A Specialized Learner for Inferring Structured cis–Regulatory Modules

Keith Noto and Mark Craven

noto@cs.wisc.edu craven@biostat.wisc.edu

Department of Computer Sciences
Department of Biostatistics and Medical Informatics
University of Wisconsin–Madison

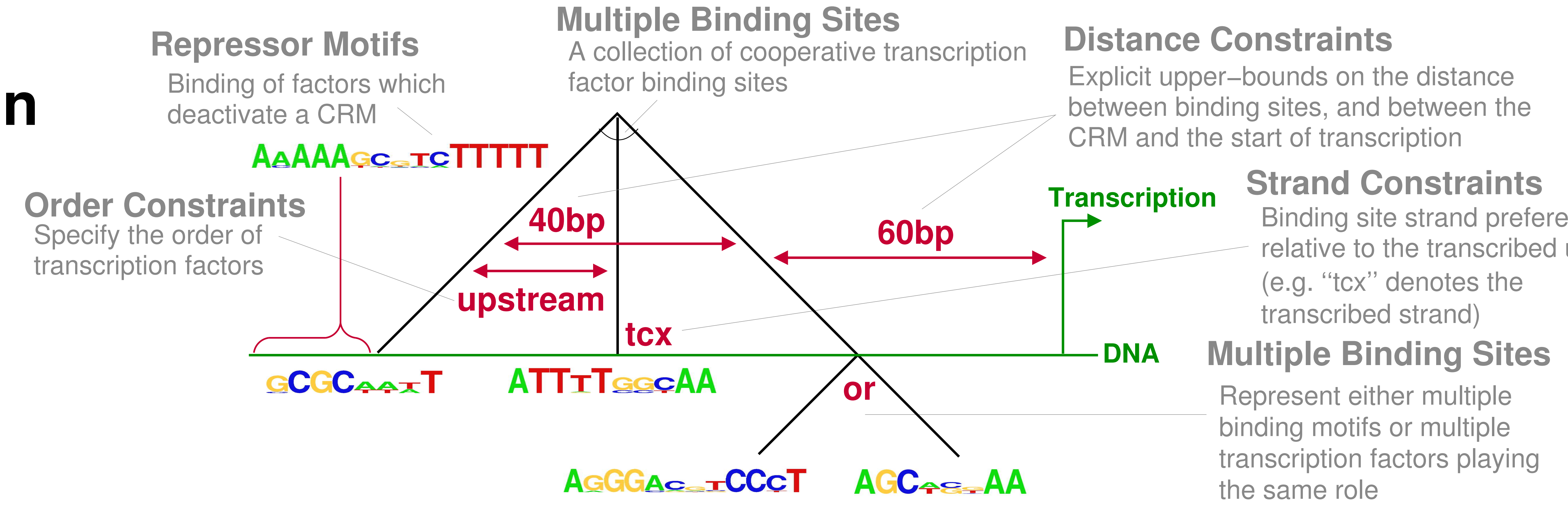
This research was supported in part by NIH/NLM training grant 5T15LM005359, NSF grand IIS–0093016, and NIH/NLM grant R01–LM07050–01.

Abstract

We present an approach to identifying cis–regulatory modules (CRMs) in terms of binding site motifs and the arrangement of their locations relative to the transcriptional start site. It is expressive enough to capture important structural aspects of a CRM, yet the search algorithm is specifically tailored to this context.

1. An Expressive CRM Representation

Transcription factors bind to DNA in specific arrangements and interact with each other. This system is called a cis–regulatory module (CRM). Our work is motivated by the need for more expressive CRM representations which can capture important structural ***aspects*** (right).



Repressor Motifs
Binding of factors which deactivate a CRM

Multiple Binding Sites
A collection of cooperative transcription factor binding sites

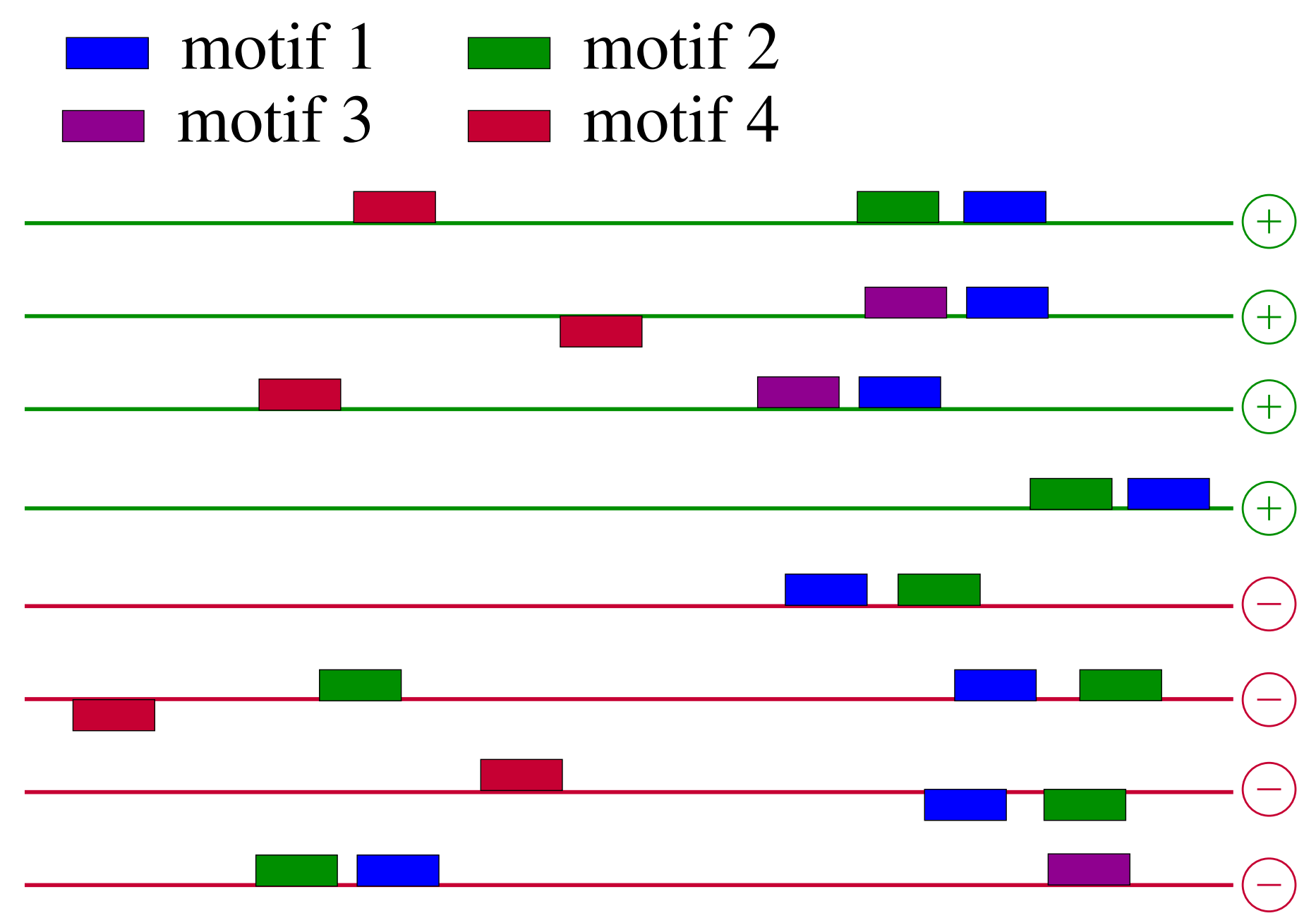
Distance Constraints
Explicit upper–bounds on the distance between binding sites, and between the CRM and the start of transcription

Order Constraints
Specify the order of transcription factors

Strand Constraints
Binding site strand preferences relative to the transcribed unit (e.g. "tcx" denotes the transcribed strand)

Multiple Binding Sites
Represent either multiple binding motifs or multiple transcription factors playing the same role

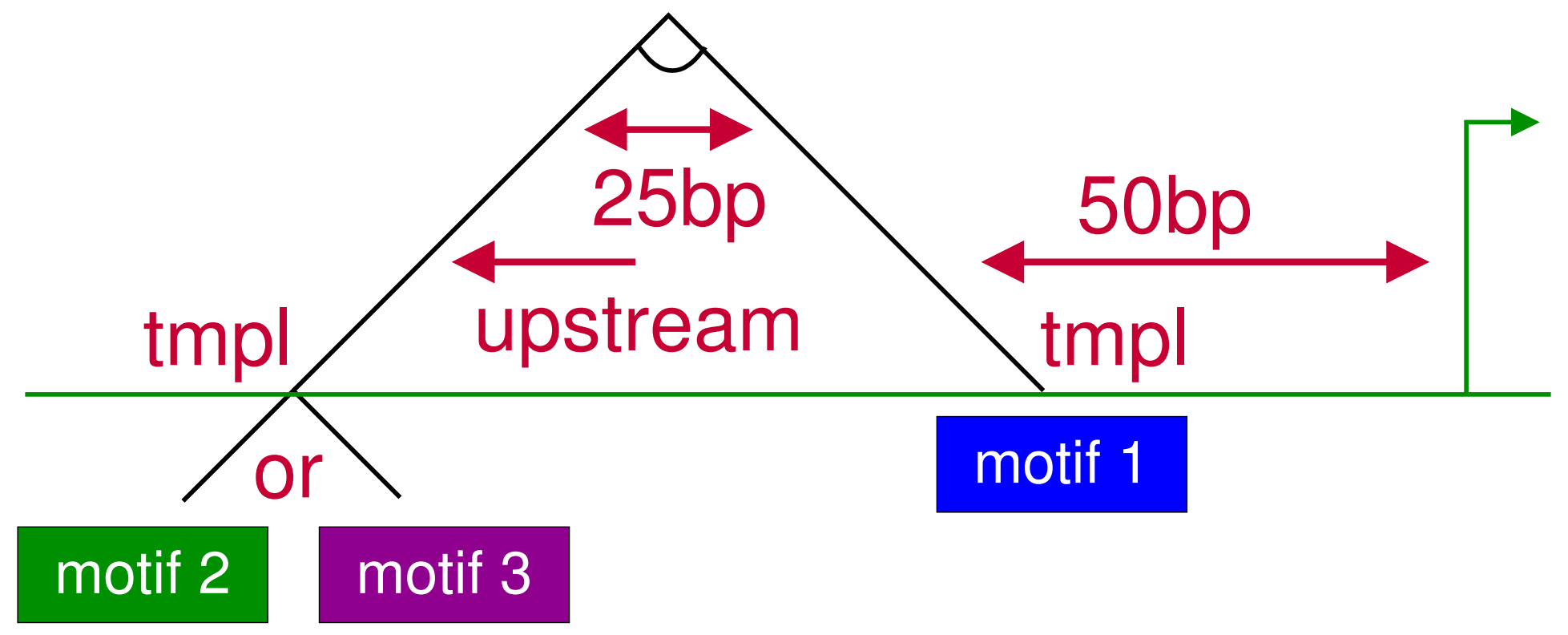
2. Task: Learn a CRM Model from data



Given: A set of DNA sequences thought to contain a CRM, a set thought not to, and a set of motifs

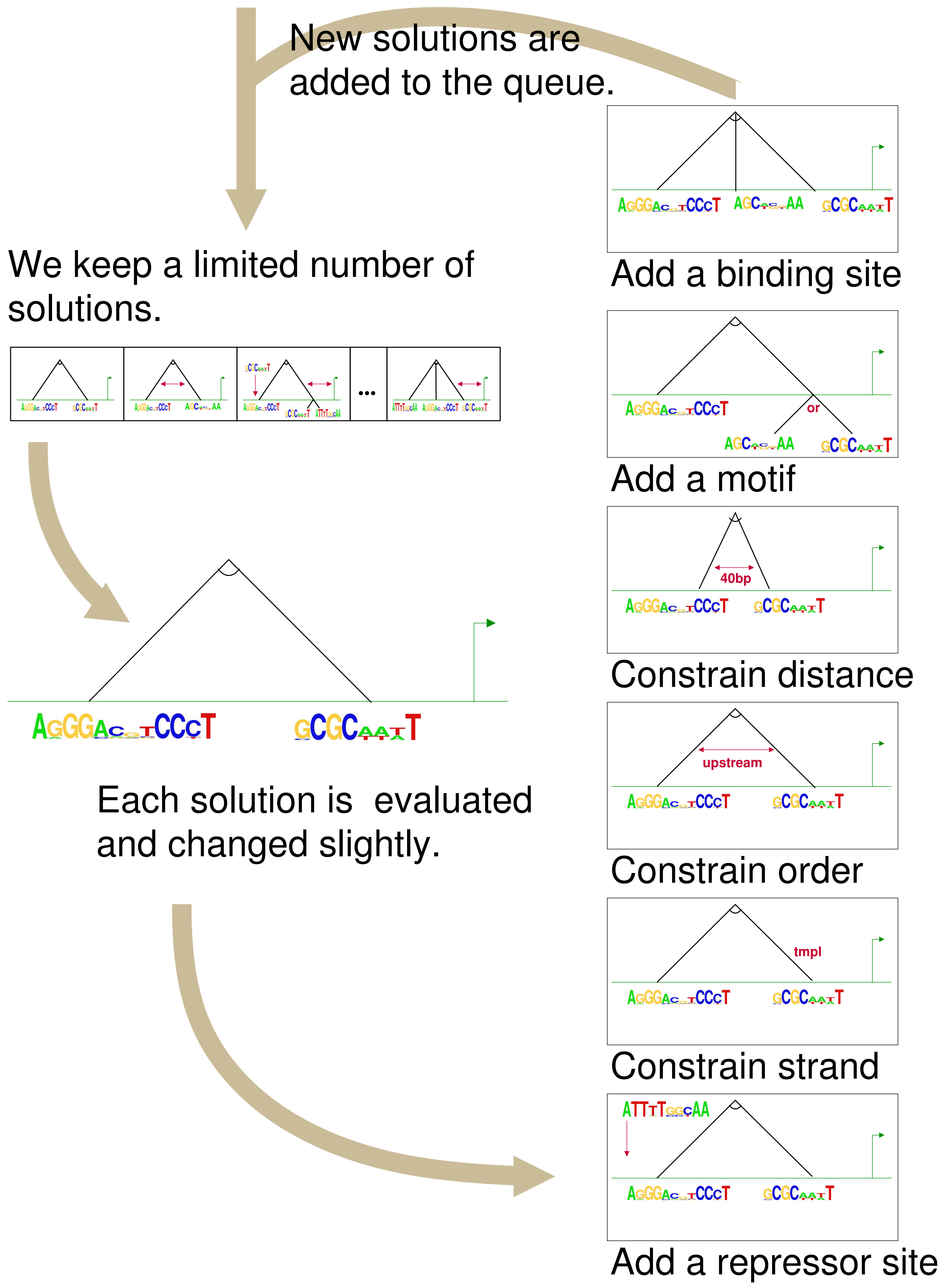
Do: Learn a CRM model that distinguishes between positive and negative examples

A CRM model for the task above:



3. Learning a CRM Model

We start with an "empty" model.



New solutions are added to the queue.

We keep a limited number of solutions.

Each solution is evaluated and changed slightly.

Add a binding site

Add a motif

Constrain distance

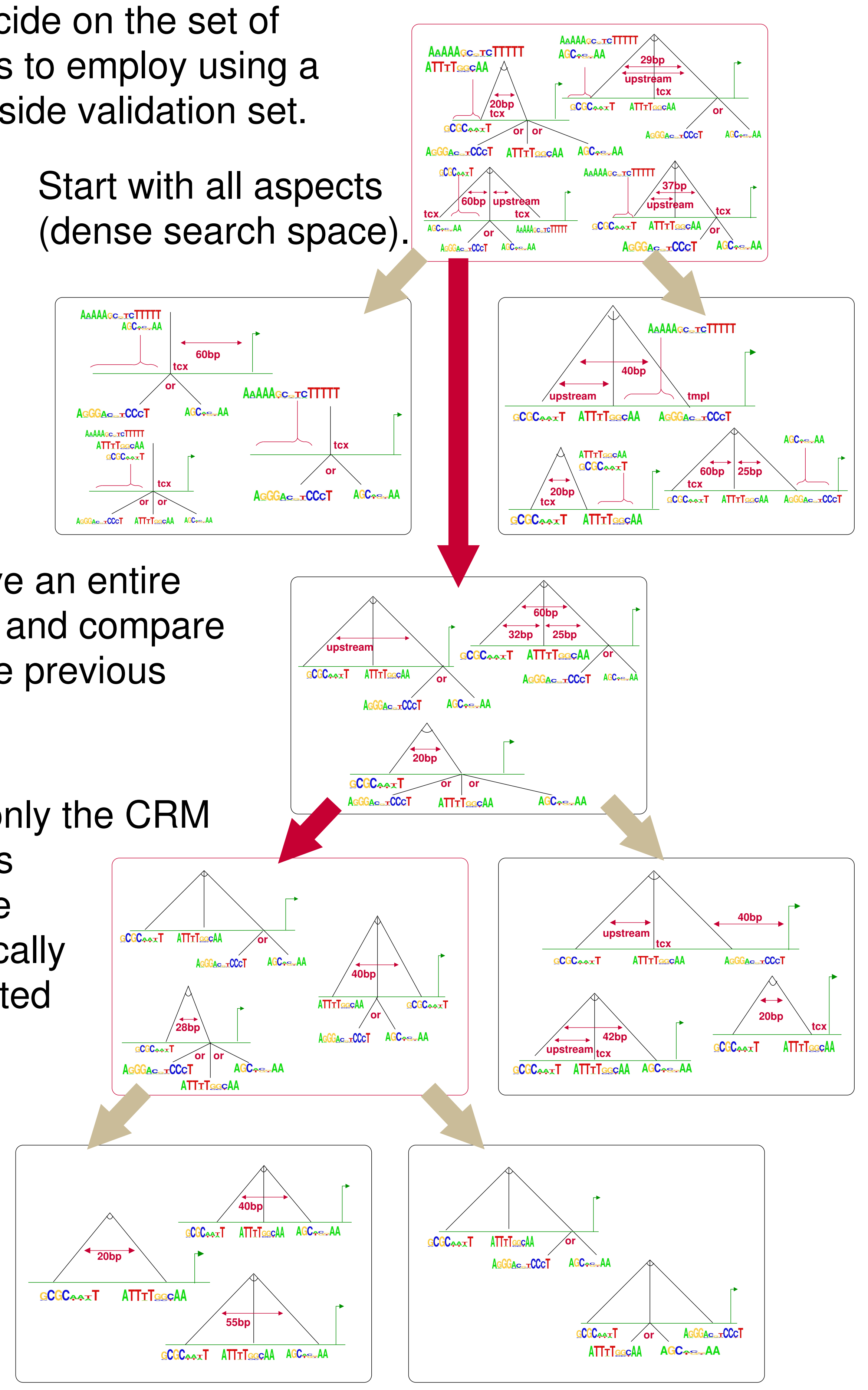
Constrain order

Constrain strand

Add a repressor site

4. Controlling Expressivity

The CRM **aspects** determine the set of possible models. We decide on the set of aspects to employ using a held–aside validation set.



Start with all aspects (dense search space).

Remove an entire aspect and compare with the previous space.

Keep only the CRM aspects that are statistically supported by the data.

5. Results

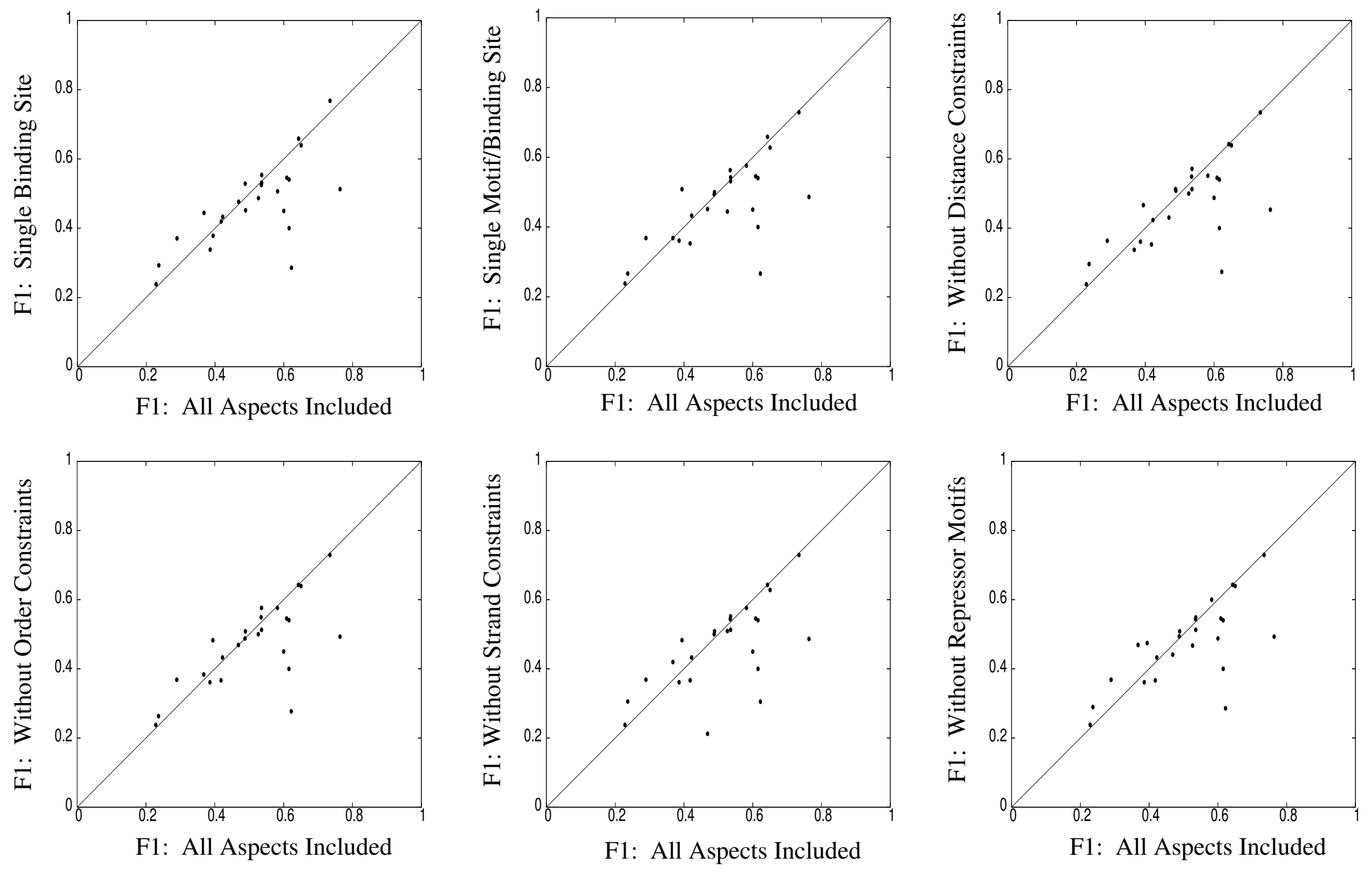
We find significant CRMs (test set p–value < 0.01) in 17 of 25 data sets from *Saccharomyces cerevisiae.*

Data set	TP	FP	TN	FN	P	R	F1	p-value	Data set	TP	FP	TN	FN	P	R	F1	p-value
GAT3, RGM1	4	16	84	11	0.200	0.267	0.229	0.246783	GAL4, YAP5	8	2	98	8	0.800	0.500	0.615	7.95E-007
GAT3, PDR1	12	10	90	5	0.545	0.706	0.615	3.27E-007	CIN5, NRG1	14	13	87	4	0.519	0.778	0.622	7.11E-008
RGM1, YAP5	9	3	97	9	0.750	0.500	0.600	9.47E-007	NDD1, SWI4	11	12	88	11	0.478	0.500	0.489	0.000206
SKN7, SWI4	11	60	40	11	0.155	0.500	0.237	0.864014	PDR1, YAP5	11	23	77	12	0.324	0.478	0.386	0.018610
FKH2, SWI4	14	33	67	10	0.298	0.583	0.394	0.020587	PHD1, YAP6	15	25	75	9	0.375	0.625	0.469	0.000692
FHL1, YAP5	15	16	84	10	0.484	0.600	0.536	2.36E-005	FKH2, MCM1	15	16	84	10	0.484	0.600	0.536	2.36E-005
MBP1, NDD1	11	40	60	14	0.216	0.440	0.289	0.442532	ACE2, SWI5	42	17	83	9	0.712	0.824	0.764	4.22E-015
FKH2, MBP1	20	35	62	7	0.364	0.741	0.488	0.000460	MCM1, NDD1	21	20	80	7	0.512	0.750	0.609	1.26E-007
RAP1, YAP5	16	10	90	13	0.615	0.552	0.582	1.03E-006	NRG1, YAP6	16	41	59	14	0.281	0.533	0.368	0.162475
GAT3, YAP5	27	18	82	12	0.600	0.692	0.643	1.79E-008	CIN5, YAP6	25	30	70	15	0.455	0.625	0.526	0.000410
MBP1, SWI4	27	34	66	13	0.443	0.675	0.535	0.000305	SWI4, SWI6	28	63	37	15	0.308	0.651	0.418	0.482183
MBP1, SWI6	40	39	61	4	0.506	0.909	0.650	1.73E-009	FKH2, NDD1	34	77	23	16	0.306	0.680	0.422	0.915385
FHL1, RAP1	94	48	52	20	0.662	0.825	0.734	8.43E-008									

Data sets from Segal and Sharan, *A Discriminative Model for Identifying Spatial cis–Regulatory Modules,* RECOMB 2004.

Precision P=TP/TP+FP. Recall R=TP/TP+FN. F1 score = 2PR/P+R.

All CRM aspects have predictive value. Each aspect, when removed from consideration during the search phase (previous section), tends to decrease the accuracy (test set F1 score) of the learned models.



F1: Single Binding Site
F1: Single Motif/Binding Site
F1: Without Distance Constraints
F1: Without Order Constraints
F1: Without Strand Constraints
F1: Without Repressor Motifs
F1: All Aspects Included