

Pre-mRNA Secondary Structure Prediction Aids Splice Site Recognition

Donald J. Patterson, Ken Yasuhara, Walter L. Ruzzo

January 3-7, 2002

Pacific Symposium on Biocomputing



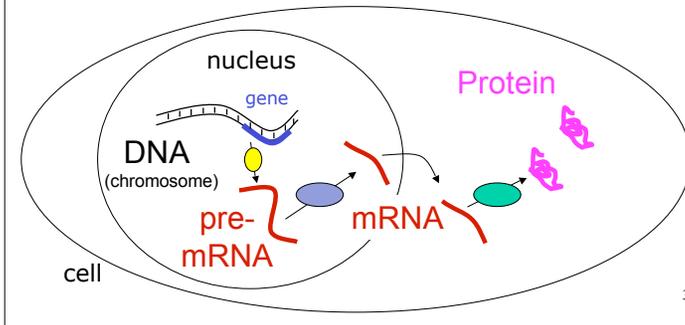
University of Washington Computational Molecular Biology Group

1

Gene Expression: Details

(Eukaryotes)

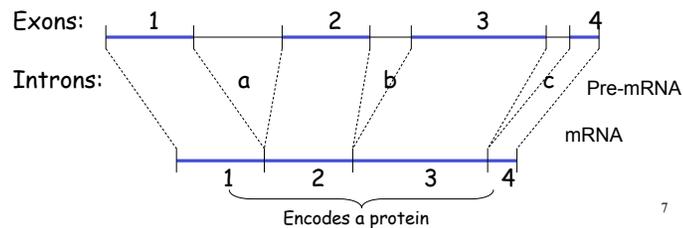
DNA → pre-mRNA → mRNA → Protein



3

Architecture of a Gene

- pre-mRNA's transcribed from most genes contain *introns*, which must be *spliced* out to form useful mRNAs



7

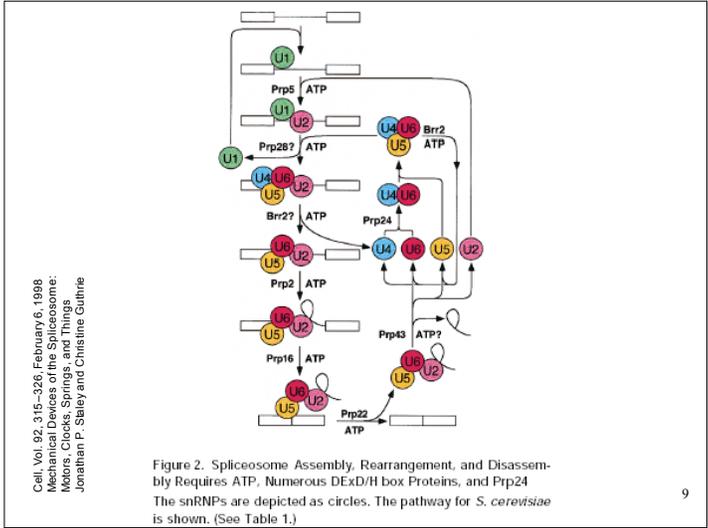
Characteristics of human genes

(Nature, 2/2001, Table 21)

	Median	Mean	Sample (size)
Internal exon	122 bp	145 bp	RefSeq alignments to draft genome sequence, with confirmed intron boundaries (43,317 exons)
Exon number	7	8.8	RefSeq alignments to finished sequence (3,501 genes)
Introns	1,023 bp	3,365 bp	RefSeq alignments to finished sequence (27,238 introns)
3' UTR	400 bp	770 bp	Confirmed by mRNA or EST on chromo 22 (689)
5' UTR	240 bp	300 bp	Confirmed by mRNA or EST on chromo 22 (463)
Coding seq	1,100 bp	1340bp	Selected RefSeq entries (1,804)*
(CDS)	367 aa	447 aa	
Genomic extent	14 kb	27 kb	Selected RefSeq entries (1,804)*

* 1,804 selected RefSeq entries were those with full-length unambiguous alignment to finished sequence

8



Relevance of Splice Prediction

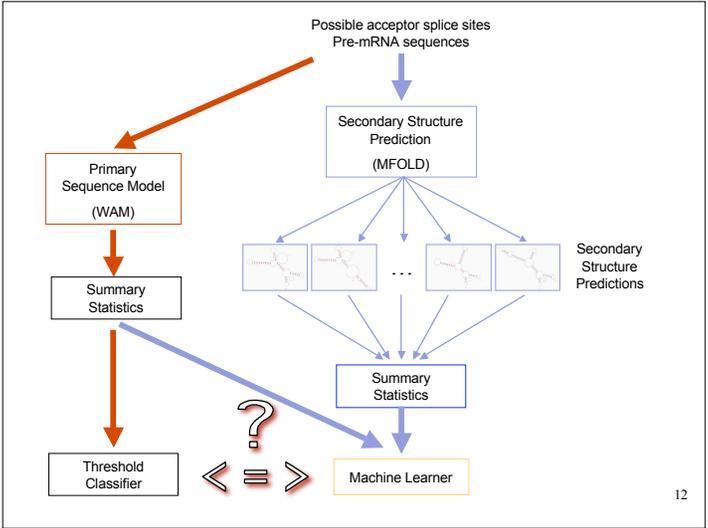
- Splice site prediction is critical to eukaryotic gene prediction.
 - Average human gene has 8.8 exons
 - Genes with over 175 exons known
 - Current primary sequence models do not display the same discriminatory power that cells exhibit *in vivo*
 - Small per-site error rate compounds

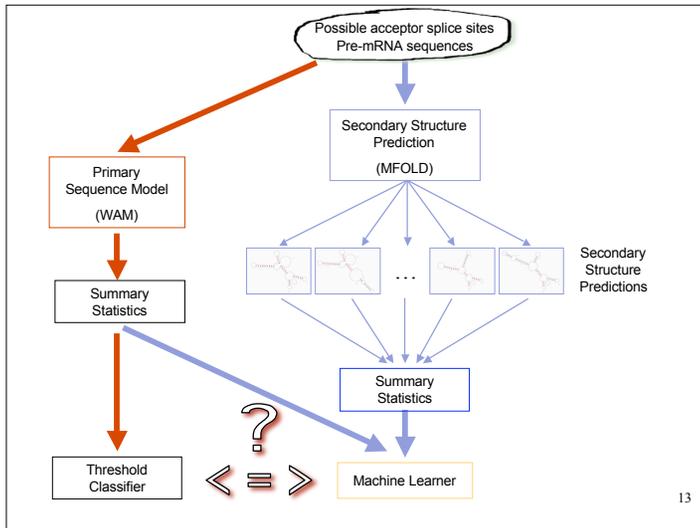
10

Hypothesis

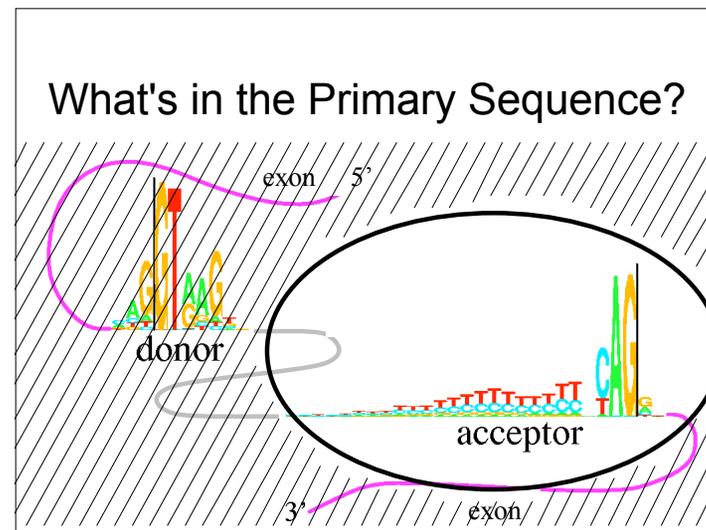
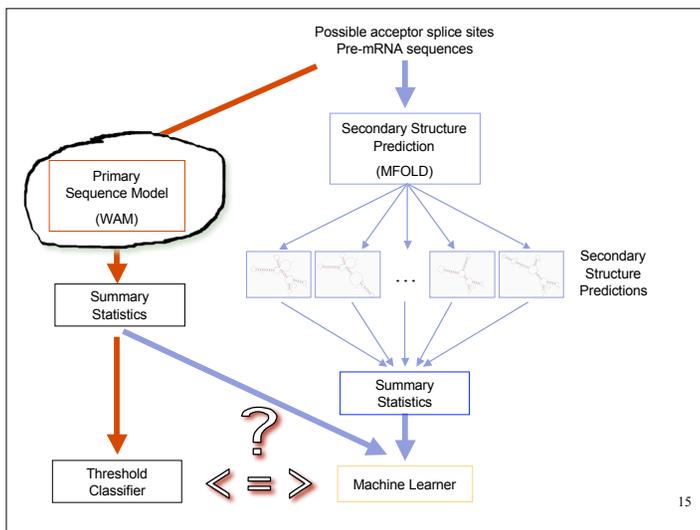
- Secondary structure contains information useful for predicting splice site location.
- This information is in addition to primary sequence information.
 - Specific instances of secondary structure variation affecting the splicing process.

11



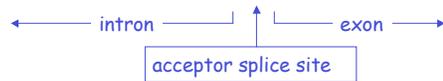


- ## Data Set
- Drawn from 462 unrelated, annotated, multi-exon human genes with standard splicing. (Reese 97)
 - 1,980 acceptor splice sites (3' end of intron)
 - 1,980 non-sites selected randomly
 - Aligned to an “AG” consensus
 - Located within 100 bases of an annotated acceptor splice site.
- 14



What's in the Primary Sequence?

	-4	-3	-2	-1	+1	+2	+3
A	22	4	100	0	25	25	27
C	33	74	0	0	13	21	27
G	22	0	0	100	52	22	24
T	22	21	0	0	9	32	23



Weight Matrix Model (0th order Markov Model)

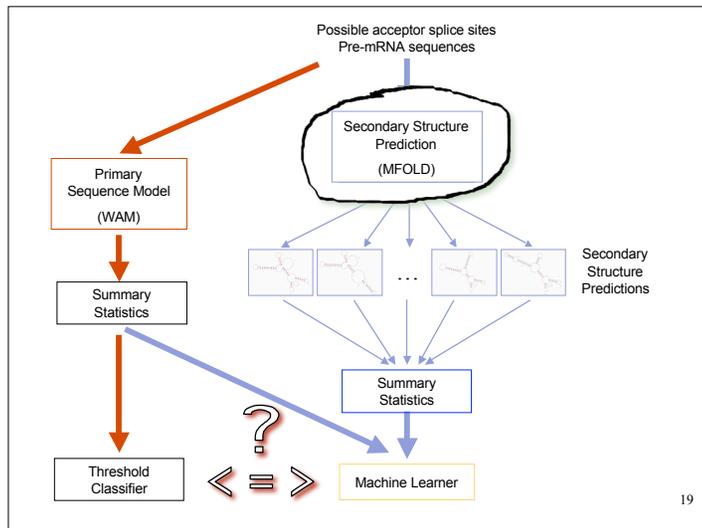
17

Sequence-based Metric

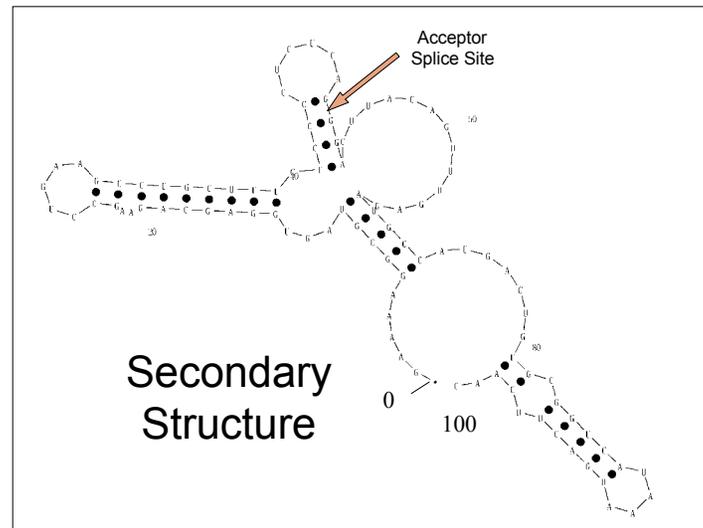
- 1st order Weight Array Matrix (WAM) / Markov Model
 - $P_i(N_i=\{A,C,G,U\} | N_{i-1}=\{A,C,G,U\})$
- Training
 - Generate two conditional probability tables for positions (-21,+3), one from positive examples and one from negative examples.
- Testing
 - For each sequence, x, calculate its likelihood ratio:

$$\log_{10} \left(\frac{P_{WAM}^+(x)}{P_{WAM}^-(x)} \right)$$

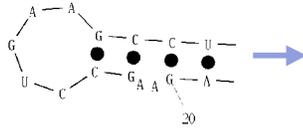
18



19



3. Neighbor Pairing Correlation Model

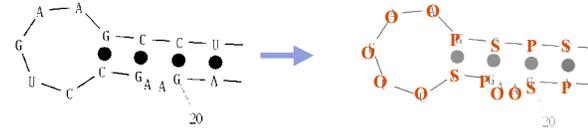


Change the pre-mRNA alphabet from nucleotides to structural symbols

- O** Unpaired base
- P** Paired base
- S** Paired and stacked base

25

3. Neighbor Pairing Correlation Model



Change the pre-mRNA alphabet from nucleotides to structural symbols

- O** Unpaired base
- P** Paired base
- S** Paired and stacked base

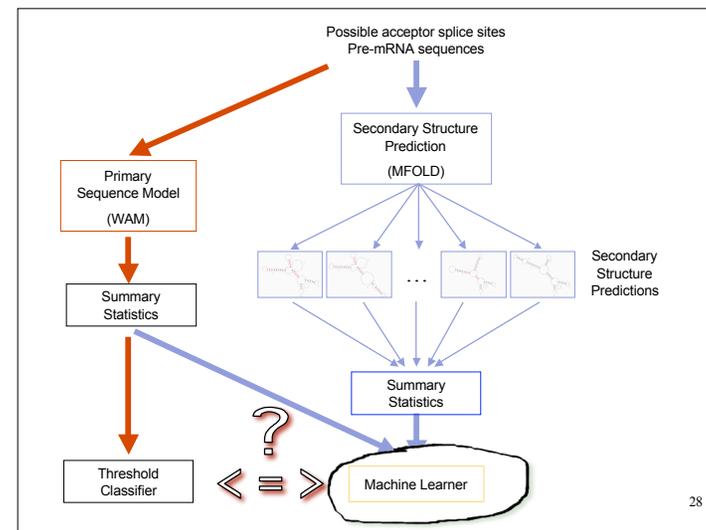
26

3. Neighbor Pairing Correlation Model

- 2nd order Markov Model
 - $P_i(N_i=\{O,P,S\} \mid N_{i-1}=\{O,P,S\} \wedge N_{i-2}=\{O,P,S\})$
- Training
 - Generate two conditional probability tables for positions $(-50,+3)$, one from positive examples and one from negative examples.
- Testing
 - For each sequence, x , calculate its log likelihood ratio:

$$\log_{10} \left(\frac{P_{NPCM}^+(x)}{P_{NPCM}^-(x)} \right)$$

27

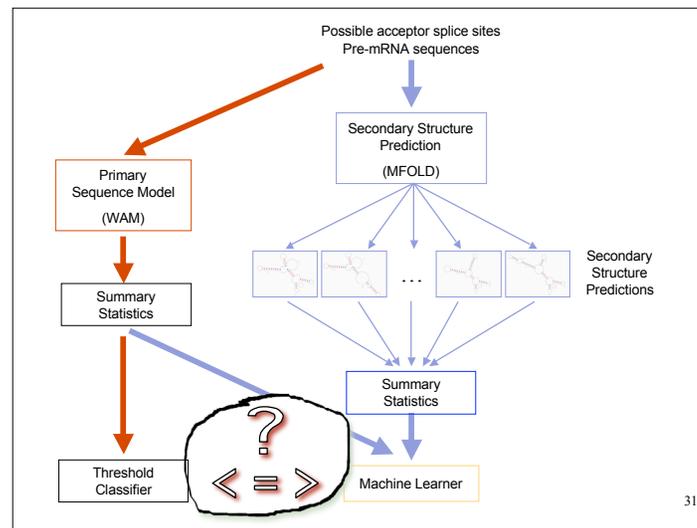


28

Machine Learners

- Decision Trees
 - Quinlan's C4.5
- Support Vector Machines
 - Noble's svm 1.1
 - Radial Basis Kernel degree 2
- Both take a vector of statistics and produce a yes/no binary classifier.

29



31

Results (Decision Trees)

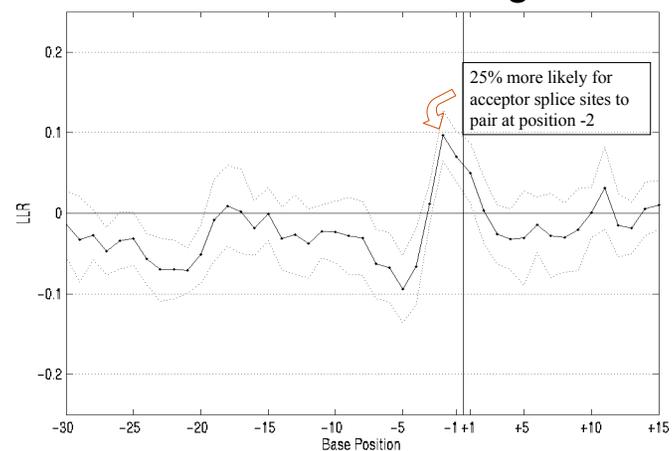
Features	Mean Accuracy (%)	% Error Reduction	p
WAM (baseline)	92.73		
WAM,OFE	93.13	5.5	0.066
WAM,OFE,NPCM	93.16	5.9	0.022
WAM,OFE,MH	93.21	6.6	0.009
WAM,OFE,NPCM,MH	93.13	5.5	0.016

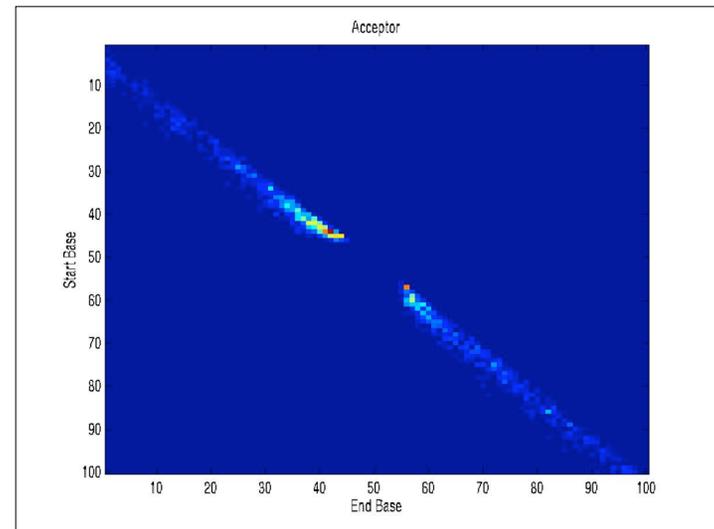
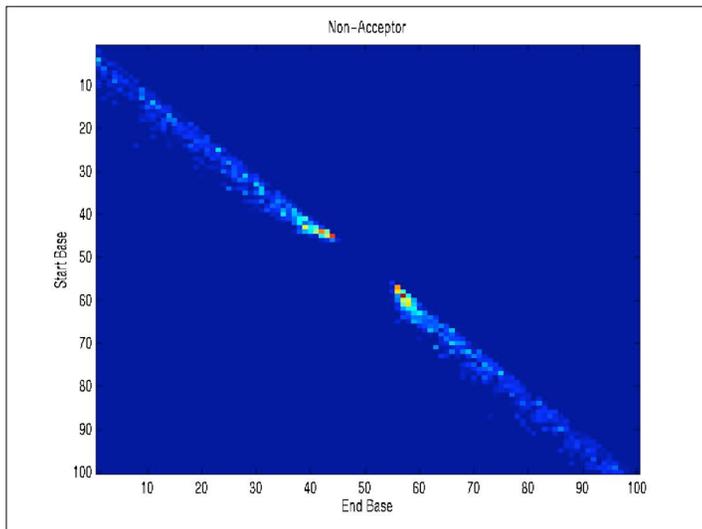
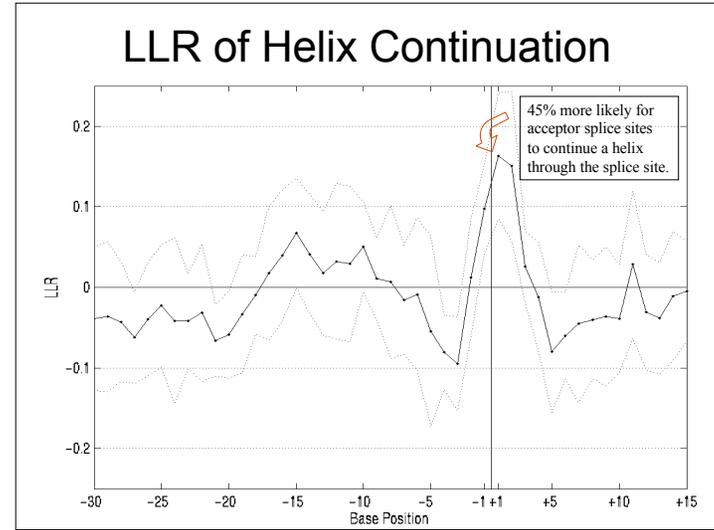
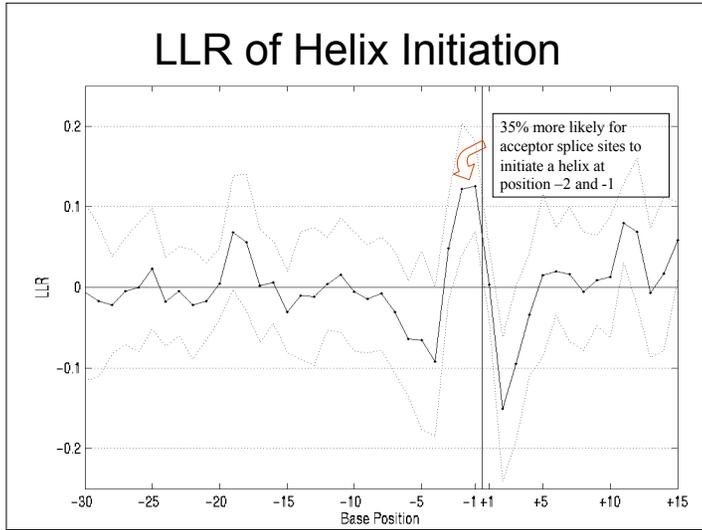
WAM = Weight Array Matrix (Primary Sequence Method)
 OFE = Optimal Free Energy
 MH = Max Helix
 NPCM = Neighbor Pairing Correlation Matrix

Wilcoxon p-value
 under 10-fold
 cross-validation

32

LLR of Base Pairing





Helix Formed at Splice Site

	Acceptor	Non-Acceptor
Pr(No Helix)	0.37	0.48
Pr(Helix)	0.63	0.52
Pr(Folds Left)	0.35	0.26
Pr(Folds Right)	0.28	0.26

38

Conclusions

- Secondary structure statistics correlate with splice site location.
- Our models (Max Helix, NPCM) can represent some of the relevant secondary structure.
- These models capture correlations that current primary sequence models don't capture.

39

Future Work

- Other organisms
 - *Oryza sativa* (rice) in progress
- Donor splice sites
- Other features?
- More structure models
 - Stochastic Context Free Grammars?

40

Acknowledgements

- Don Paterson
- Ken Yasuhara
- Jeff Stoner
- Kevin Chu

More Info

<http://www.cs.washington.edu/homes/ruzzo>



UW CSE Computational Biology Group

41