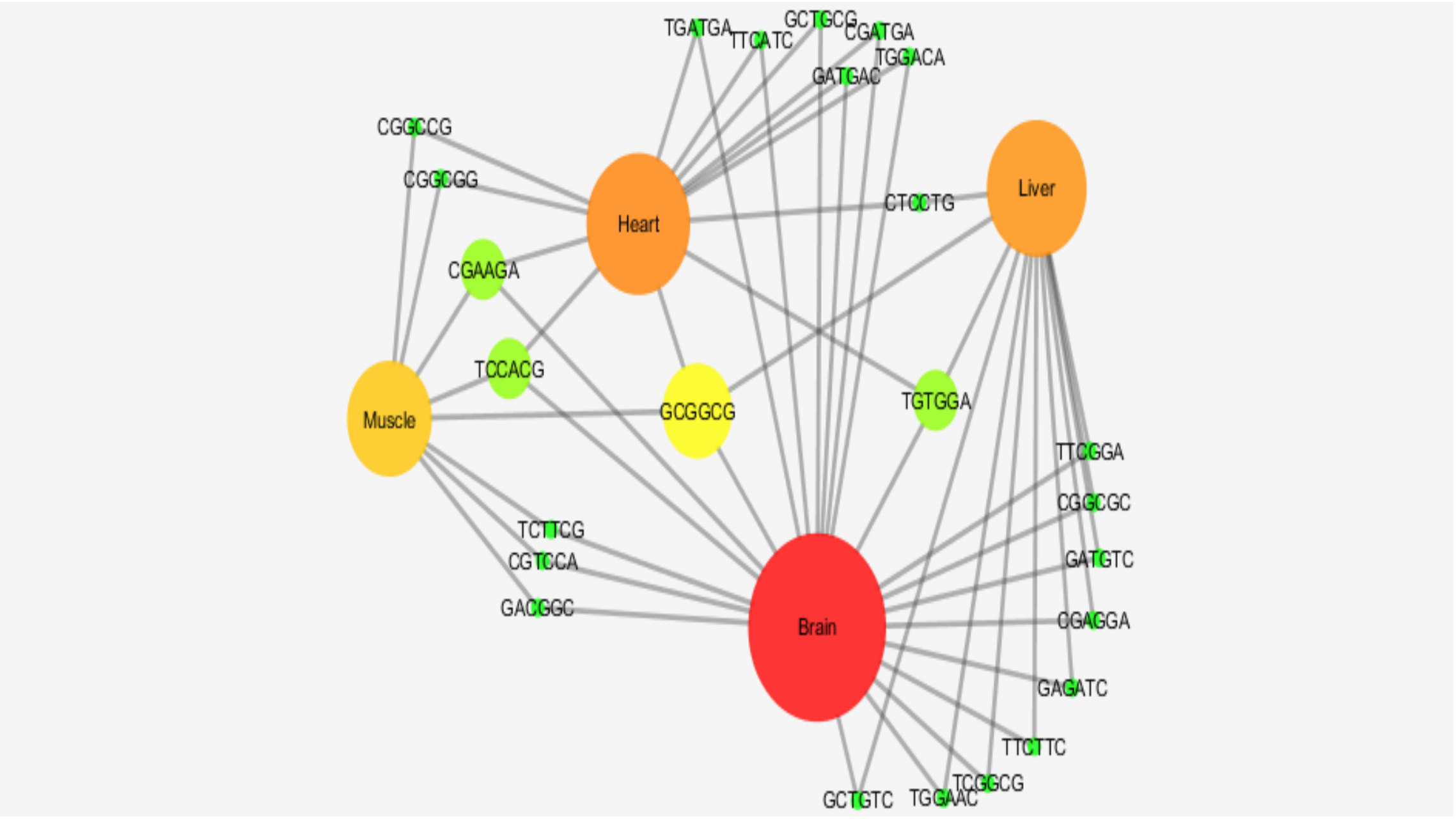


Fig 6. Regulatory network for different tissues

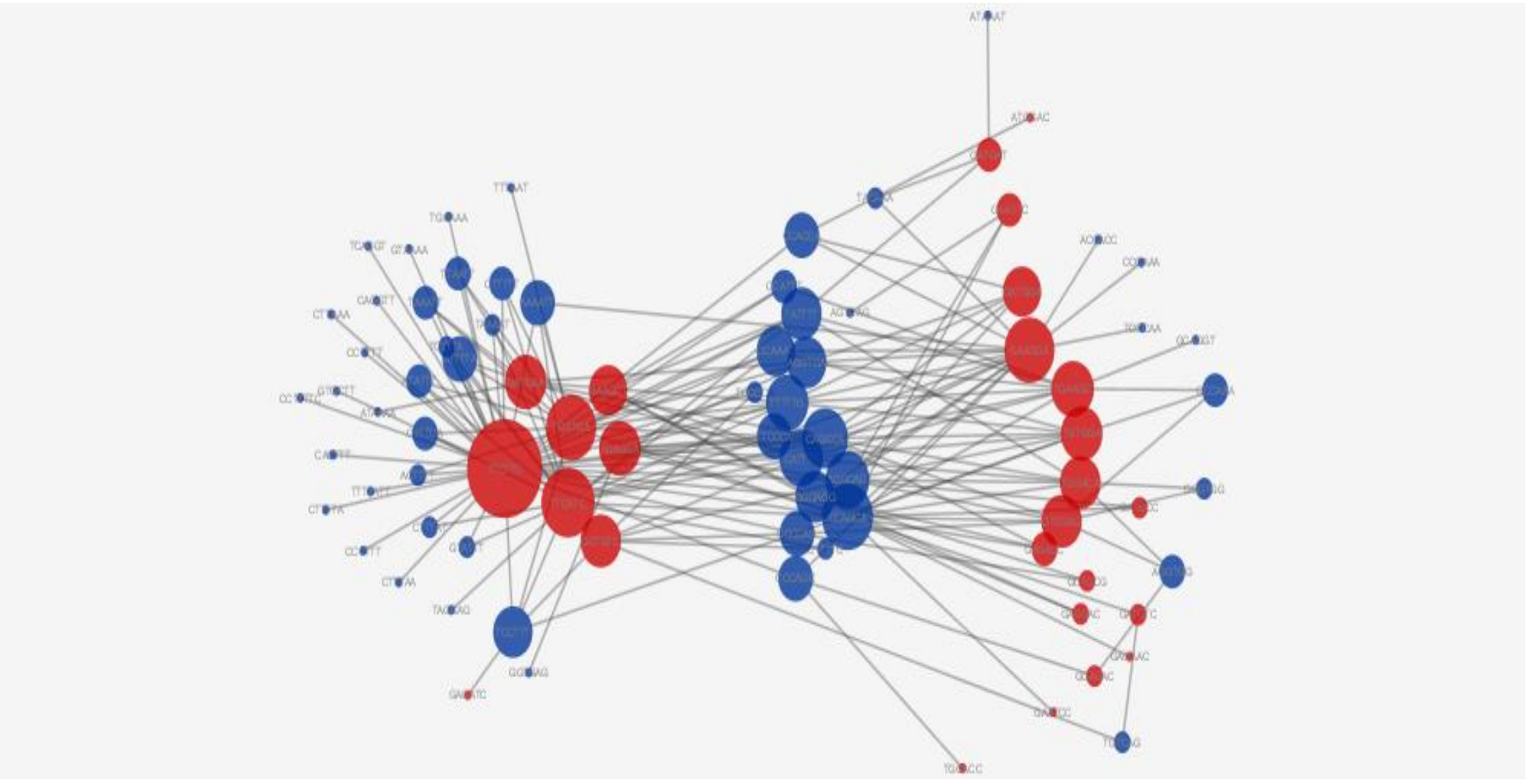


Fig 7. Combinatorial SRE network in brain tissue