

# Interpolated Markov Models for Gene Finding

BMI/CS 776

[www.biostat.wisc.edu/~craven/776.html](http://www.biostat.wisc.edu/~craven/776.html)

Mark Craven

[craven@biostat.wisc.edu](mailto:craven@biostat.wisc.edu)

February 2002

## Announcements

- HW #1 out; due March 11
- class accounts ready
  - [quasar-1.biostat.wisc.edu](mailto:quasar-1.biostat.wisc.edu), [quasar-2.biostat.wisc.edu](mailto:quasar-2.biostat.wisc.edu)
- class mailing list ready
  - [bmi776@biostat.wisc.edu](mailto:bmi776@biostat.wisc.edu)
  - please check mail regularly and frequently, or forward it to wherever you can do this most easily
- reading for next week
  - Bailey & Elkan, *The Value of Prior Knowledge in Discovering Motifs with MEME* (on-line)
  - Lawrence et al., *Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment* (handed out in class)
- talk tomorrow
  - Bioinformatics Tools to Study Sequence Evolution: Examples from HIV*
  - Keith Crandall, Dept. of Zoology, BYU
  - 10am, Thursday 2/28
  - Biotech Center Auditorium (425 Henry Mall)

## Approaches to Finding Genes

- **search by sequence similarity:** find genes by looking for matches to sequences that are known to be related to genes
- **search by signal:** find genes by identifying the sequence *signals* involved in gene expression
- **search by content:** find genes by statistical properties that distinguish protein-coding DNA from non-coding DNA
- **combined:** state-of-the-art systems for gene finding combine these strategies

## Gene Finding: Search by Content

- encoding a protein affects the statistical properties of a DNA sequence
  - some amino acids are used more frequently than others (Leu more popular than Trp)
  - different numbers of codons for different amino acids (Leu has 6, Trp has 1)
  - for a given amino acid, usually one codon is used more frequently than others
    - this is termed *codon preference*
    - these preferences vary by species

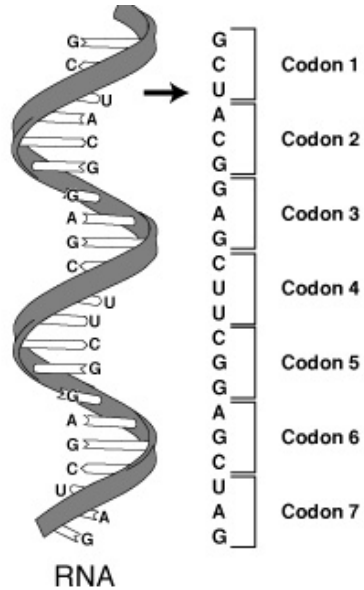
## Codon Preference in E. Coli

AA	codon	/1000
Gly	GGG	1.89
Gly	GGA	0.44
Gly	GGU	52.99
Gly	GGC	34.55
Glu	GAG	15.68
Glu	GAA	57.20
Asp	GAU	21.63
Asp	GAC	43.26

## Search by Content

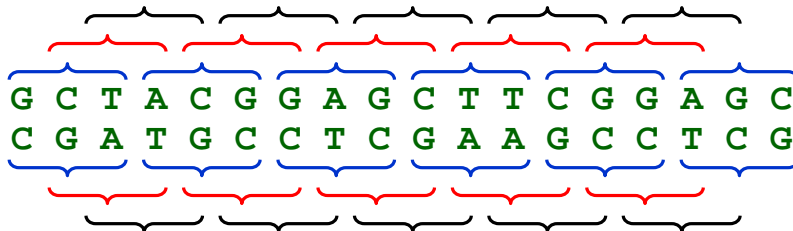
- common way to search by content
  - build Markov models of coding & noncoding regions
  - apply models to ORFs or fixed-sized windows of sequence
- GeneMark [Borodovsky et al.]
  - popular system for identifying genes in bacterial genomes
  - uses 5<sup>th</sup> order inhomogenous Markov chain models

## Reading Frames



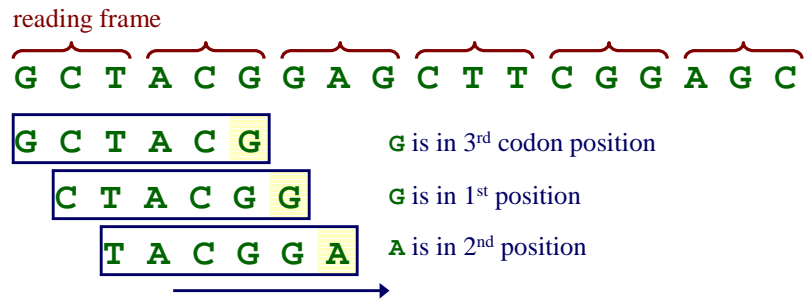
## Reading Frames

- a given sequence may encode a protein in any of the six reading frames

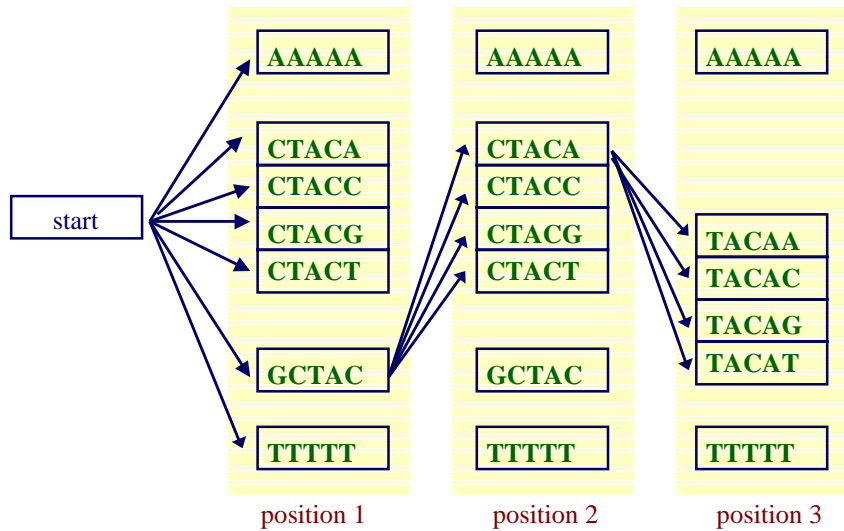


# Markov Models & Reading Frames

- consider modeling a given coding sequence
- for each “word” we evaluate, we’ll want to consider its position with respect to the reading frame we’re assuming



## A Fifth Order Inhomogenous Markov Chain



## Selecting the Order of a Markov Chain Model

- higher order models remember more “history”
- additional history can have predictive value
- example:
  - predict the next word in this sentence fragment  
“...finish \_\_” (up, it, first, last, ...?)
  - now predict it given more history  
“Nice guys finish \_\_”

## Selecting the Order of a Markov Chain Model

- but the number of parameters we need to estimate grows exponentially with the order
  - for modeling DNA we need  $O(4^{n+1})$  parameters for an  $n$ th order model
- the higher the order, the less reliable we can expect our parameter estimates to be
  - estimating the parameters of a 2<sup>nd</sup> order homogenous Markov chain from the complete genome of E. Coli, we’d see each word > 72,000 times on average
  - estimating the parameters of an 8<sup>th</sup> order chain, we’d see each word ~ 5 times on average

## Interpolated Markov Models

- the IMM idea: manage this trade-off by interpolating among models of various orders
- *simple* linear interpolation:

$$\begin{aligned}\Pr_{\text{IMM}}(x_i | x_{i-1}, \dots, x_{i-n}) &= \lambda_0 \Pr(x_i) \\ &+ \lambda_1 \Pr(x_i | x_{i-1}) \\ &\dots \\ &+ \lambda_n \Pr(x_i | x_{i-1}, \dots, x_{i-n})\end{aligned}$$

- where  $\sum_i \lambda_i = 1$

## Interpolated Markov Models

- we can make the weights depend on the history
  - for a given order, we may have significantly more data to estimate some words than others
- *general* linear interpolation

$$\begin{aligned}\Pr_{\text{IMM}}(x_i | x_{i-1}, \dots, x_{i-n}) &= \lambda_0 \Pr(x_i) \\ &+ \lambda_1(x_{i-1}) \Pr(x_i | x_{i-1}) \\ &\dots \\ &+ \lambda_n(x_{i-1}, \dots, x_{i-n}) \Pr(x_i | x_{i-1}, \dots, x_{i-n})\end{aligned}$$

## The GLIMMER System

- Salzberg et al., 1998
- system for identifying genes in bacterial genomes
- uses 8<sup>th</sup> order, inhomogeneous, interpolated Markov chain models

## IMMs in GLIMMER

- how does GLIMMER determine the  $\lambda$  values?
- first, let's express the IMM probability calculation recursively

$$\Pr_{\text{IMM},n}(x_i | x_{i-1}, \dots, x_{i-n}) = \lambda_n(x_{i-1}, \dots, x_{i-n}) \Pr(x_i | x_{i-1}, \dots, x_{i-n}) + [1 - \lambda_n(x_{i-1}, \dots, x_{i-n})] \Pr_{\text{IMM},n-1}(x_i | x_{i-1}, \dots, x_{i-n+1})$$

- let  $c(x_{i-1}, \dots, x_{i-n})$  be the number of times we see the history  $x_{i-1}, \dots, x_{i-n}$  in our training set

$$\lambda_n(x_{i-1}, \dots, x_{i-n}) = 1 \text{ if } c(x_{i-1}, \dots, x_{i-n}) > 400$$

## IMMs in GLIMMER

- if we haven't seen  $x_{i-1}, \dots, x_{i-n}$  more than 400 times, then compare the counts for the following:

<i>n</i> th order history + base	<i>(n-1)</i> th order history + base
$x_{i-n}, \dots, x_{i-1}, a$	$x_{i-n+1}, \dots, x_{i-1}, a$
$x_{i-n}, \dots, x_{i-1}, c$	$x_{i-n+1}, \dots, x_{i-1}, c$
$x_{i-n}, \dots, x_{i-1}, g$	$x_{i-n+1}, \dots, x_{i-1}, g$
$x_{i-n}, \dots, x_{i-1}, t$	$x_{i-n+1}, \dots, x_{i-1}, t$

- use a statistical test ( $\chi^2$ ) to get a value  $d$  indicating our confidence that the distributions represented by the two sets of counts are different

## IMMs in GLIMMER

- putting it all together

$$\lambda_n(x_{i-1}, \dots, x_{i-n}) = \begin{cases} 1 & \text{if } c(x_{i-1}, \dots, x_{i-n}) > 400 \\ d \times \frac{c(x_{i-1}, \dots, x_{i-n})}{400} & \text{else if } d \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

where  $d \in (0,1)$

## GLIMMER Experiment

- 8<sup>th</sup> order IMM vs. 5<sup>th</sup> order Markov model
- trained on 1168 genes (ORFs really)
- tested on 1717 annotated (more or less known) genes

## Accuracy Metrics

		actual class	
		positive	negative
predicted	positive	true positives (TP)	false positives (FP)
	negative	false negatives (FN)	true negatives (TN)

$$\text{sensitivity} = \frac{\text{TP}}{\text{all pos}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{specificity} = \frac{\text{TN}}{\text{predicted pos}} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

## GLIMMER Results

	TP	FN	FP
Model	Genes found	Genes missed	Additional genes
GLIMMER IMM	1680 (97.8%)	37	209
5 <sup>th</sup> -Order Markov	1574 (91.7%)	143	104

The first column indicates how many of the 1717 annotated genes in *H.influenzae* were found by each algorithm. The 'additional genes' column shows how many extra genes, not included in the 1717 annotated entries, were called genes by each method.

$$\text{GLIMMER} \quad \text{sensitivity} = \frac{1680}{1680+37} = 0.978 \quad \text{specificity} = \frac{1680}{1680+209} = 0.889$$

$$\text{5<sup>th</sup> Order} \quad \text{sensitivity} = \frac{1574}{1574+143} = 0.917 \quad \text{specificity} = \frac{1574}{1574+104} = 0.938$$