

# *Comparative Genomics & Annotation*

## *The Foundation of Comparative Genomics*

### *The main methodological tasks of CG Annotation:*

*Protein Gene Finding*

*RNA Structure Prediction*

*Signal Finding*

---

### *Overlapping Annotations:*

*Protein Genes*

*Protein-RNA*

### *Combining Grammars*

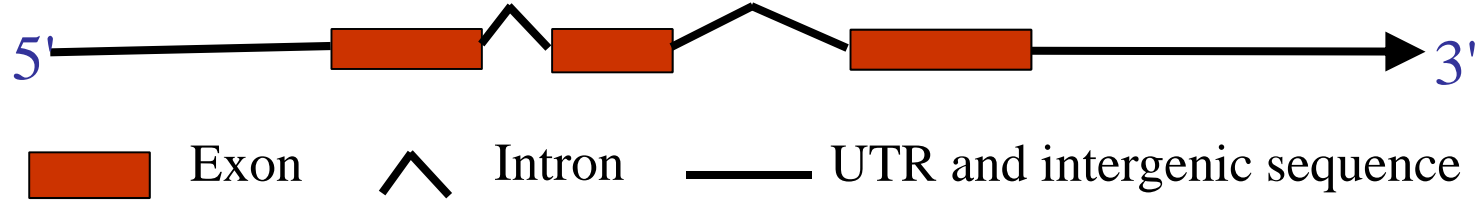
# Ab Initio Gene prediction

**Ab initio gene prediction:** prediction of the location of genes (and the amino acid sequence it encodes) given a **raw DNA sequence**.

... tttttgcagtagctcccgggccctctgttggggcctccccttctctccagggtggagtcgaggaggcggggctcggggcctccttatctctagagccggccctggctctctggcgccggggcccttagtccgggctttttgccatggggctctctgttccctctgtcgctgctgttttttttggcggccgcctaccgggagttgggagcgcgctgggacgccggactaagcggggcgaagccccaagggtagccctctcgcgccctccgggacctcagtgcccttctgggtgcgcatgagcccggagttcgtggctgtgcagccggggaagtcaagtgcagctcaattgcagcaacagctgtccccagccgcagaattccagcctccgcagcccgctgcggaaggcaaacgctcagagtcgggggtgggtgtcttaccagctgctcgacgtgagggcctggagctccctcgcgactgctcgtgacctgcgaggaacacgctgggcatctcagatcactgactacagtgaggatggggtctcccggctgggggtgaggggagggggctggaagaggtggggaaagggtagttgacagtcgctctatagggagcctgggtgggctctcaggggtccccttggctggcagcctggagcgtgattttggagcctccggctctaaagggcaggaatacactttgcgctgccacgtgacgcaggtgttcccgggtgggctacttgggtgggtgacctgagggcatggaagccgggtcatctattccgaaagcctggagcgttaccggcctggatctggccaacgtgacctgacctacgagtttgctgctggaccccgcgacttctggcagcccgtgatctgccacgcgcgctcaatctcgacggcctgggtggctcgcaacagctcggcaccattacactgatgctcgggtgagggcaccctgtaaccctggggactaggaggaagggggcagagagagttatgaccccagagggcgcacagaccaagcgtgagctccacgcgggtcgacagacctccctgtgttccgcttctaattctcgccttctgctcccagcttggagcccgcgcccacagcttggcctccgggtccatcgctgcccttgtagggatcctcctcactgtggggcgtgctgacctatgcaagtgccctagctatgaagtcccaggcgtaaaggggatgttctatgccggctgagcgagaaaaagaggaatatgaaacaatctggggaaatggccatacatggtg...

Input data

Output:

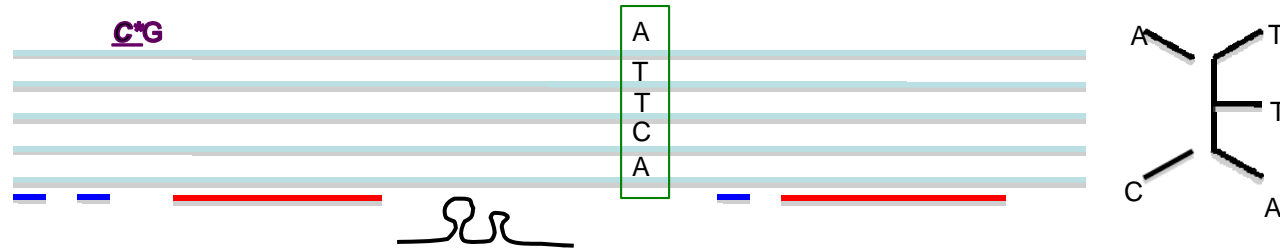


5'... tttttgcagtagctcccgggccctctgttggggcctccccttctctccagggtggagtcgaggaggcggggctcggggcctccttatctctagagccggccctggctctctggcgccggggcccttagtccgggctttttgccATGGGGTCTCTGTTCCCTCTGTCGCTGCTGTTTTTTTTTGGCGGCCGCCTACCCGGGAGTTGGGAGCGCGCTGGGACGCCGGACTAAGCGGGCGCAAAGCCCAAGGGTAGCCCTCTCGCCCTCCGGGACCTCAGTGCCCTTCTGGGTGCGCATGAGCCCAGGAGTTCGTGGCTGTGCAGCCGGGGAAGTCAGTGCAGCTCAATTGCAGCAACAGCTGTCCCCAGCCGAGAATTCAGCCTCCGCACCCCGCTGCGGCAAGGCAAGACGCTCAGAGGGCCGGGTTGGGTGTCTTACCAGCTGCTCGACGTGAGGGCCTGGAGCTCCCTCGCACTGCCTCGTGACCTGCGCAGGAAAAACACGCTGGGCCACCTCCAGGATCACCGCTACAgtaggggacaggggctcgggtcccggctgggggtgaggggagggggctggaagaggtggggaaaggtagttgacagtcgctctatagggagcggcccgacactcactcagaggtccccttgccttagAACCGCCCCACAGCGTGATTTTGGAGCCTCCGGTCTTAAAGGGCAGGAAATACACTTTGCGCTGCCACGTGACGCAGGTGTTCCCGGTGGGCTACTTGGTGGTGACCCTGAGGCATGGAAGCCGGGTATCTATTCCGAAAGCCTGGAGCGCTTACCCGGCCTGGATCTGGCCAACGTGACCTTGACCTACGAGTTTGCTGCTGGACCCCGGACTTCTGGCAGCCCCGTGATCTGCCACGCGCCTCAATCTCGACGGCCTGGTGGTCCGCAACAGCTCGGCACCCATTACACTGATGCTCGgtgagggcaccctgtaaccctggggactaggaggaagggggcagagagagttatgaccccagagggcgcacagaccaagcgtgagctccacgcgggtcgacagacctccctgtgtccgcttctaattctcgccttctgctcccagcttggagcccgcgcccacagcttggcctccgggtccatcgctgcccttgtagggatcctcctcactgtggggcgtgctgacctatgcaagtgccctagctatgaagtcccaggcgtaaaggggatgttctatgccggctgagcgagaaaaagaggaatatgaaacaatctggggaaatggccatacatggtg.... 3'

# Levels of Annotation

**“Annotation”**: Tagging regions and nucleotides with information about function, structure, knowledge, additional data,....

## Homologous Genomes



## Annotation levels

**Protein coding genes including alternative splicing**

**RNA structure**

**Regulatory signals – fast/slow, prediction of TF, binding constants,...**

**Selection Strength,...**

**Epigenomics – methylation, histone modification**

## Further complications

**Integration of levels – RNA structure of mRNA, signals in coding regions,..**

**Knowledge and annotation transfer – experimental knowledge might be present in other species**

**Evolution of Feature – regulatory signals > RNA > protein**

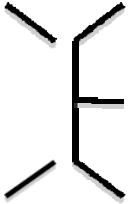
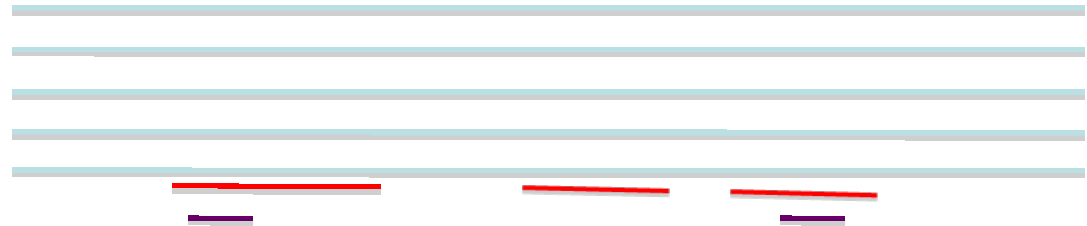
**Combining with non-homologous analysis – tests for common regulation.**

**Combining specie and population perspective**

# Observables, Hidden Variables, Evolution & Knowledge

*Observables*

$$P(X) = \pi(X)$$

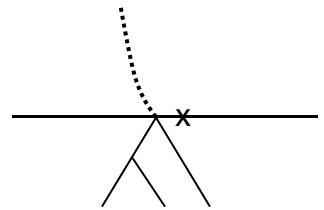


*Hidden Variable*

$$P(X) = \sum_H P(X|H)P(H) = \sum_H \pi(X|H)P(H)$$

*Evolution*

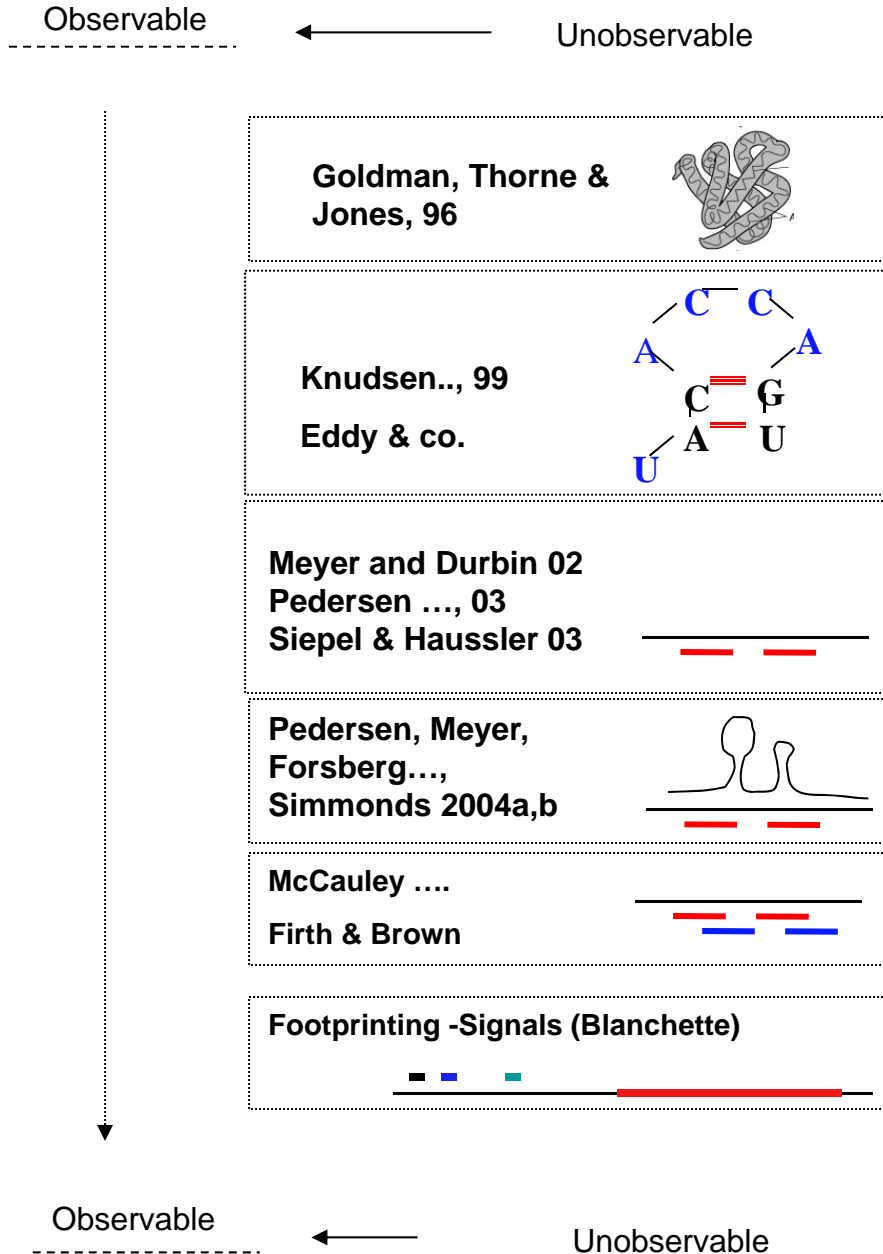
$$P(X) = \sum_H \pi(X|H)P(X_{dyna}|H)P(H)$$



*Knowledge (Constraints)*

$$P(X) = [PW]^{-1} \sum_H P(X|H)P(H)w(H) \stackrel{\text{If knowledge deterministic}}{\downarrow} [PW]^{-1} \sum_{H \cap w=1} P(X|H)P(H)$$

# Co-Modelling and Conditional Modelling



AGGTATATA**ATGCG**.....  $P_{\text{coding}}\{\text{ATG-->GTG}\}$  or  
AGCCATTTA**GTGCG**.....  $P_{\text{non-coding}}\{\text{ATG-->GTG}\}$



## • Conditional Modelling

$$P(\text{Sequence}|\text{Structure})P(\text{Structure}) =$$

$$P(\text{Structure}|\text{Sequence})P(\text{Sequence})$$

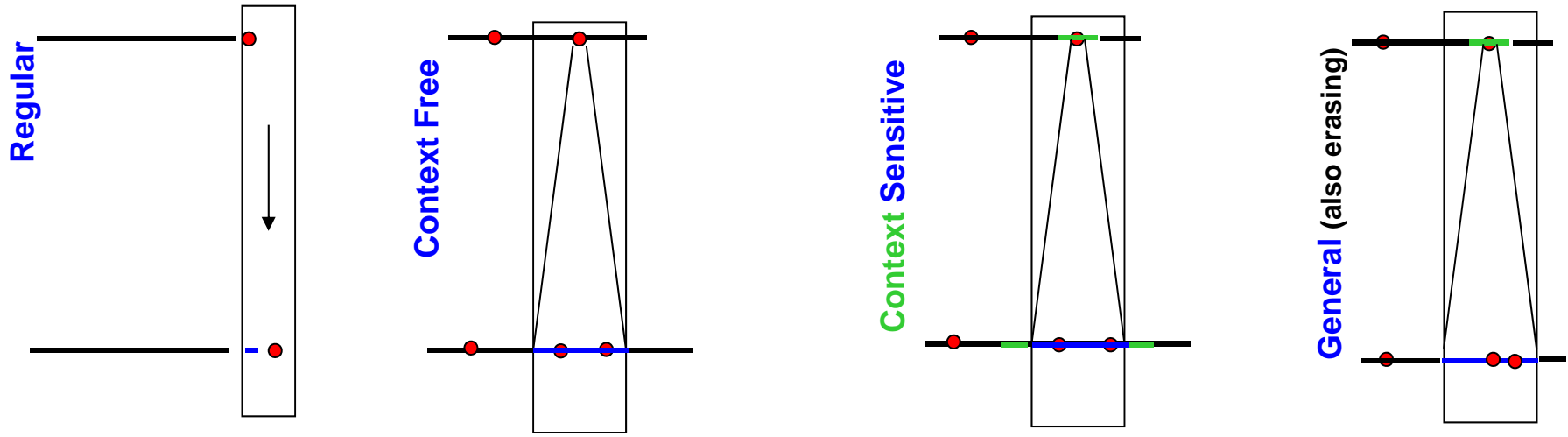
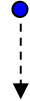
## Needs:

i.  $P(\text{Sequence}|\text{Structure})$

ii.  $P(\text{Structure})$

# Grammars: Finite Set of Rules for Generating Strings

- i. A starting symbol: • Ordinary letters: — & Variables: •
- ii. A set of substitution rules applied to variables • in the present string: —•—•—



— finished – no variables

# Simple String Generators

Variables (capital)

Letters (small)

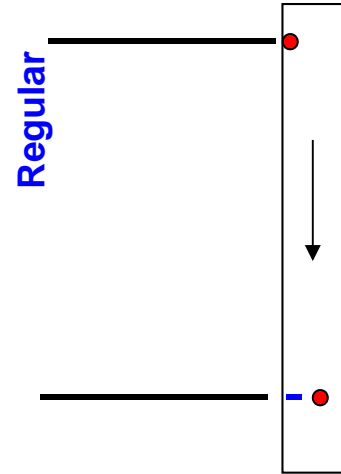
Regular Grammar:

Start with  $S$

$$S \rightarrow aT \quad bS$$
$$T \rightarrow aS \quad bT \quad \epsilon$$

One sentence - odd # of  $a$ 's:

$S \rightarrow aT \rightarrow aaS \rightarrow aabS \rightarrow aabaT \rightarrow aaba$

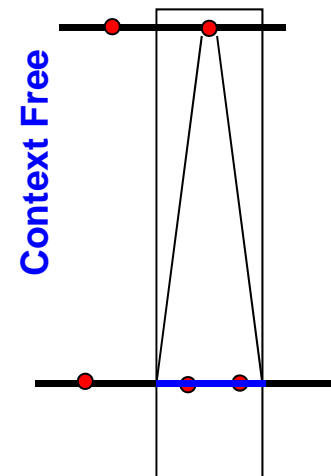


Context Free Grammar

$S \rightarrow aSa \quad bSb \quad aa \quad bb$

One sentence (even length palindromes):

$S \rightarrow aSa \rightarrow abSba \rightarrow abaaba$



# Stochastic Grammars

The grammars above classify all string as belonging to the language or not.

All variables has a finite set of substitution rules. Assigning probabilities to the use of each rule will assign probabilities to the strings in the language.

If there is a 1-1 derivation (creation) of a string, the probability of a string can be obtained as the product probability of the applied rules.

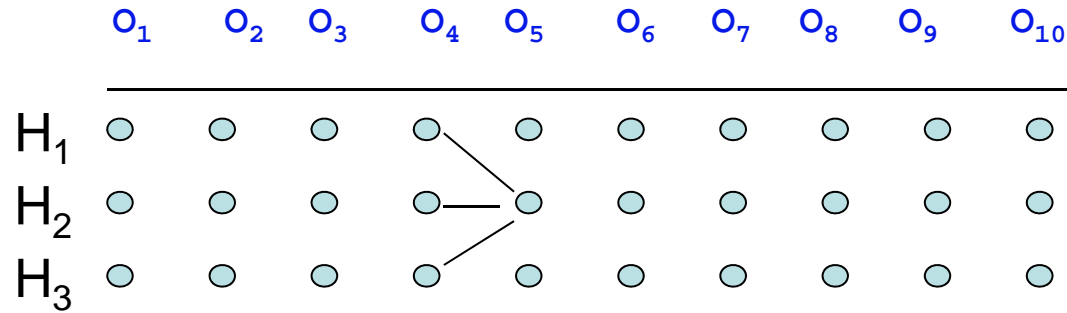
**i.** Start with **S**.  
 $S \rightarrow (0.3)aT \quad (0.7)bS$   
 $T \rightarrow (0.3)aS \quad (0.4)bT \quad (0.3)\epsilon$

$S \xrightarrow{*0.3} aT \xrightarrow{*0.3} aaS \xrightarrow{*0.7} aabS \xrightarrow{*0.3} aabaT \xrightarrow{*0.3} aaba$

**ii.**  $S \rightarrow (0.3)aSa \quad (0.5)bSb \quad (0.1)aa \quad (0.1)bb$

$S \xrightarrow{*0.3} aSa \xrightarrow{*0.5} abSba \xrightarrow{*0.1} abaaba$

# Hidden Markov Models in Bioinformatics



## Definition

## Four Key Algorithms

- **Summing over Unknown States**
- **Most Probable Unknown States**
- **Marginalizing Unknown States**
- **Optimizing Parameters**

# What is the probability of the data?

The probability of the observed is  $P(\vec{O}) = \sum_{\vec{H}} P(\vec{O}|\vec{H})P(\vec{H})$ , which could be hard to calculate. However, these calculations can be considerably accelerated. Let  $P_{O_k=i}^{H_k=j}$  the probability of the observations  $(O_1, \dots, O_k)$  conditional on  $H_k=j$ . Following recursion will be obeyed:

$$i. \quad P_{O_k=i}^{H_k=j} = P(O_k = i | H_k = j) \sum_{H_{k-1}=r} P_{O_{k-1}=r}^{H_{k-1}=r} P_{r,i}$$

$$ii. \quad P_{O_1=i}^{H_1=j} = P(O_1 = i | H_1 = j) \pi_j \quad (\text{initial condition})$$

$$iii. \quad P(O) = \sum_{H_n=j} P_{O_n=i}^{H_n=j}$$

	O <sub>1</sub>	O <sub>2</sub>	O <sub>3</sub>	O <sub>4</sub>	O <sub>5</sub>	O <sub>6</sub>	O <sub>7</sub>	O <sub>8</sub>	O <sub>9</sub>	O <sub>10</sub>
H <sub>1</sub>	○	○	○	●	○	○	○	○	○	○
H <sub>2</sub>	○	○	○	●	●	○	○	○	○	○

$$P_{O_5=i}^{H_5=2} = P(O_5 = i | H_5 = 2) \sum_{H_4=j} P_{O_4=j}^{H_4=j} P_{j,i}$$



# Genscan

■ State with length distribution

Exons of phase 0, 1 or 2

Introns of phase 0, 1 or 2

Initial exon

Terminal exon

Exon of single exon gene

5' UTR

3' UTR

Poly-A signal

Promoter

Forward (+) strand

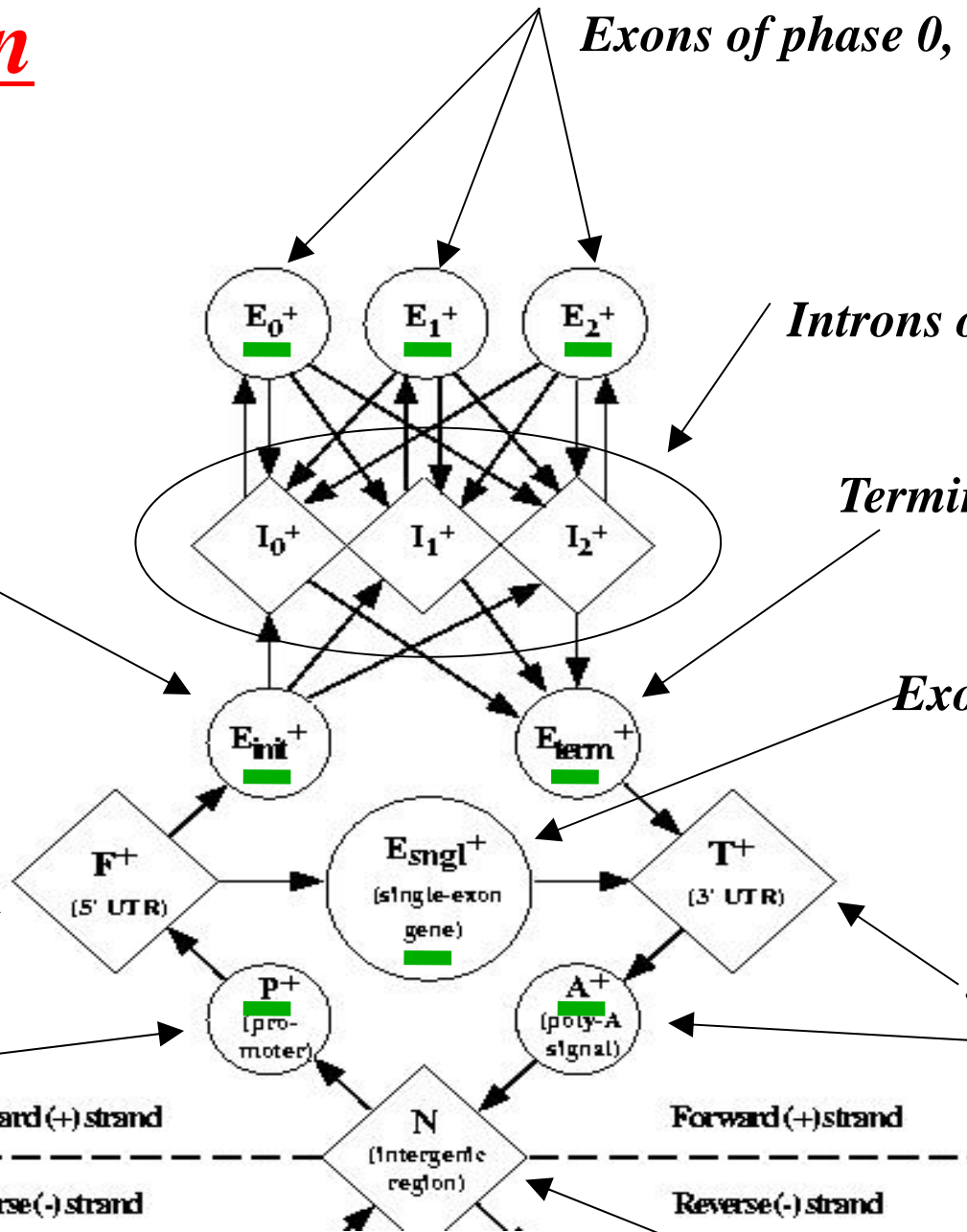
Forward (+) strand

Reverse (-) strand

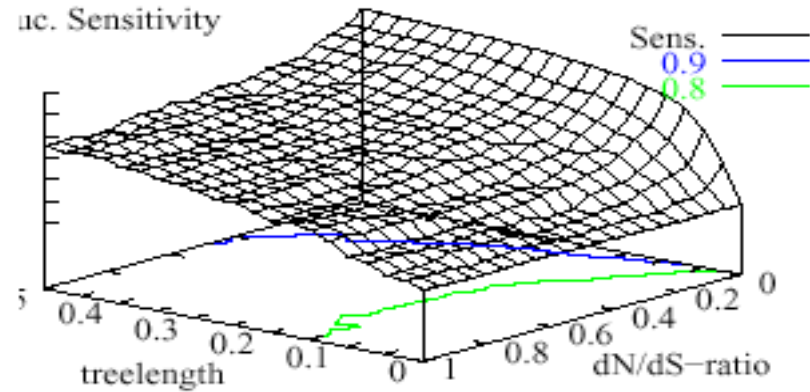
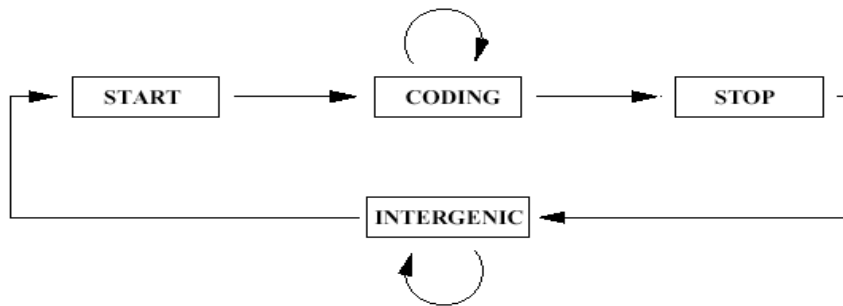
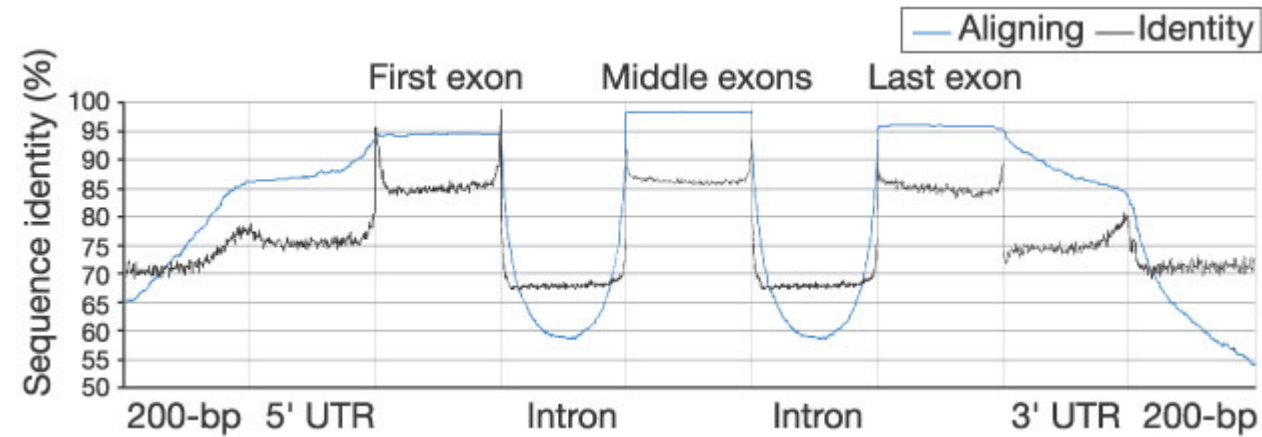
Reverse (-) strand

Intergenic sequence

Omitted: reverse strand part of the HMM



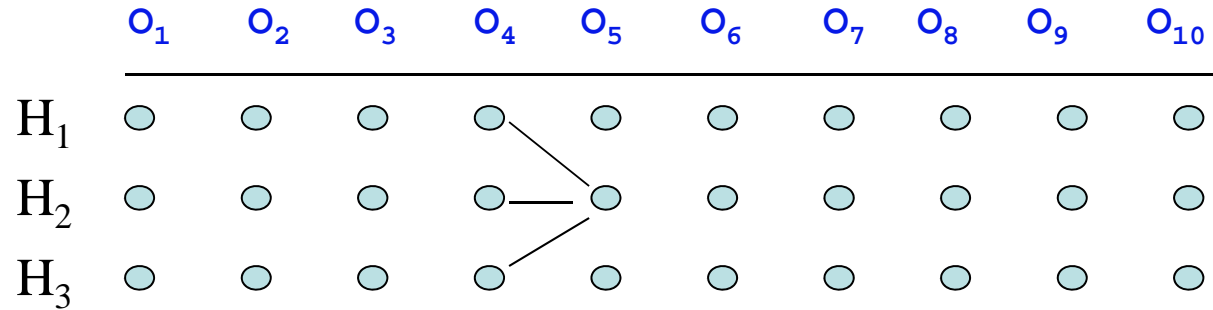
# Comparative Gene Annotation



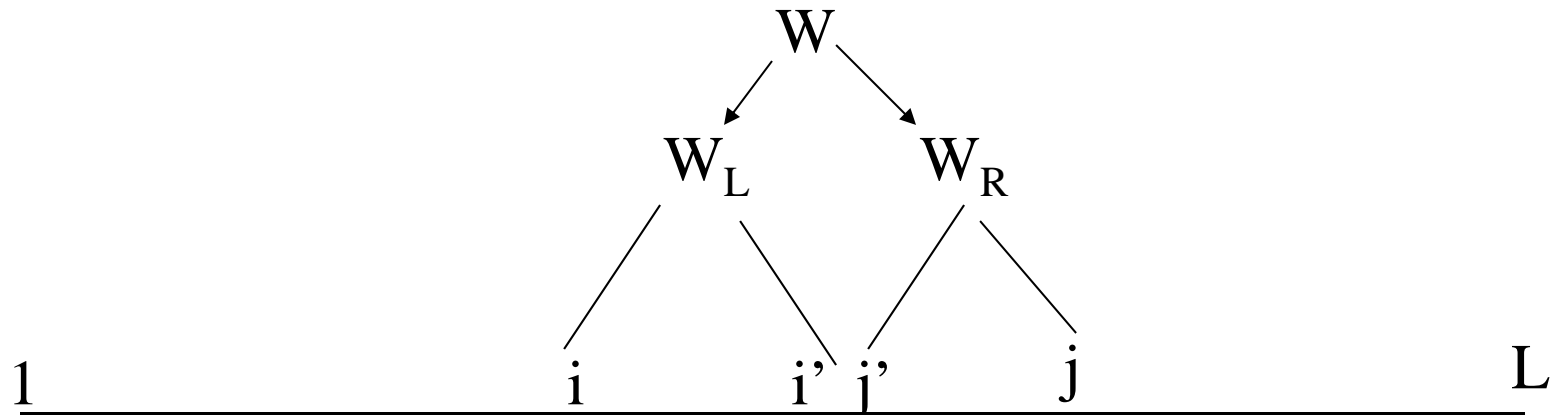
AGGTATATAATGCG.....  $P_{\text{coding}}\{\text{ATG-->GTG}\}$  or  
 AGCCATTTAGTGCG.....  $P_{\text{non-coding}}\{\text{ATG-->GTG}\}$

# SCFG Analogue to HMM calculations

## HMM/Stochastic Regular Grammar:



## SCFG - Stochastic Context Free Grammars:



# Secondary Structure Generators

<b>S</b>	-->	<b>LS</b>	<b>L</b>	.869	.131
<b>F</b>	-->	<b>dFd</b>	<b>LS</b>	.788	.212
<b>L</b>	-->	<b>s</b>	<b>dFd</b>	.895	.105

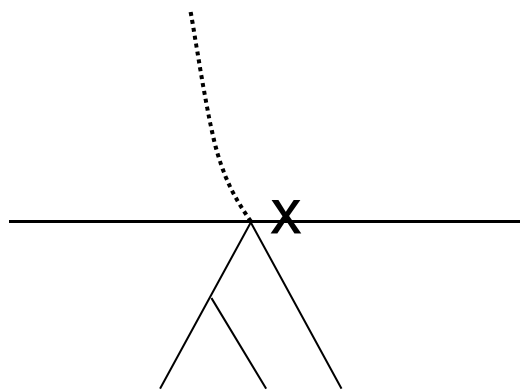
*S* → *LS* → *LLLLLLLS* → *LLLLLLLLL*  
 → *ssLsssss* → *ssdFdsssss*  
 → *ssdddFdddsssss*  
 → *ssdddLSdddsssss*  
 → *ssdddLLLdddsssss*  
 → *ssdddssssdddsssss*

*s* <sup>*SS*</sup> *s*  
*d-d*  
*d-d*  
*SS* *d-d* *SSSSS*



# Observing Evolution has 2 parts

$P(x)$ :



$P(\text{Further history of } x)$ :

