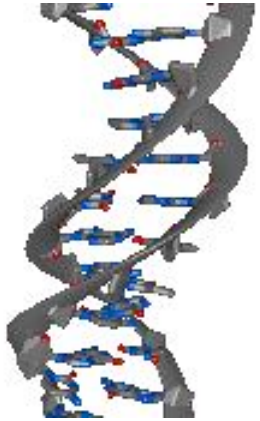
Finding Short DNA Motifs Using Permuted Markov Models

Xiaoyue Zhao

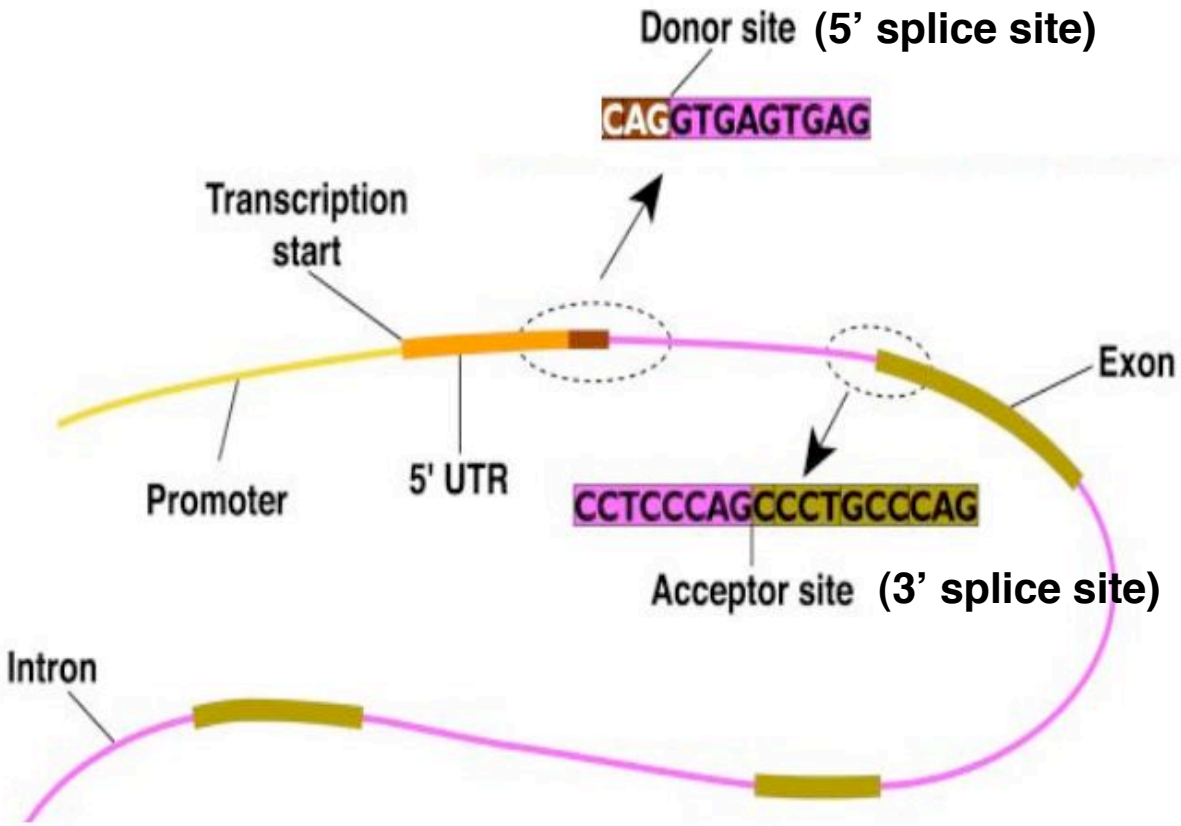


Department of Statistics
University of California at Berkeley

Outline

- Some Biological Background
- Models and Modelling
- Results and Discussions
- Future Work

Cartoon of gene structure



12 examples of 5'splice (donor) sites

exon	intron
TCG	GTGAGT
TGG	GTGTGT
CCG	GTCCGT
ATG	GTAAGA
TCT	GTAAGT
CAG	GTAGGA
CAG	GTAGGG
AAG	GTAAGG
AGG	GTATGG
TGG	GTAAGG
GAG	GTTAGT
CAT	GTGAGT

Probability models for short DNA motifs

Short: ~6-20 base pairs

DNA motifs: splice sites, transcription factor binding sites, translation initiation sites, enhancers, promoters,...

Why probability models?

- to ***characterize*** the motifs
- to help ***identify*** them
- for ***incorporation*** into larger models

Aim: given a number of instances of a DNA sequence motif, we want a ***model*** for the ***probability*** *of that motif*.

Weight matrix models, Staden (1984)

Base	-3	-2	-1	0	+1	+2	+3	+4	+5
A	33	61	10	0	0	53	71	7	16
C	37	13	3	0	0	3	8	6	16
G	18	12	80	100	0	42	12	81	22
T	12	14	7	0	100	2	9	6	46

A weight matrix for donor sites.
Essentially a **mutual independence** model.
An improvement over the **consensus** CAGGTAAGT.
But we have to go **beyond independence**…

Some notation

L: length of the DNA sequence motif.

X_i: discrete random variable at position i, taking values from the set $\chi = \{A, C, G, T\}$.

$\mathbf{x} = (x_1, \ldots, x_L)$: a DNA sequence motif of length L.

$\mathbf{x}_1^j$ ($j>1$): the sequence $(x_1, x_2, \ldots, x_j)$.

$P(\mathbf{x})$: probability of $\mathbf{x}$.

Variable Length Markov Models
(Rissanen (1986), Buhlmann & Wyner (1999))

Factorize $P(\boldsymbol{x})$ in the usual telescopic way:

$$P(X_1 = x_1) \times \prod_{l=2}^{L} P(X_l = x_l \mid X_1^{l-1} = x_1^{l-1})$$

We now simplify this using **context functions** $c_l : \chi^{l-1} \rightarrow \cup_{i=0}^{l-1} \chi^i$, $l = 2,\ldots,L$, to

$$P(X_1 = x_1) \times \prod_{l=2}^{L} P(X_l = x_l \mid c_l(X_1^{l-1}) = c_l(x_1^{l-1}))$$

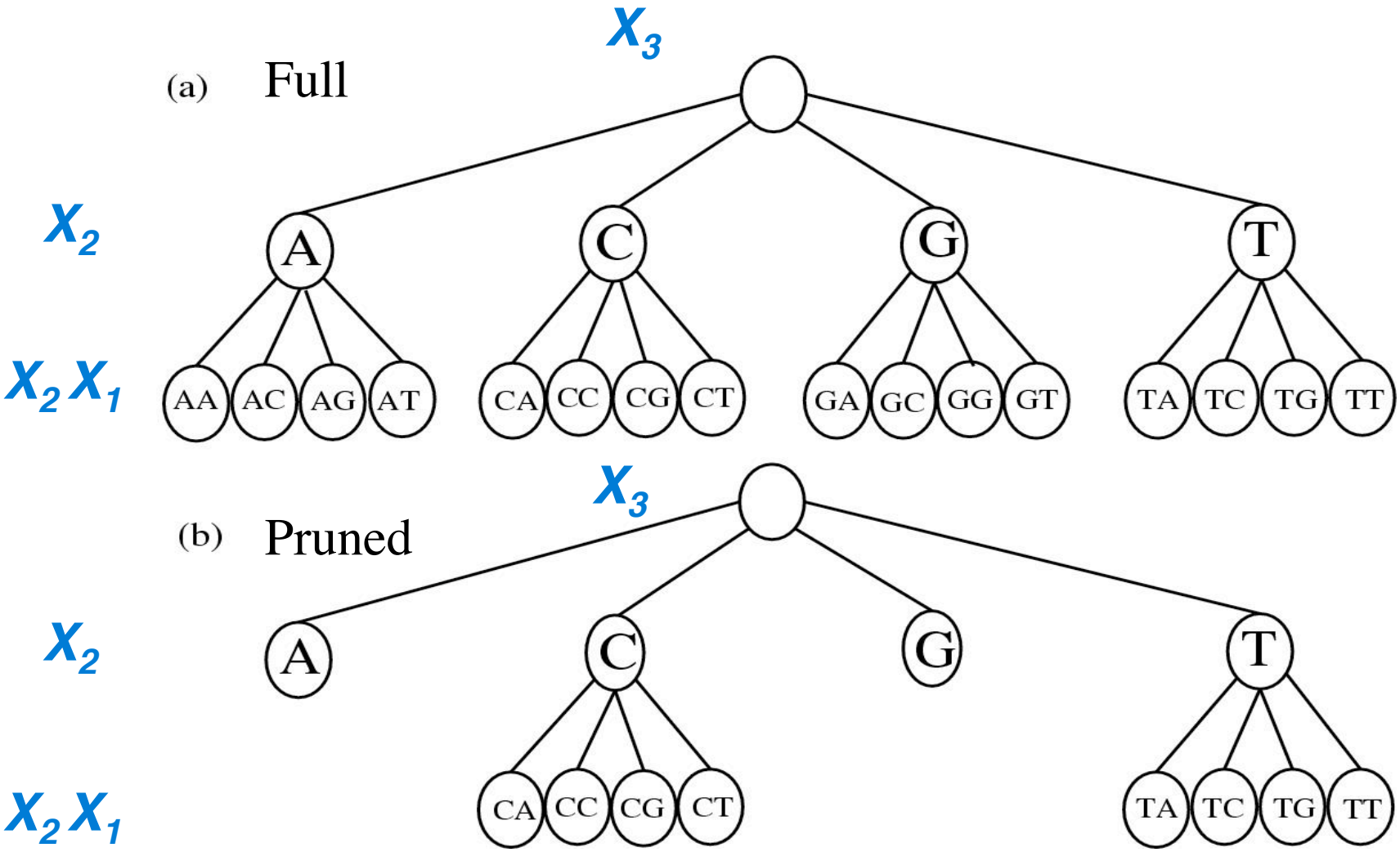
VLMM cont.

A VLMM for a DNA sequence motif of length L is specified by

- a distribution for X_1, and, for $l = 2,\ldots L$,
- a **constrained** conditional distribution for X_l given $X_{l-1},\ldots,X_1$.

That is, we need ***L-1* context functions**.

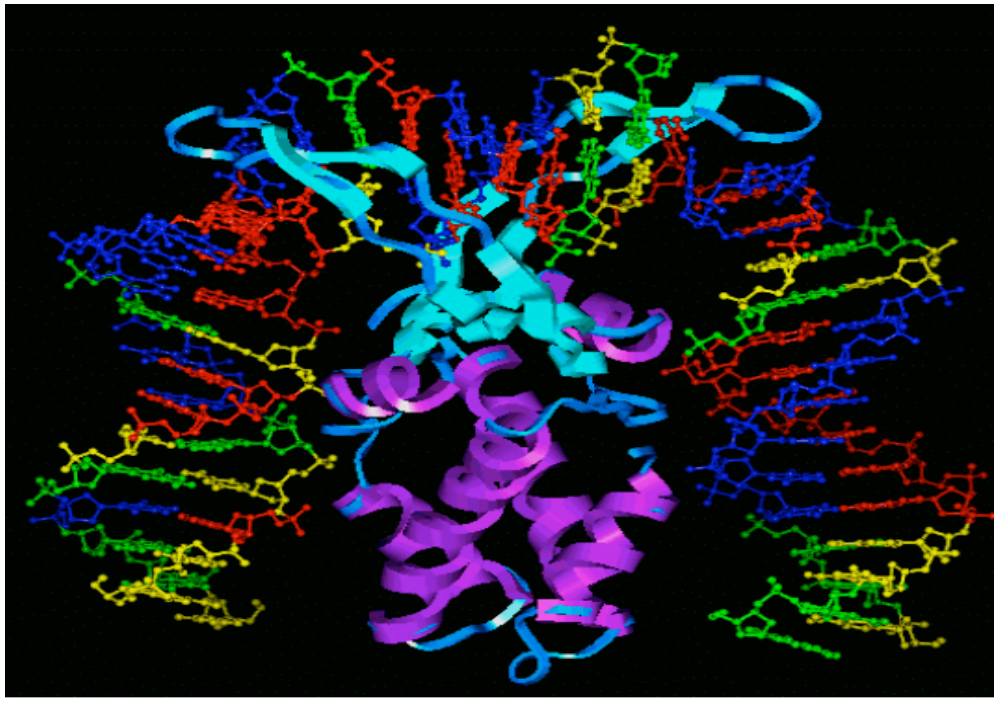
VLMM: an illustrative example



Pruned context: $P(X_3|X_2=A, X_1) = P(X_3|X_2=A)$, etc.

Sequence dependencies (interactions) are not always local

3-dimensional folding; DNA, RNA & protein interactions



The methods outlined so far all fail to incorporate **long-range** (≥4 bp) interactions. **New model types** are needed!

Modelling “long-range” dependency

The principal work in this area is Burge & Karlin’s (1997) **maximal dependence decomposition (MDD)** model.

Cai *et al* (2000) and Barash *et al* (2003) used **Bayesian networks (BN)**.

Yeo & Burge (2003) applied **maximum entropy models**.

Ellrott *et al* (2002) **ordered** the **sequence** and applied Markov models to the motif.

We have adapted this last idea, to give **permuted variable length Markov models (PVLMM).**

Permuted variable length MMs (PVLMM)

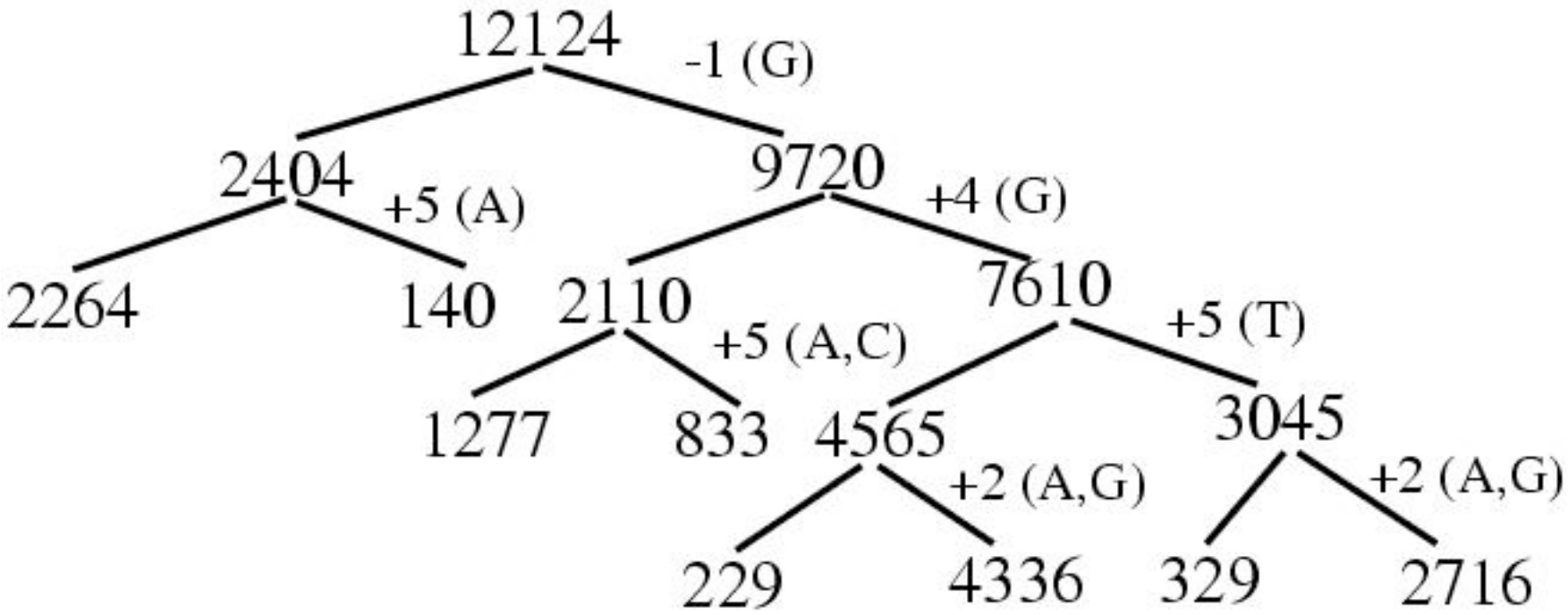
Let $\pi = (\pi(1), \pi(2), ..., \pi(L))$ be a permutation. By using VLMM, we can simplify the context terms in the equation:

$$P(X_{\pi(1)} = x_{\pi(1)}) \times \prod_{l=2}^{L} P(X_{\pi(l)} = x_{\pi(l)} \mid X_{\pi(1)}^{\pi(l-1)} = x_{\pi(1)}^{\pi(l-1)})$$

The simplified equation has the form:

$$P(X_{\pi(1)} = x_{\pi(1)}) \times \prod_{l=2}^{L} P(X_{\pi(l)} = x_{\pi(l)} \mid c_l(X_{\pi(1)}^{\pi(l-1)}) = c_l(x_{\pi(1)}^{\pi(l-1)})).$$

Part of a context (decision) tree for position -2 of a splice donor PVLMM



Node #s: counts
Edge #s: split variables

Sequence order: +2 +5 -1 +4 -2 +3 -3

Issues in statistical modelling of short DNA motifs

In any study of this kind, essential items are:

- the **model class** (e.g. WMM, PVLMM, or MDD)
- the way we **search** through the model class (e.g. by forward selection & simulated annealing)
- the way we **compare** models when searching (e.g. by AIC, BIC, or MDL), and finally,
- the way we **assess** the final model in relation to our aims (e.g. by cross-validation).

We always need interesting, high-quality **datasets.**

Model assessment: Stand-alone recognition

M: ***motif*** model *B:* ***background*** model

Given a sequence $\boldsymbol{x} = (x_1, \ldots, x_L)$, we ***predict*** $\boldsymbol{x}$ to be a ***motif*** if

$$\log \{P(\boldsymbol{x} \mid M) / P(\boldsymbol{x} \mid B)\} > c,$$

for a suitably chosen ***threshold*** value *c*.

Transcription factor binding sites

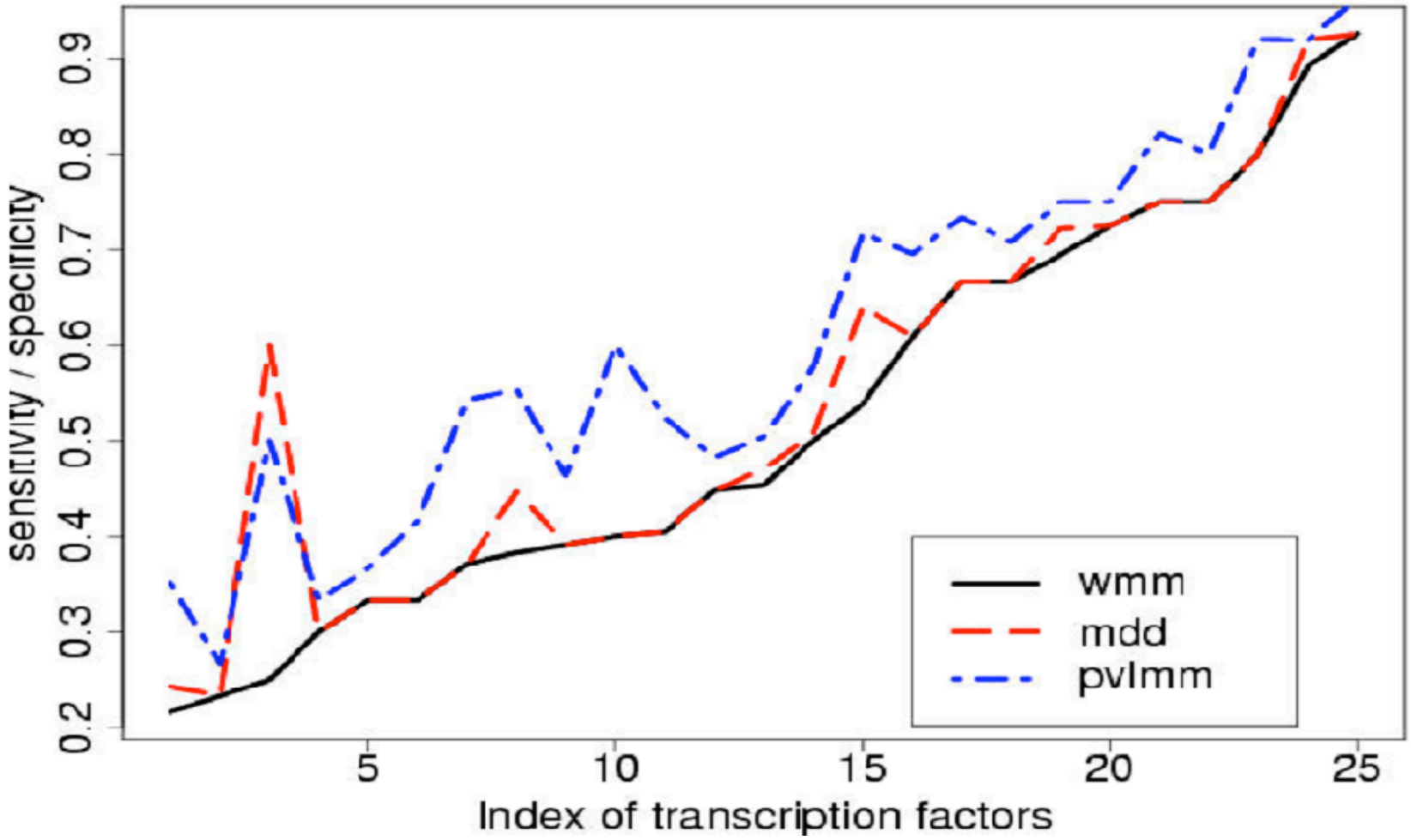
These are of **great interest**, their signals are **very weak**, and we typically have only a **few instances**.

TRANSFAC (Wingender et al 2000).

We have studied 61 TFs with effective length ≤ 9 and ≥ 20 instances, 2,238 sites in all.

In **25/61** cases we are able to improve upon WMM.
In the remaining, PVLMM performs similarly to WMM.

TFBS: results for 25/61



Three transcription factors

Sensitivity

wmm: 0.37

pvlmm: **0.54**

mdd: 0.37

wmm: 0.38

pvlmm: **0.55**

mdd: 0.45

wmm: 0.54

pvlmm: **0.72**

mdd: 0.64

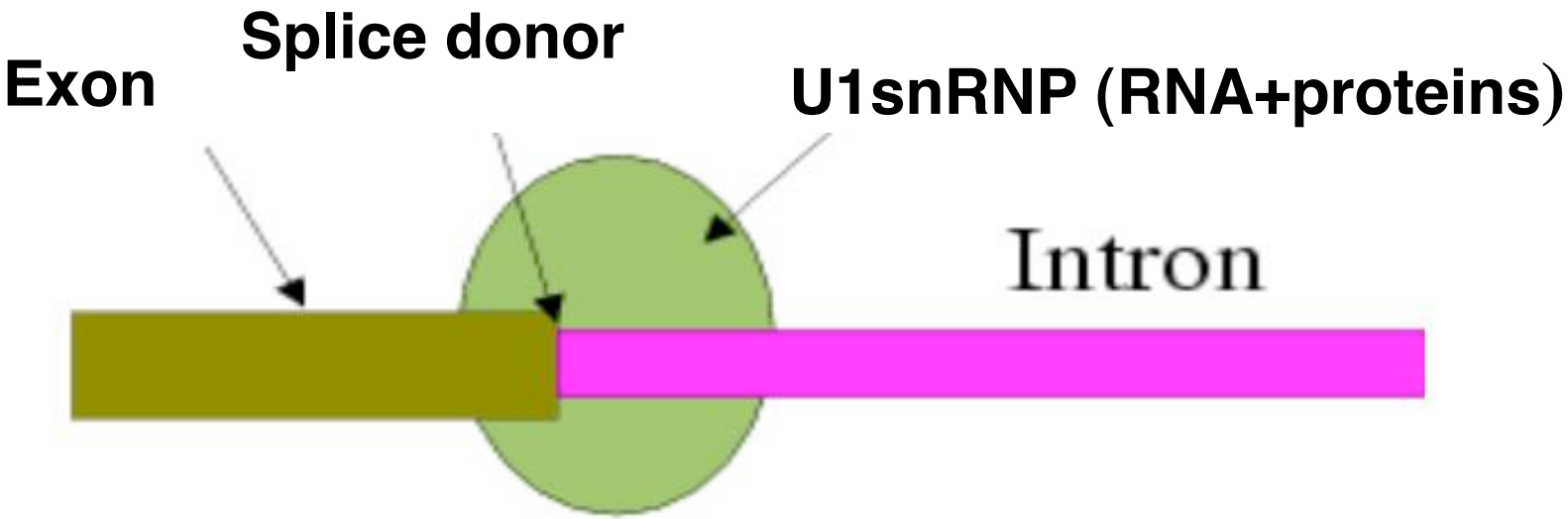
Splice donor sites

Mammalian donor sequences from **SpliceDB,** Burset *et al* (2001).

15,155 **canonical donor sites** of length 9, with GT conserved at position 0 and 1.

Mammalian donor site

Beginning of the splicing process

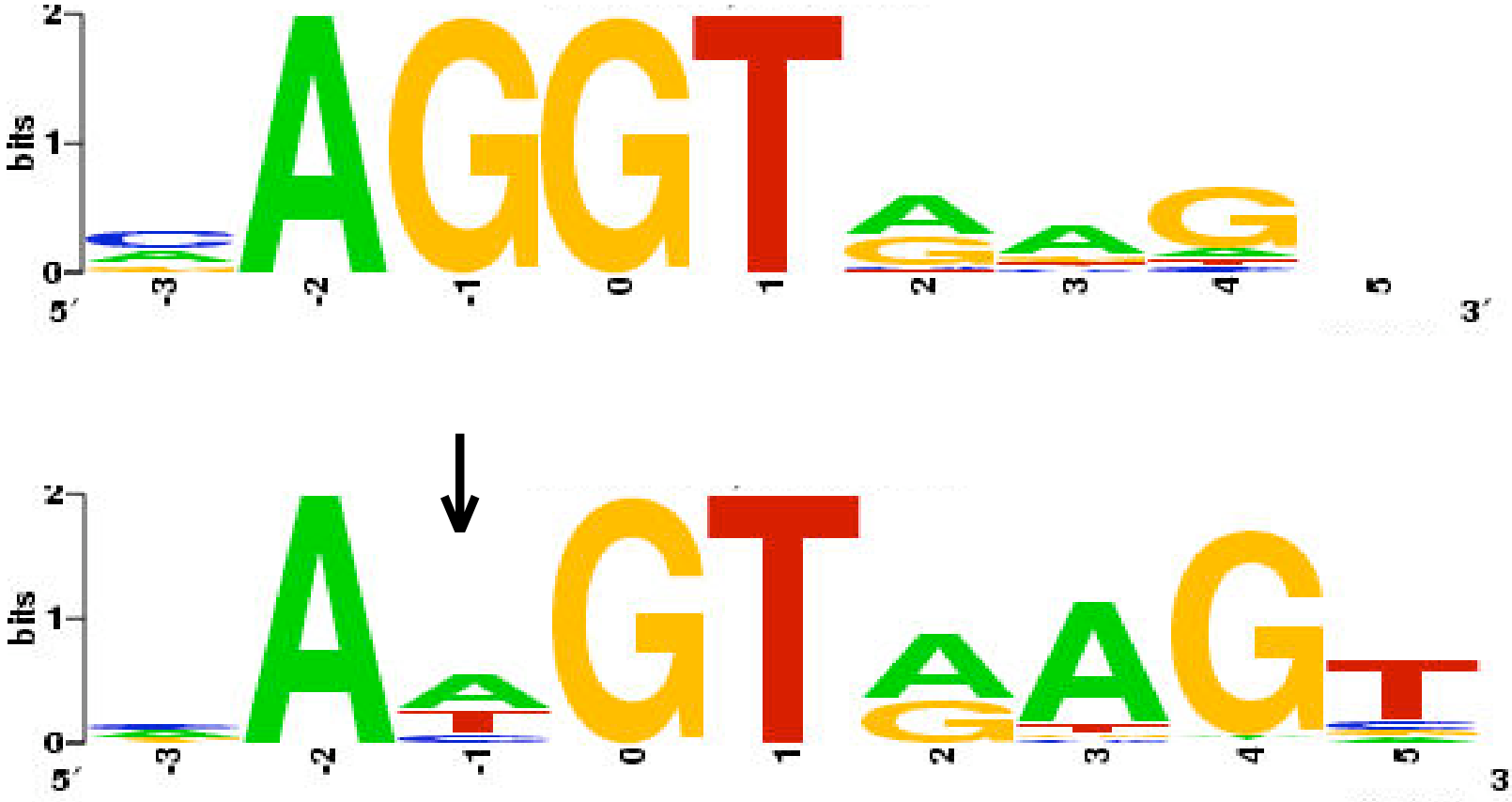


Interpretation of PVLMM model selected

Mammalian donor site

Base	-3	-2	-1	0	+1	+2	+3	+4	+5
A	33	**61**	10	0	0	**53**	**71**	7	16
C	**37**	13	3	0	0	3	8	6	16
G	18	12	**80**	**100**	0	42	12	**81**	22
T	12	14	7	0	**100**	2	9	6	**46**
U1 snRNA	**G**	**U**	**C**	**C**	**A**	**U**	**U**	**C**	**A**

“Long-range” dependence in the model: 5’/3’ compensation



U1 sn RNA: G U C C A U U C A

Optimal permutation: +2 +3 +4 -1 -2 **+5** -3

Integrating into a gene finder

SLAM (Pachter *et al*, 2002): a eukaryotic cross-species gene finder (generalized pair hidden Markov model (GPHMM)).

Results at the exon level

	VLMM	MDD	PVLMM
Sensitivity	360/465	365/465	**371/465**
Specificity	360/470	365/465	**371/464**

Conclusions

- **Systematic** framework for probabilistic modeling of short DNA motifs

  - **Statistical** modeling issues
  - Better prediction performance by modeling **local** & **non-local** dependence
  - Good **biological** interpretations

Future work

- Model the dependence in **protein** motifs
- Extend to find ***de novo*** motifs
- Find a way to deal with **indels**
- Incorporate **local** sequence information into PVLMM-based predictions
- **Jointly model** multiple motifs in one species and sites across species

Acknowledgements

Terry Speed, UCB & WEHI
Haiyan Huang, UCB

The SLAM team:
Simon Cawley, Affymetrix
Lior Pachter, UCB
Marina Alexandersson, FCC

Sourav Chatterji, UCB
WEHI bioinformatics lab

Thank you!

Assessing our models for TFBS recognition

We randomly inserted each of the 2,238 sites into a background sequence of length 1,000 simulated from a stationary 3rd order MM, with parameters estimated from a large collection of human sequence upstream of genes.

10-fold cross-validation (CV) was used for each TF. At each round of CV, 90% of the true sites were used to fit the models (PVLMM, MDD, WMM). We then scanned the sequences containing the remaining 10% of the sites, and selected the N top-scoring sequences as putative binding sites, where N was the size of the remaining 10%. Thus we have *sn = sp.*

Assessing our models for TFBS recognition

We randomly inserted each of the 2,238 sites into a background sequence of length 1,000 simulated from a stationary 3rd order MM, with parameters estimated from a large collection of human sequence upstream of genes.

10-fold cross-validation (CV) was used for each TF. At each round of CV, 90% of the true sites were used to fit the models (PVLMM, MDD, WMM). We then scanned the sequences containing the remaining 10% of the sites, and selected the N top-scoring sequences as putative binding sites, where N was the size of the remaining 10%. Thus we have *sn = sp.*

Maximal Dependence Decomposition

MDD starts with a mutual independence model as with WMMs. The data are then iteratively subdivided, at each stage splitting on the **most dependent** position, suitably defined. At the tips of the tree so defined, a **mutual independence model** across all remaining positions is used.

The details can vary according to the *splitting criterion* (Burge & Karlin used χ^2), the *actual splits* (binary, etc), and the *stopping rule*.

However, the result is always **a single tree.**

Longer sites: finding the best PVLMM needs simulated annealing

V$AP1_Q2.01 has 46 known sites of length 12

"Optimal" order: 8 7 6 4 3 5 10 11 9 1

Sn/Sp: WMM 0.5 PVLMM 0.69

V$CHOP_01 has 39 known sites of length 13

"Optimal" order: 6 12 11 2 5 1 10 9 7 4 3 8 0

Sn/Sp: WMM 0.69 PVLMM 0.84