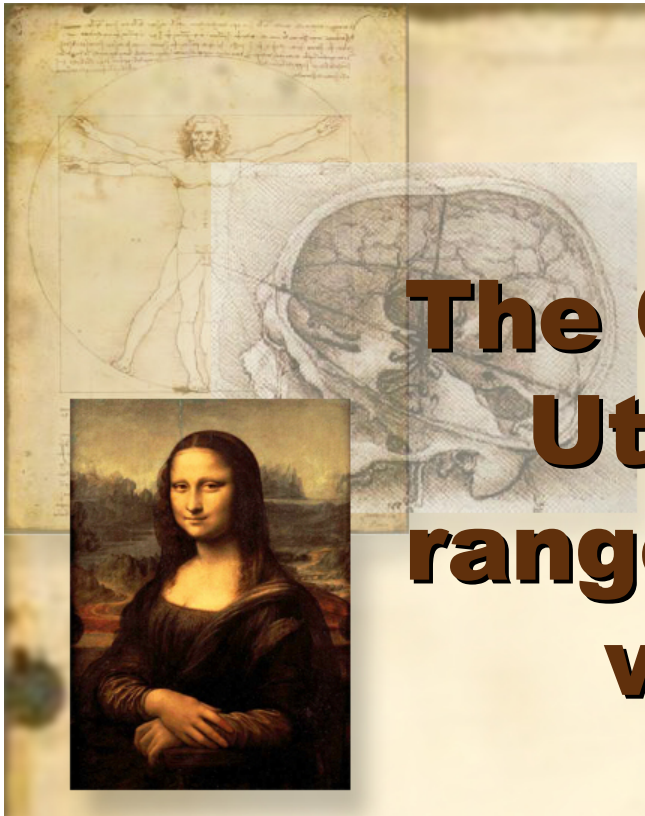


The Characterization and Utilization of Middle-range Sequence Patterns within the Human Genome

Presentation by Samuel Shepard

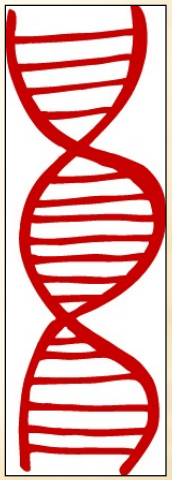
April 6th, 2010

Outline

- Discuss the non-randomness of the human genome in regions between 30 to 10,000 nucleotides.
- Introduce a new algorithm for exon-intron prediction.

The Human Genome

- You are 23 chromosome pairs, over 3 billion nucleotides of ATCG, & ~23,000 protein-coding genes!
- The Human genome sequence allows biologists to study DNA with *computers*.
- What do we find in the genome?



It's got "STUFF."

- Protein-coding genes.
- Repetitive & transposable elements.
- Functional non-coding RNAs.
- Transcription factor binding sites.
- Splicing enhancers/silencers.
- *And much, much more!*

Genome ≠ Random.

- Non-randomness is seen in DNA at 3 scales:
  - Short < 30 nucleotides
  - Middle 30 to 10,000 nucleotides
  - Long 300,000+ base pairs

- Includes inhomogeneous regions, non-uniform frequency distributions, & mosaic structures.

Bias of Genomic “Words” (short)

- Beyond 5 or 6 nt distance, base choice is essentially uncorrelated (though coding regions are a special case).
- “Genomic Signatures” (dinucleotide biases) exist in species (Karlin *et al* 98).
- Codon structure, we have *RNY* periodicity (Shepherd 81) and correlations at multiples of 3 (see Guigó rev).
- “Pyknons” are longers words at unexpected frequencies (Rigoutsos *et al*. 2006).

Long-range Mosaic Structure

- The human genome is a mosaic of “isochores.”
- Isochores have the same G+C composition down to 300,000 bp windows.

Human Chromosome 21 Isochore Map

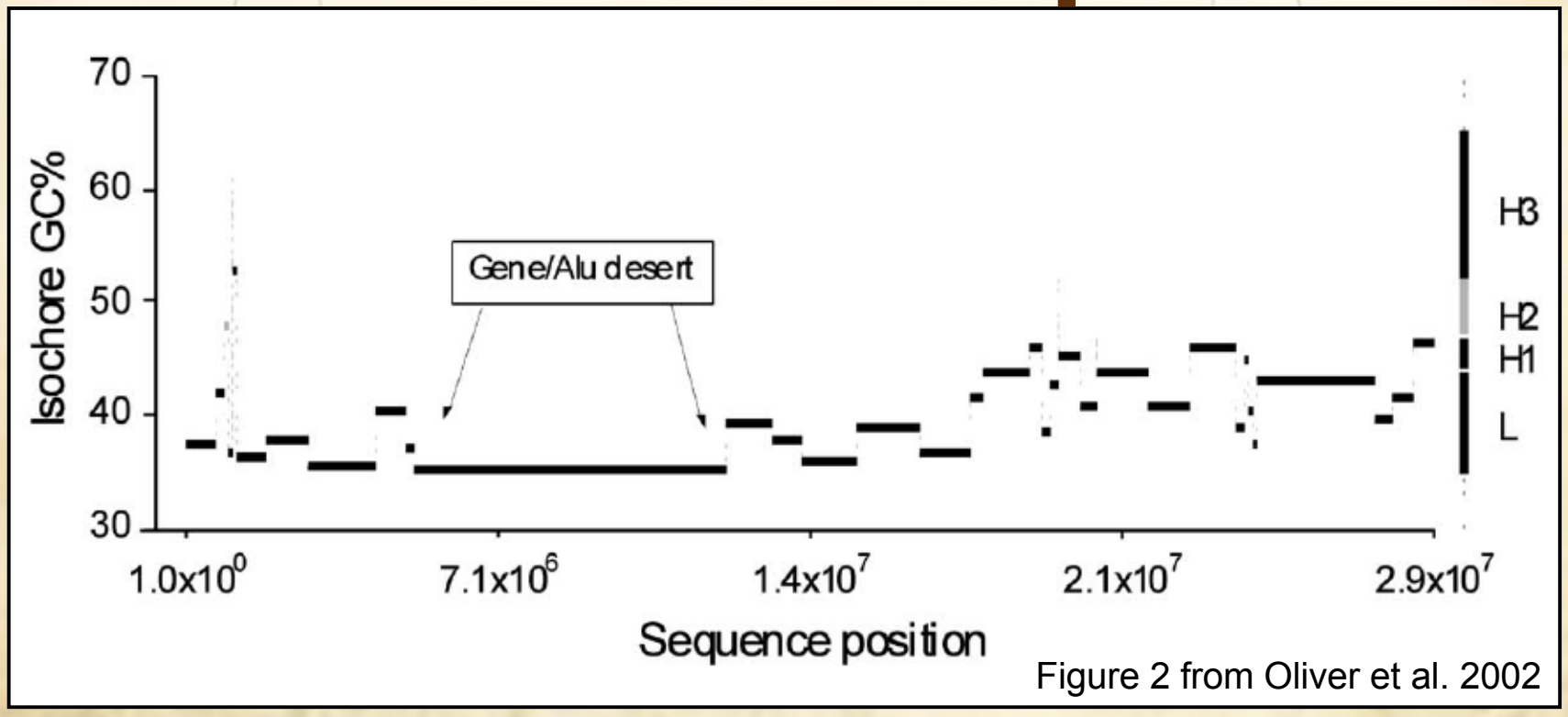
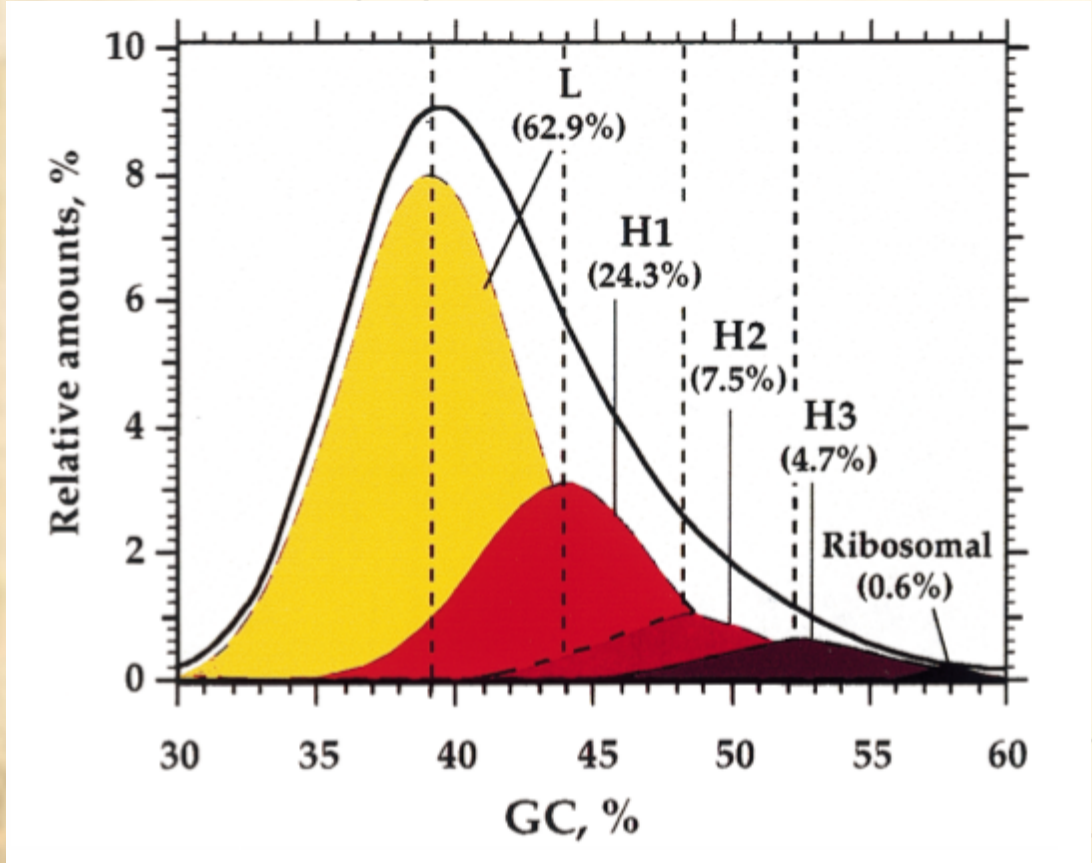


Figure 2 from Oliver et al. 2002

Human Isochore Families



- Isochores divided by <u>G+C-content.</u>
- H associated with gene density.

Figure 2.
S Zoubak, O Clay, and G Bernardi. The gene distribution of the human genome. Gene, 174(1):95-102, 1996 Sep 26.

What happens in the middle?

- The middle-range (30 to 10,000 bp) is *inhomogeneous* in terms of sequence composition.

"Inhomo-genie" what?

- Take pizza.
  - Sauce
  - Pepperoni
  - Crust
  - Cheese.

- This is an *inhomogeneous* structure.

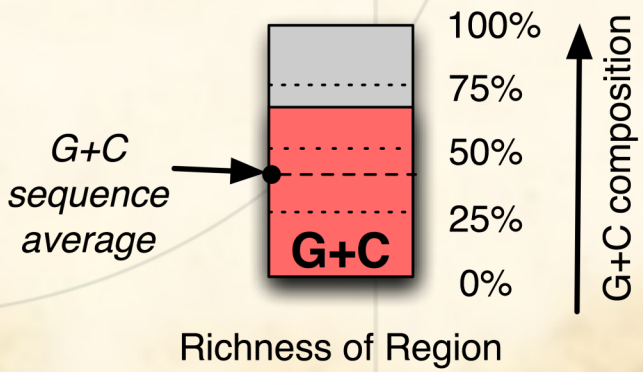
- Put pizza in a blender (add water) and you get a *homogeneous* pizza shake.

Middle-range Inhomogeneity in the Human Genome

DNA Strand

...ACGGCTGCGGC..GCGGCCGCTACGACGTGACTCGTA...

"GC-rich" MRI region
30+ nucleotides long



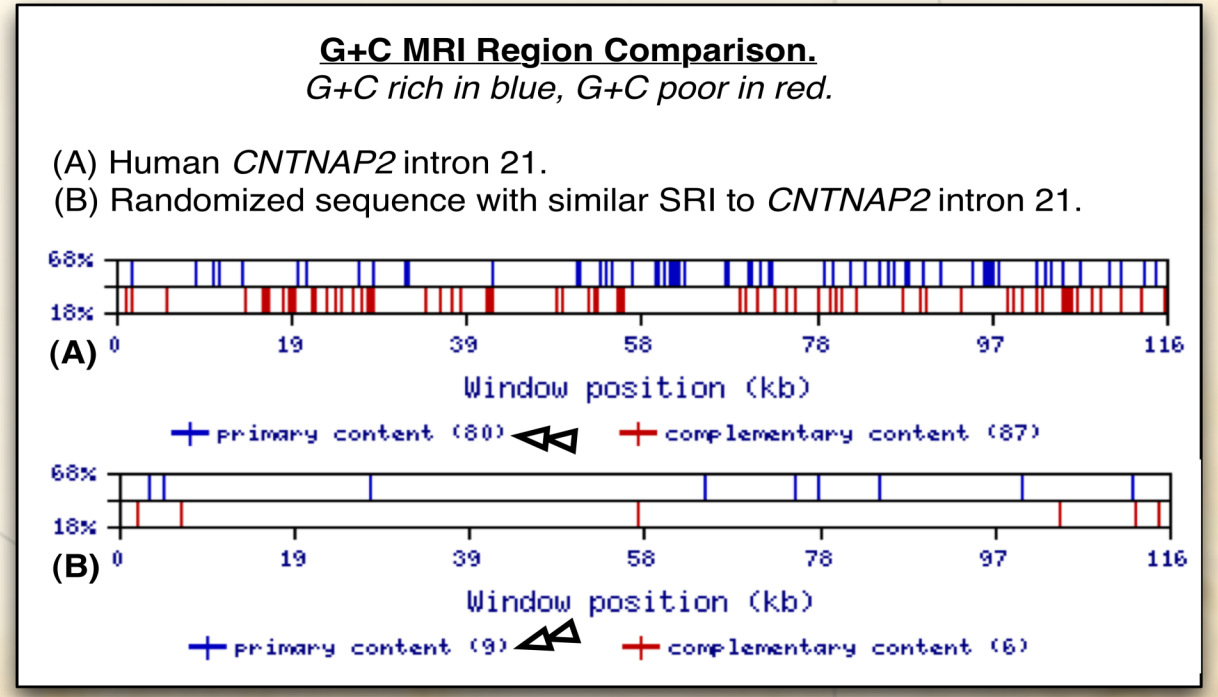
I got a website for that.

- Genomic MRI web-site: *http://bpg.utoledo.edu/gmri/*

Graph of Mid-range Inhomogeneity (GC-rich/poor)

A. large human intron

B. random sequence

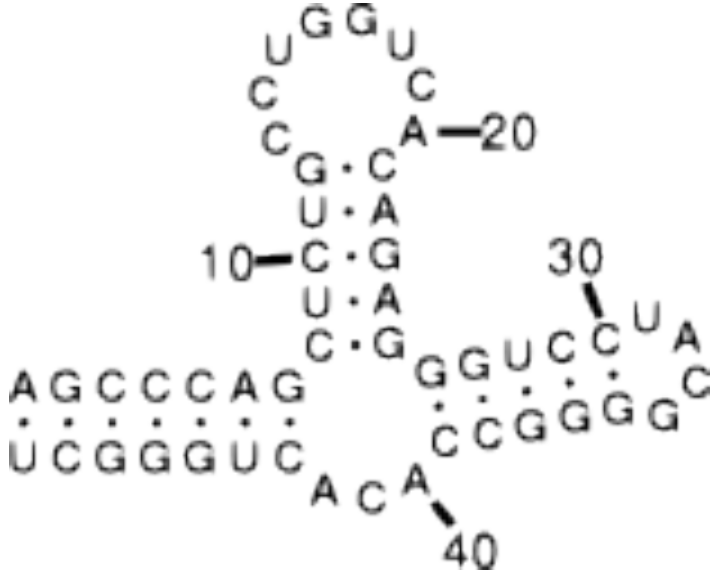


But I like the genomic "word" idea better!

- Human genome has ~3.4 billion base pairs.
- For 16-mers:
  - 4^{16} ~ 4.3 billion possible combinations
- Unless some biological feature, expect longer and longer words to be unique.
- Sequence composition is a fuzzy way to measure nucleotides bias at the mid-range.

RNA secondary structure

In 2008. Middle-range inhomogeneity is associated with predicted strong, local RNA secondary structures.



Example RNA secondary structure with a predicted mfe of ***-27.2 kcal/mol***

Mid-range Inhomogeneity is Everywhere

- In all human genomic regions: 5'-UTRs, 3'-UTRs, introns, intergenic regions, coding sequences

- In many species: human, mouse, cow, dog, rat, fly, etc.

The Maintenance of MRI

- Can look at MRI regions within the whole human genome and see what mutations do to them.

MRI within human populations.

- Mutation is *always* happening.
- Theoretically, "better ones" (less bad?) get saved.
- Over time some of these changes become *fixed* in a population.
- So how many fixed mutations "want to preserve" middle-range inhomogeneity?

Single Nucleotide Polymorphisms (SNP)

SNP example: A ⇨ G

```
..GCATTGCATGAATACCCGCTA..   Chimpanzee        } Ancestral allele: A
..GCATTGCATGAATACCCGCTA..   38% of Humans      } Polymorphic alleles: A/G
..GCATTGCATGGATACCCGCTA..   62% of Humans      }
```

-10 +10

The 20 base pairs flanking the site must have 90% identity for human & chimp.

- **G** is the *mutant* allele.
- **G** is between 20% and 80%.
- Thus, **G** is a medium SNP.

Classifying the Mutant Allele

The mutant SNP allele(s) classified according to their frequency within the human population:

Rare SNP	< 3%
Minor SNP	3 to 20%
Medium SNP	20 to 80%
Major SNP	> 80%

Processing changes in the Human Genome

- 3.9 million SNPs from dbSNP
- 18.8 million fixed point substitutions (human-chimp-macaque)
- 6.9 million bp of insertions/deletions

Using “Extremely Easy Math”

- Consider a G+T-rich region is 70% G+T.
- How does that percentage change over time because of mutations?

$$Q_X = \frac{P_X \cdot N_{nonX \to X}}{N_{X \to nonX} \cdot (1 - P_X) + N_{nonX \to X} \cdot P_X}$$

Just one example.

Equilibrium for X-percentage computed from each substitution rate						
Type of region	Observed X-percentage	rare SNPs	minor SNPs	Medium SNPs	major SNPs	fixed substit.
GT-rich	69.8%	56.9	60.7	64.6	70.8	70.4
nonGT-rich	30.1%	41.7	37.7	36.5	29.2	30.1
GT-average	50.0%	49.9	50.0	50.0	50.0	50.0

Conclusion: Maintenance of Middle-range Inhomogeneity

- MRI regions have similar levels of *new* mutations as control genomic sequences.
- *New* mutations quickly erode MRI regions by bringing their nucleotide composition toward genome-average levels.
- Mutations that favor the maintenance of MRI tend to spread throughout the entire human population.
- *Insertions/deletions* tend to maintain MRI features but have a smaller impact than substitutions.

Time to Apply MRI

- Middle-range inhomogeneity is ubiquitous as well as important to the genome.
- Can we exploit MRI for the prediction protein-coding genes?

Markov Chains

- Markov models are the basis for many gene prediction programs such as GeneMark.
- We base our approach on Markov chain algorithms.

Welcome to class!

- Today you will become human Markov models.
- Markov models can *generate* & *discriminate* sequence data.
- Ready to begin?

Fill in the blank...

i. th[ ]

ii. gol[ ]

iii. fluff[ ]

iv. dinosau[ ]

q r t e d u o s f m a y

Fill in the blank...

q r t e d u o s f m a y

i. th**e**

ii. gol

iii. fluff

iv. dinosau

Fill in the blank...

q r t e d u o s f m a y

i. the

ii. gold

iii. fluff

iv. dinosau

Fill in the blank...

i. the

ii. gold

iii. fluffy

iv. dinosau

q r t e d u o s f m a y

Fill in the blank...

q r t e d u o s f m a y

i. the
ii. gold
iii. fluffy
iv. dinosaur

Markov chain fundamentals

- The number of “letters” remembered by the Markov chain are known as its order.

- Markov chains can *generate* the next letter based on the model frequencies.

Markov chain fundamentals

- Longer words like “dinosaur” were easier to guess than shorter ones like “gold” (could have been “golf”).

- Larger order Markov chains generally do *better* prediction.

Markov chains for Prediction

- Earlier you became human Markov models to generate words using your knowledge of *English.*

- *What if I <u>only</u> gave you a sequence of characters & wanted to know which language it was???*

Español or English?

tsnottearittheysaidtoon
eanothertetsdecidebylo
twhowillgetitthishappen
edthatthescripturemigh
tbefulfilledthatsaidthey
dividedmyclothesamon
gthemandcastlotsformy
garmentsothisiswhatth
esoldi

?
?
?

idamossedijeronunosa
otrosechemossuertesp
araveraquienletocayasi
lohicieronlossoldadose
stosucedioparaquesec
umplieralaescrituraque
diceserepartieronentre
ellosmimantoysobremir
opaecharonsuer

Español or English?

tsnottearit**they**saidtoon
eanothertetsdecidebylo
twhowillgetitthishappen
edthatthescripturemigh
tbefulfilledthatsaid**they**
divided**my**clothesamon
gthemandcastlotsformy
garmentso**this**iswhatth
esoldi

?
?
?

idamossedijeronunosa
otrosechemossuertesp
araveraquienletocayasi
lohicieron**los**soldadose
stosucedioparaquesec
umplier**a****la**escrituraque
diceserepartieronentre
ellosmimantoysobre**mi**
ropaecharonsuer

Doing Prediction

- Frequent patterns (words) help you see the *language* or model classification.
- It's difficult to make sense of the sentences without knowing where to start reading.

Help with Reading Frame

tsnottearittheysaidtooneanother**L**ets
decidebylotwhowillgetit**T**hishappene
dthatthescripturemightbefulfilledthat
said**T**heydividedmyclothesamongth
emandcastlotsformygarment**S**othisi
swhatthesoldi

Doing Prediction

- *Inhomogeneous* Markov models can “see” multiple reading frames.
  - Helps detect coding sequences.
  - More accurate.

- *Homogeneous* Markov chains don’t care.

Training for the Unknown

- Suppose you **don't know** either language.
- How do you do prediction without learning the meaning of every word in each language?

...beschlossensiediesesuntergewandwollen…

Training a Model

- You'd read lots of books in each language & learn the frequent words!

Training Markov chains

- Our algorithm gets to read **12 million** nucleotides of exons and introns each.
- **3 million** are used to test prediction.

Training Markov chains

- Our tests based on whole intron and exon sequences.
- 72,000 training & 18,000 test EXONS.
- 2,500 training & 600 test INTRONs.

Moving toward a new Approach

- Remember that longer and longer words will be unique.
- We've been using **short** "words" for prediction, but the mid-range patterns are also non-random!

Enter: Binary-abstracted Markov models

- Mid-range nucleotide sequences need more “books” of information than the human genome can provide.

- We *reduce* sequence information to do mid-range Markov model analysis.

Abstraction is Sometimes Good Enough

- Why do we say “doctors make good money” instead of “orthopedic surgeons earn significantly above the mean wage”?
- Why would I say “I go to school” when I go to the “University of Toledo: Health Science Campus”?

Our Abstraction Process for Nucleotide Sequences

"G" or not "G", that is the question:

Binary-abstraction process.

AGCTGTAATGTG..

↓↓↓↓↓↓↓↓↓↓↓↓

010010000101..

The abstraction rule.

1 if G
0 otherwise

Markov Chain
Training/Testing

Our Abstraction Process for Nucleotide Sequences

"G" or not "G", that is the question:

Binary-abstraction process.

AGCTGTAATGTG..

0100100000101..

The abstraction rule.

1 if G
0 otherwise

Markov Chain
Training/Testing

Our Abstraction Process for Nucleotide Sequences

"G" or not "G", that is the question:

Binary-abstraction process.

AGCTGTAATGTG..

0100100000101..

The abstraction rule.

1 if G
0 otherwise

Markov Chain
Training/Testing

Our Abstraction Process for Nucleotide Sequences

"G" or not "G", that is the question:

Binary-abstraction process.

AGCTGTAATGTG..

01001000101..

The abstraction rule.

1 if G
0 otherwise

Markov Chain
Training/Testing

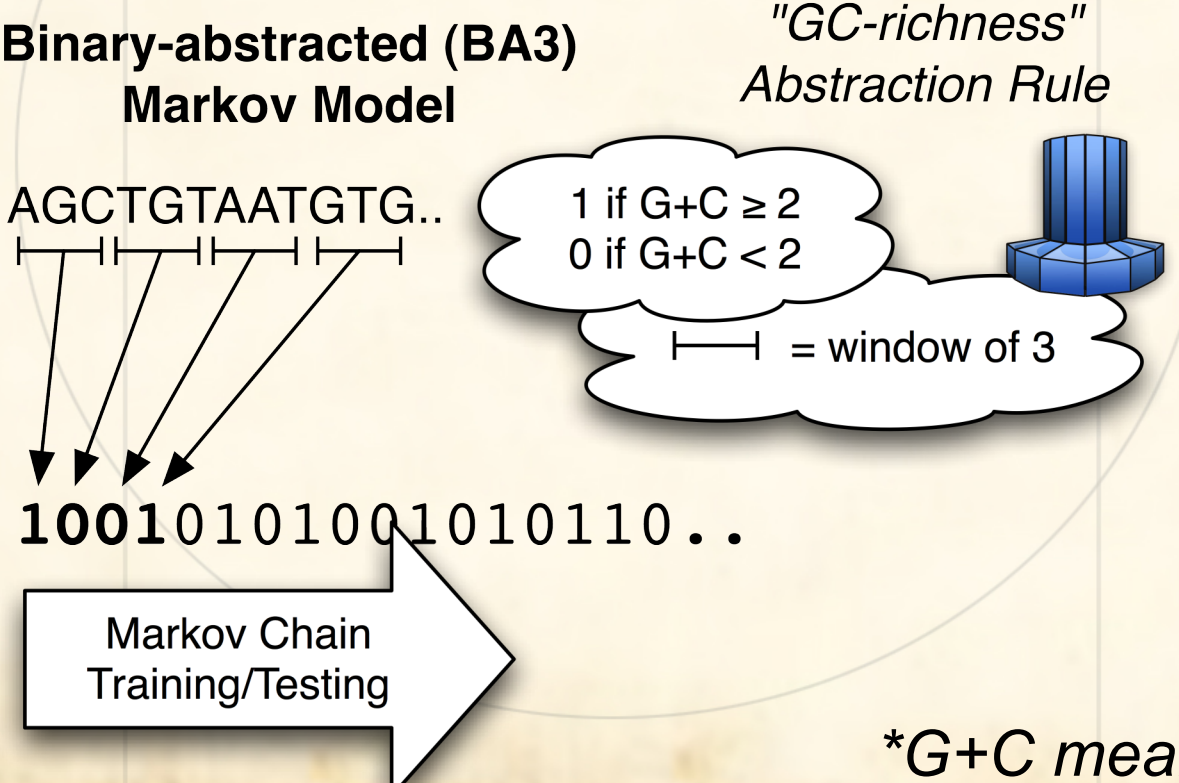
Abstraction Rule

- Abstraction rules indicate how to reduce nucleotide information into a binary code.
- Abstraction rules depend on the nucleotide word length.

Nucleotides Words of Length 3

Binary-abstracted (BA3) Markov Model

"GC-richness" Abstraction Rule



**G+C means "G or C"*

How many ways can I reduce nucleotide information?

Word Length	# Words	# Abstraction Rules
1	4	16
2	16	65,536
3	**64**	**1.84×10^{19}**
4	**256**	**1.16×10^{77}**

How do I get the best abstraction rules?

- Abstraction lengths of 1 & 2 are okay.
- For 3 & 4, need MORE POWER!

The Ohio Supercomputer Center & the Glenn Cluster

- 4,212 Opteron CPUs.
- 75 trillion floating point operations per second.
- We typically used only 512 computer cores.

Optimization of Abstraction Rules for “4-mers”

- Tested over 324 million 4-mer abstraction rules.
- Took about 11 days of supercomputer time in total.
- Would have taken over 3 and half years on a single core desktop computer.

The best individual results.

Best Abstraction Rule for Words:	*Exon Accuracy*	*Intron Accuracy*
Length 1	77%	79%
Length 2	75%	88%
Length 3	77%	93%
Length 4	80%	92%

**Accuracy is the percentage of correct predictions.*

Other Ideas

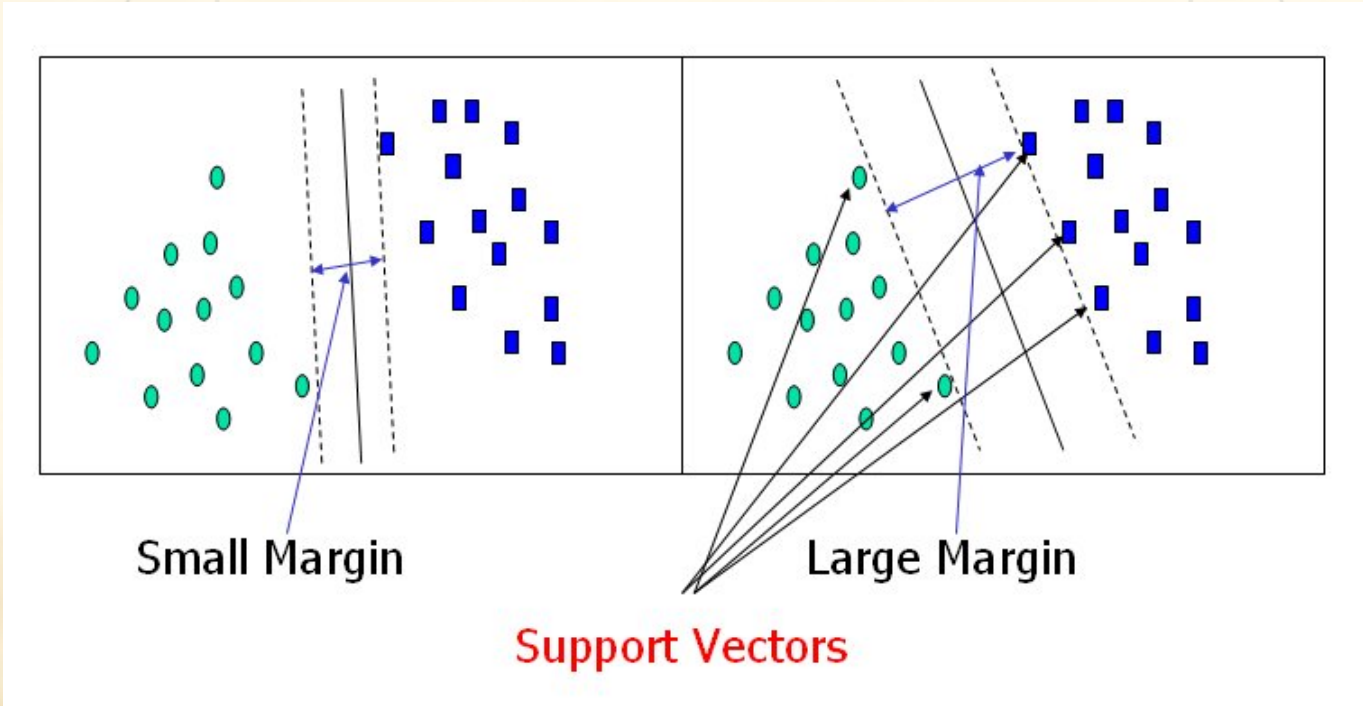
- Abstraction rules based on frame.
- Abstraction rules based on repetitive sequences.
- Abstraction rules based on splicing signals.

Machine-learning to Optimize Prediction

- Support Vector Machines can learn to draw better boundary lines between two classes of data.

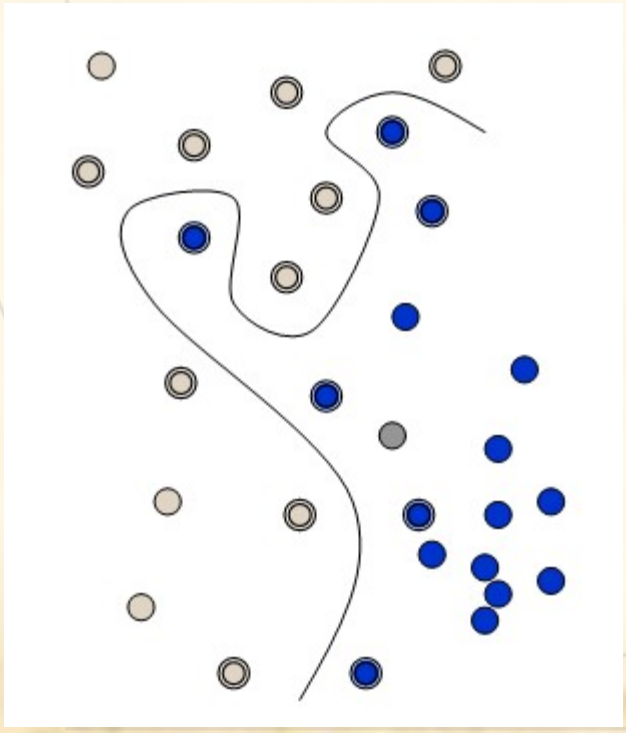
- Multiple binary-abstracted Markov model predictions can be used as input.

Support Vector Machines



From www.dtreg.com/svm.htm

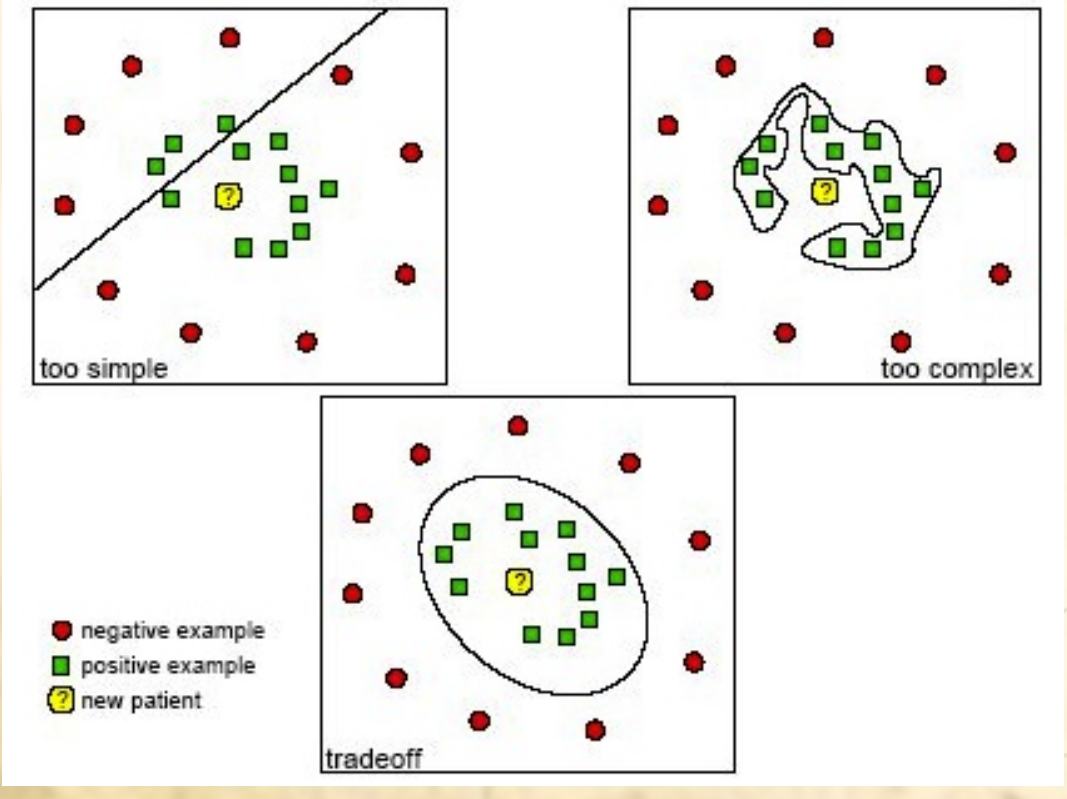
Support Vector Machines



Non-linear division boundary.

From www.dtreg.com/svm.htm

Over & under fitting.



From www.dtreg.com/svm.htm

Model Optimization

	Original Values			***SVM Optimized Values***		
Abstraction Rule	*Exon Acc.*	*Intron Acc.*	M-value	*Exon Acc.*	*Intron Acc.*	M-value
Markov Model 6	89%	83%	0.854	94%	80%	0.855
G-map (BA1)	77%	79%	0.779	94%	72%	0.801
BA2 Best	75%	88%	0.806	94%	81%	0.860
BA3 Best	77%	93%	0.831	94%	86%	0.893
BA4 Best	80%	92%	0.849	95%	84%	0.883
A priori 3	76%	69%	0.726	93%	75%	0.817
SP Top 24 Pos	73%	86%	0.782	94%	76%	0.822
GT-rich	65%	83%	0.725	94%	70%	0.781
Duplication	77%	86%	0.807	95%	76%	0.829
Purine-pyrimidine	79%	65%	0.707	93%	69%	0.777

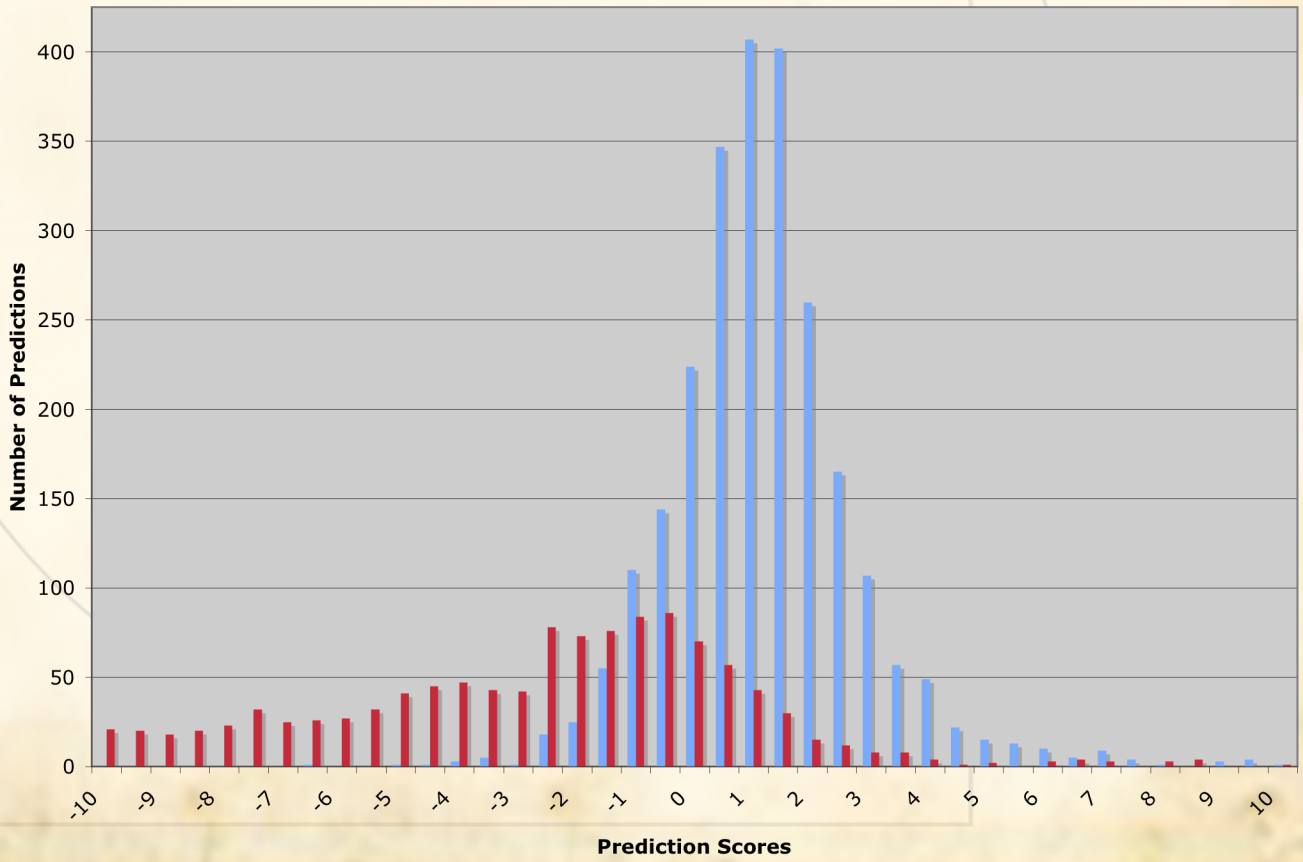
**M-value combines the total accuracy of predictions.*

Optimization Consequence

- Lose fewer points of intron accuracy to gain many points of exon accuracy.
- Exon accuracy emphasis may be due to the variation in the prediction data.

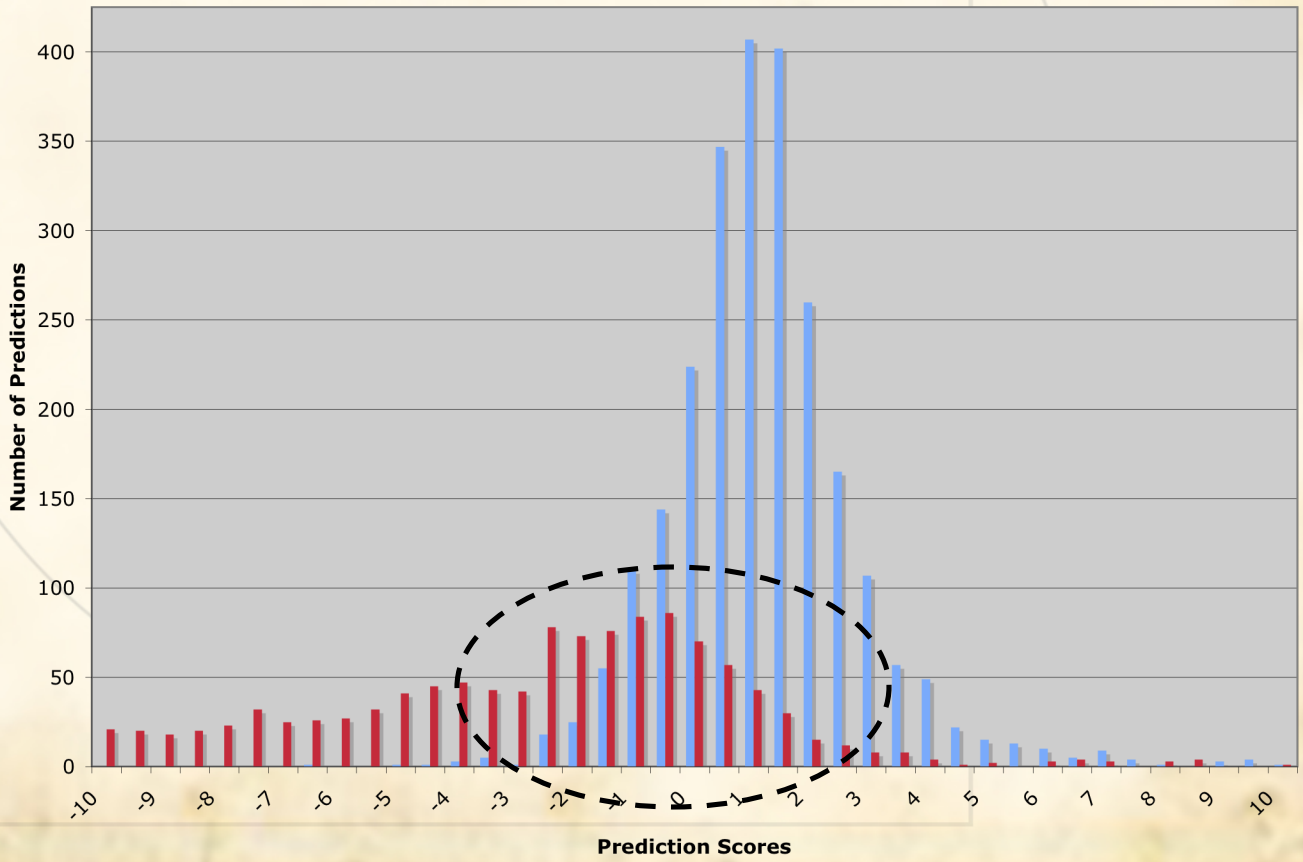
Zoomed Histogram of Prediction Scores

- Exons **blue**; Introns **red**.
- Only scores for -10 to +10 shown.
- Best abstraction rule for triplets.



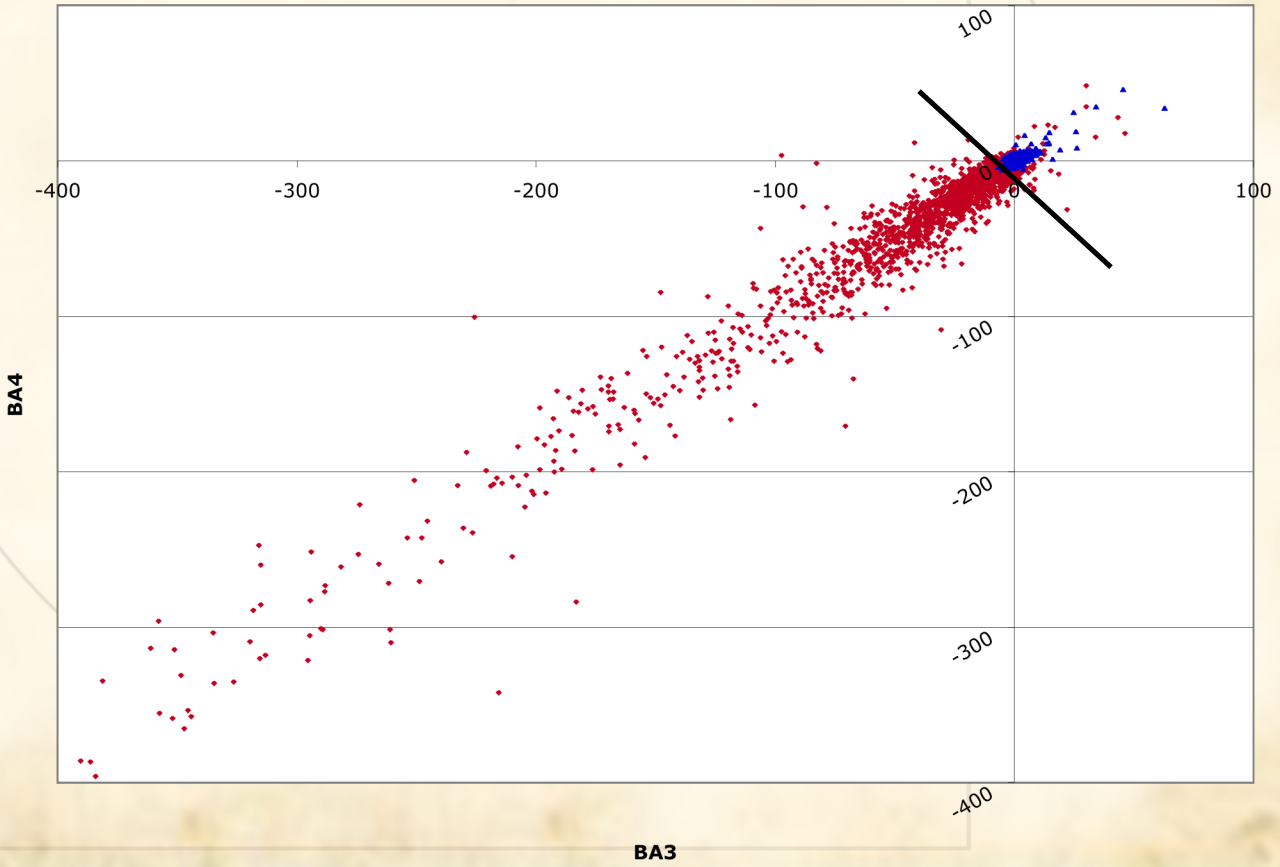
Zoomed Histogram of Prediction Scores

- Exons **blue**; Introns **red**.
- Overlap of introns around the mean histogram scores for exons.
- Intron mean further out.



Zoomed Plot of 2 Models

- Exons **blue**; Introns **red**.
- Best abstraction rule for 4-mers versus 3-mers.



Combining Different Models

- From 10 different abstraction rules/models we chose combinations of 1, 2, 3, *etc. (K)*

- Found best combination for each group that differed no more than $M = 0.003$ between the test & validation sets.

Model Combination Results

- High accuracy for K models > 2.
- Little accuracy difference between test & validation sets.

		Validation Set			Test Set		
$\binom{n}{k}$	Best Combo	%ex	%in	M-val	%ex	%in	M-val
$K = 1$	BA3 Best (BA3)	93.9	86.6	0.896	93.9	86.2	0.893
$K = 2$	BA3 + BA4	94.7	89.2	0.915	94.5	89.8	0.918
$K = 3$	BA1 + BA3 + AP3	94.8	91.9	0.932	94.6	92.2	0.933
$K = 4$	MM6 + BA3 + BA4 + AP3	97.1	92.7	0.944	96.6	93.3	0.947
$K = 5$	MM6 + BA3 + BA4 + AP3 + DUP	96.8	93.6	0.950	96.5	94.0	0.951
$K = 6$	MM6 + BA2 + BA4 + AP3 + POS + DUP	96.6	93.8	0.950	96.3	94.4	0.952
$K = 7$	MM6 + BA2 + BA3 + BA4 + AP3 + POS + DUP	96.9	94.3	0.954	96.7	94.7	0.956

Model Combination Results

- K = 5 models about 95% accurate!
- Over-fitting due to abstraction rules not large.

		Validation Set			Test Set		
$\binom{n}{k}$	Best Combo	%ex	%in	M-val	%ex	%in	M-val
$K = 1$	BA3 Best (BA3)	93.9	86.6	0.896	93.9	86.2	0.893
$K = 2$	BA3 + BA4	94.7	89.2	0.915	94.5	89.8	0.918
$K = 3$	BA1 + BA3 + AP3	94.8	91.9	0.932	94.6	92.2	0.933
$K = 4$	MM6 + BA3 + BA4 + AP3	97.1	92.7	0.944	96.6	93.3	0.947
$K = 5$	MM6 + BA3 + BA4 + AP3 + DUP	96.8	93.6	0.950	96.5	94.0	0.951
$K = 6$	MM6 + BA2 + BA4 + AP3 + POS + DUP	96.6	93.8	0.950	96.3	94.4	0.952
$K = 7$	MM6 + BA2 + BA3 + BA4 + AP3 + POS + DUP	96.9	94.3	0.954	96.7	94.7	0.956

Not just for Coding Exons

- Untranslated regions (UTR) are also spliced, especially in the 5-prime UTR.
- Can we train on coding sequences and predict 5-prime UTR exons?

Jensen-Shannon Divergence

- Divergence measures the difference in the 5-mer frequency distributions.

$D(F_i, F_j)$	3′ EX	5′ EX	CDS	INTER	3′ IN	5′ IN	CDS IN
3′ UTR Exon	-	0.231	0.103	0.012	0.008	0.007	0.008
5′ UTR Exon	0.231	-	0.097	0.208	0.274	0.256	0.281
CDS Exon	0.103	0.097	-	0.086	0.129	0.125	0.139
Intergenic	0.012	0.208	0.086	-	0.010	0.010	0.014
3′ UTR Intron	0.008	0.274	0.129	0.010	-	0.002	0.001
5′ UTR Intron	0.007	0.256	0.125	0.010	0.002	-	0.001
CDS Intron	0.008	0.281	0.139	0.014	0.001	0.001	-

Jensen-Shannon Divergence

- 5-prime UTR Exons extremely different from introns.

$D(F_i, F_j)$	3′ EX	5′ EX	CDS	INTER	3′ IN	5′ IN	CDS IN
3′ UTR Exon	-	0.231	0.103	0.012	0.008	0.007	0.008
5′ UTR Exon	0.231	-	0.097	0.208	0.274	0.256	0.281
CDS Exon	0.103	0.097	-	0.086	0.129	0.125	0.139
Intergenic	0.012	0.208	0.086	-	0.010	0.010	0.014
3′ UTR Intron	0.008	0.274	0.129	0.010	-	0.002	0.001
5′ UTR Intron	0.007	0.256	0.125	0.010	0.002	-	0.001
CDS Intron	0.008	0.281	0.139	0.014	0.001	0.001	-

5-prime UTR Exon Classification

- **Trained** with coding exons & introns.
- **Tested** on 5-prime UTR exons and introns.

Trained on CDS exons & all introns, tested on 5′ UTR exons & introns.			
Abstraction Rule / Model	Exon Accuracy	Intron Accuracy	M-value
GC-richness	82%	61%	0.700
GT-richness	52	88	0.650
AG-richness	59	73	0.655
BA3 Best	66	93	0.757
A priori 3	85	68	0.749
SP version 2008	76	74	0.748
SP '08 Optimized (HD4)	76	79	0.775
SP version 2009	78	76	0.770
SP '09 Optimized (HD4)	77	81	0.787
SP '09 Top 24 Positive	76	76	0.758
SP '09 Positive HD4	72	82	0.766
SP '09 Top 24 Negative	82	73	0.773
SP '09 Negative HD4	76	84	0.794

5-prime UTR Exon Classification

- Abstraction rules based on nucleotide richness.
- Abstraction rules based on splicing signals.

Trained on CDS exons & all introns, tested on 5′ UTR exons & introns.			
Abstraction Rule / Model	Exon Accuracy	Intron Accuracy	M-value
GC-richness	82%	61%	0.700
GT-richness	52	88	0.650
AG-richness	59	73	0.655
BA3 Best	66	93	0.757
A priori 3	85	68	0.749
SP version 2008	76	74	0.748
SP '08 Optimized (HD4)	76	79	0.775
SP version 2009	78	76	0.770
SP '09 Optimized (HD4)	77	81	0.787
SP '09 Top 24 Positive	76	76	0.758
SP '09 Positive HD4	72	82	0.766
SP '09 Top 24 Negative	82	73	0.773
SP '09 Negative HD4	76	84	0.794

Observations with UTR Classification

- 3-prime UTRs are difficult to predict based on composition.
- 5-prime UTR data may be too small to use for training.
- Accuracy under SVM was 87% exon accuracy and 91% intron accuracy for 4 models (BA1, BA2, BA3, SP).

Summary of Achievements

- Described how mid-range genomic signals are maintained in the human genome.
- Introduced and tested a new algorithm for genomic sequence classification.
- Optimized the method using supercomputer & machine-learning technology.
- Achieved better results than the traditional homogeneous Markov model.
- Adapted our approach for 5-prime UTR data.

Final Remarks

- Not a sequence parse, but can further develop for a full gene-finding program.
- May be able to utilize abstraction methodology for other classification: alternative splicing, nucleosome filing positions, *etc*.

Thank you for your attention.

Questions?

¿Preguntas?

вопросы?

问题?

Fragen?

質問か。

Acknowledgements

- **My Lab:** Alexei Fedorov, Andrew McSweeny, Ashwin Prakash, Maryam Nabiyouni, and David Rearick
- **My Committee:** Robert Blumenthal, Robert Trumbly, Sadik Khuder, and John Gray.
- **The Ohio Supercomputer Center.**
- **Collaborators & Colleagues:** Mark Borodovsky, Craig Zirbel, Gursel Serpen, Jie He, Peter Bazeley, & Jason Bechtel.
- **My loved ones:** thanks to Jesus Christ, my family, & my friends for all their help & support.