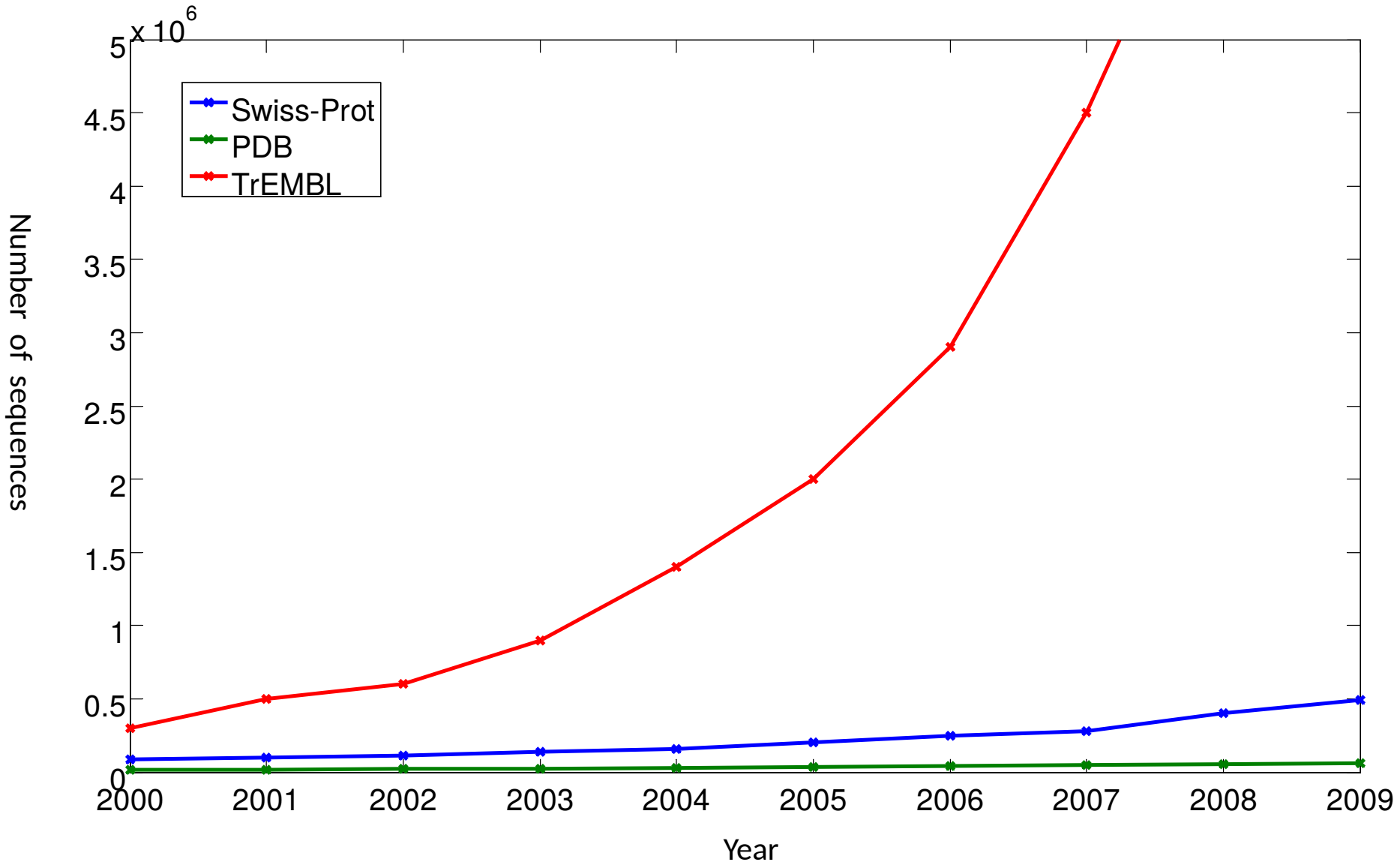
The Genomic Knowledge Gap



Representing the information

- **Sequences**(biology) and **strings** (computer science)
- DNA/RNA/Protein *sequences* are represented as *strings*
- Standard alphabets
  - DNA {A,T,G,C} {R,Y,N}
  - RNA {A,U,G,C} {R,Y,N}
  - Protein {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y} {X,U,*}

Formalization of sequence representation

- A DNA sequence **s** is a finite string from the alphabet *N*={A,T,G,C} of nucleotides
- A genome is the set of all DNA sequences associated with an organism or an organelle
- Subsequences a.k.a “slices”:
- s=ATATGTCGTGCA
- s(3:6)=ATGT
- S(8) = G

Formalization of sequence representation

- Subsequences a.k.a “slices”:
  - s=ATATGTCGTGCA
  - s(3:6)=ATGT
  - s(8) = G
- Concatenations:
  - s(3:6) + s(8) = ATGT+G = ATGTG

Probabilistic models of sequences: why?

- We want to find “interesting” elements in a genome
- Interesting → statistically significant
- Genes vs. non-genes: *length*, *base composition*
  - *Edge requirement: start/stop codons*
- Introns vs. exons
- More?

Simple model: multinomial

- Nucleotides are independent and identically distributed (i.i.d)
- p(A)+p(C)+p(G)+p(T) = 1
- Even simpler: p(A)=p(C)=p(G)=p(T) = 0.25

$$p(s) = \prod_{i=1}^{n} p(s(i))$$

Multinomial model

- p(A) = 0.19 p(C)=0.21 p(G)=0.27 p(T)=0.32
- p(AA) = 0.19*0.19 = 0.19^2 = 0.0361
- p(AAAAAAA) = 0.19^7 = 8.94×10^{-6}
- p(AATGCGT) =?

Multinomial model

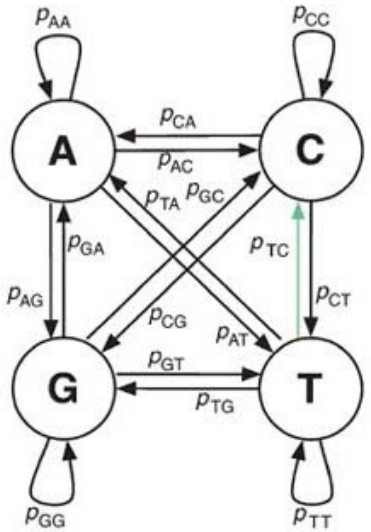
- p(A) = 0.19 p(C)=0.21 p(G)=0.27 p(T)=0.32
- p(AA) = 0.19*0.19 = 0.19^2 = 0.0361
- p(AAAAAAA) = 0.19^7 = 8.94×10^{-6}
- p(AATGCGT) =?

A human genome is 3×10^9 base-pairs. How many AATGCGT would you expect to find by chance?

Multinomial model

- Pros: simple, works quite a few times. Good as a baseline
- Con: i.i.d assumption: we know nucleotides are not independent nor identically distributed

Markov sequence model



	A	C	G	T
A	p_{AA}	p_{AC}	p_{AG}	p_{AT}
C	p_{CA}	p_{CC}	p_{CG}	p_{CT}
G	p_{GA}	p_{GC}	p_{GG}	p_{GT}
T	p_{TA}	p_{TC}	p_{TG}	p_{TT}

Look ma! No i.i.d!!

Markov Transition Matrix

	To A	To C	To G	To T
From A	0.6	0.2	0.1	0.1
From C	0.1	0.1	0.8	0
From G	0.2	0.2	0.3	0.3
From T	0.1	0.8	0	0.1

ACGCGTAATCAAAATCGGTCGTCGGAAAAAAAAATCG

Probabilistic models: summary

- “All models are wrong, but some are useful”
- Markov chain and multinomial models are both used
- Statistical anomalies discovered by the model may have biological significance

Why is GC content important?*

$$\frac{G+C}{A+T+G+C} \times 100$$

Why is GC content important?

- Identifying horizontal gene transfer elements: *change point analysis*
- Coding regions have a higher GC content
- Systematics
  - GC-rich: Actinobacteria (60-70%)
  - AT-rich: Plasmodium (~20%)
- Annealing temperature for PCR primers

k-mer frequency analysis

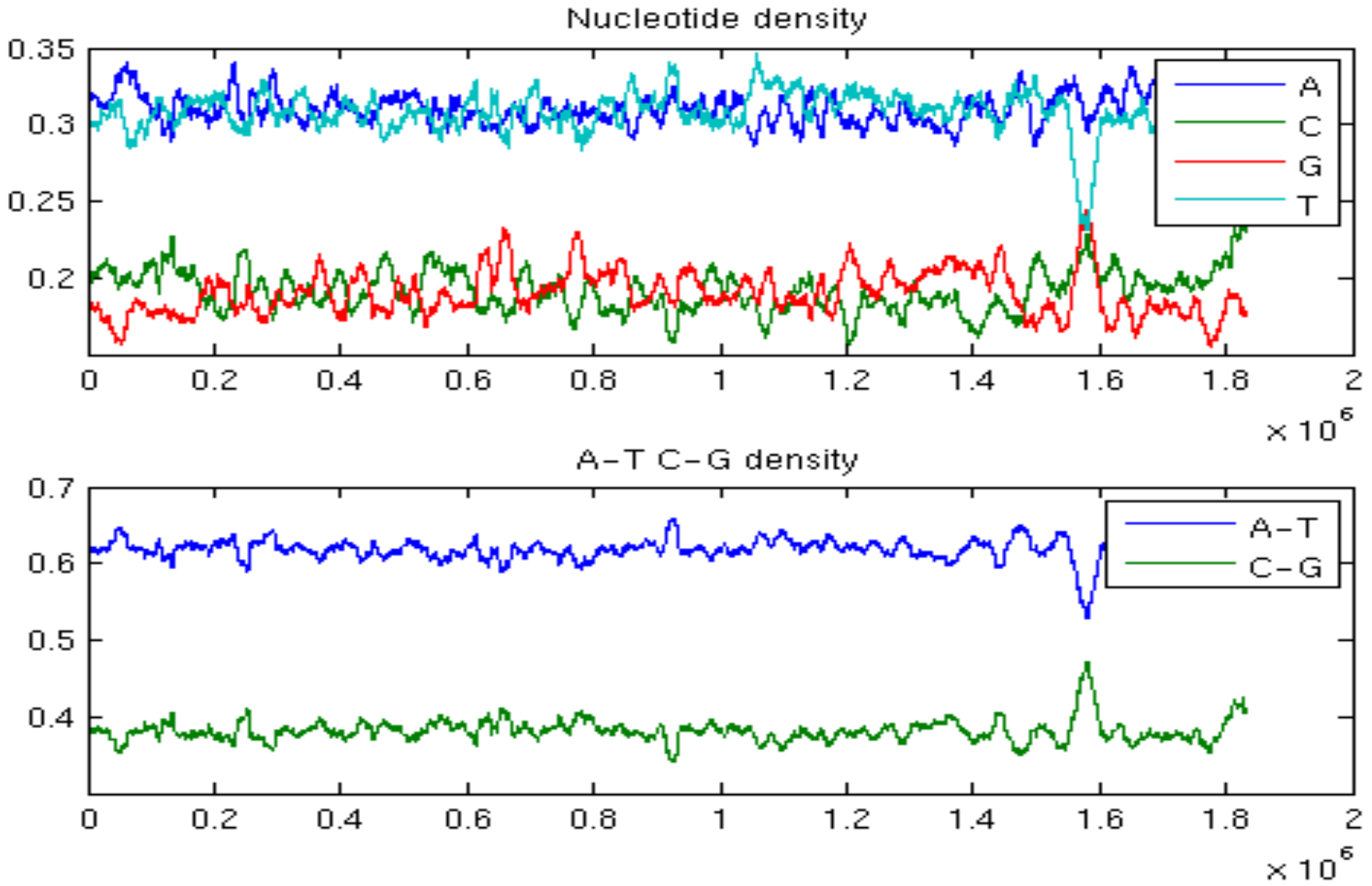
- Analyzing the frequency of “words” in the genome
- 1-mer words: A, T, G, C
- 2-mer words: AA, AT, AG, AC, TA, TT,TG,TC...
- 3-mer words: AAA, AAT, AAG...
- Number of *k*-mers = 4^k
- Number of protein k-mers = ?

Unusual *k*-mers

- Words that are highly frequent or highly infrequent may have a biological meaning
- CTAG kinks DNA
- Palindromic *k*-mers: restriction sites: EcoRI / GAATTC
- CpG islands

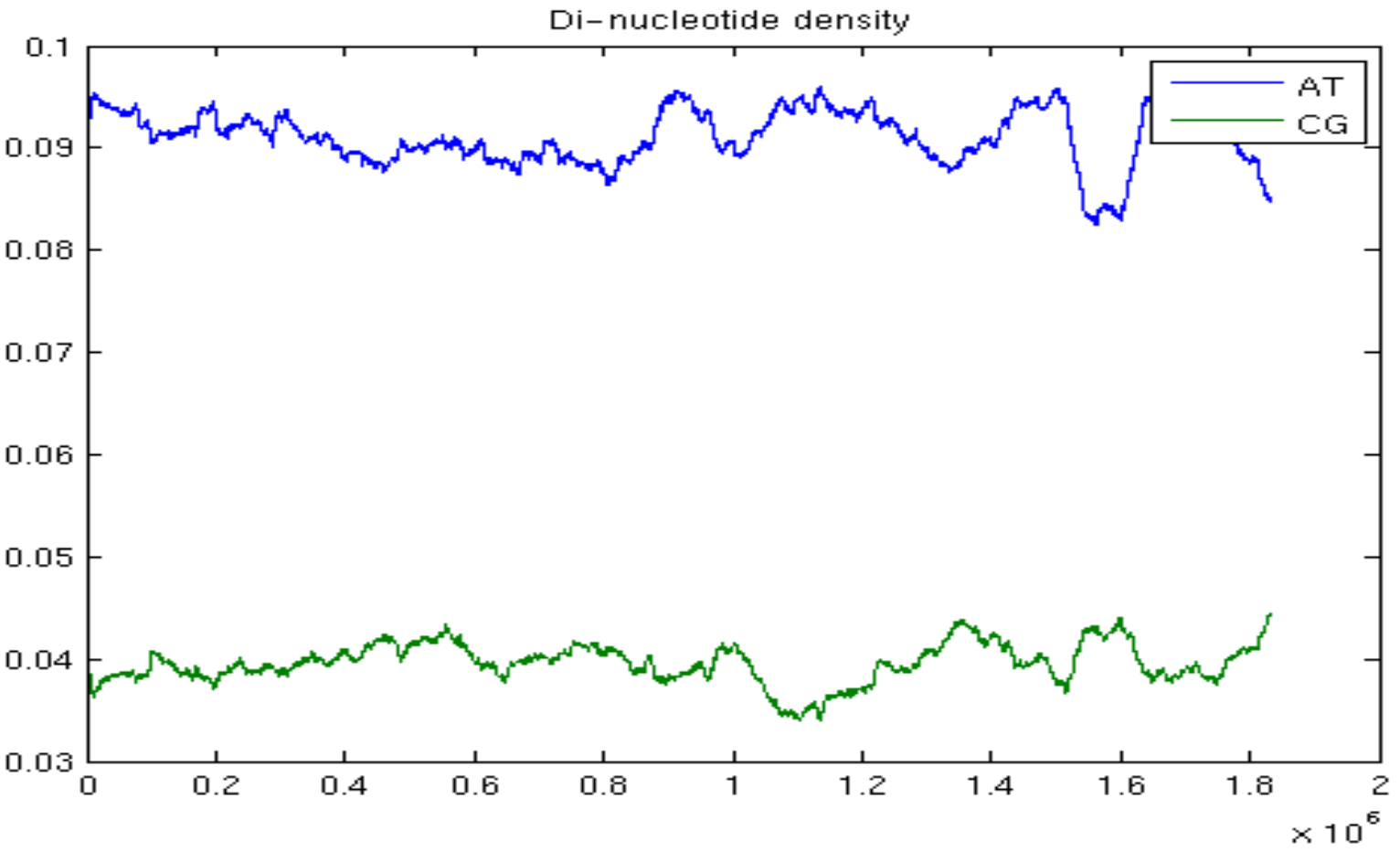
Statistical sequence analysis

- Base composition of the genome and variations in the genome of *H. influenzae* (20 Kbp window size)



Statistical sequence analysis

- Base composition of the genome and variations in the genome



The Odds Ratio

- Observed / expected

Example for the AT dimer:

$$\frac{P(AT)}{P(A)\times P(T)}$$

Odds ratio >> 1 or << 1: unusual dimer

Genome statistics: conclusion

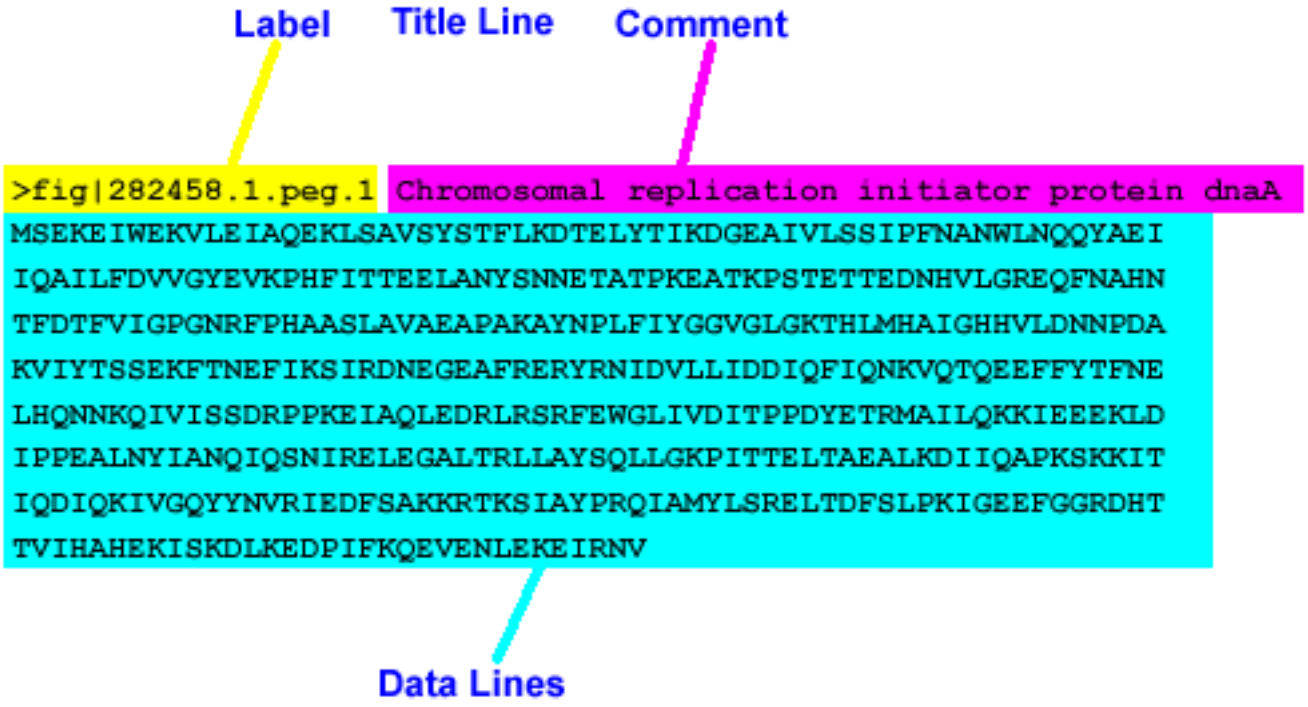
- Whole genome statistics lets us broadly typify a genome
- Statistical biases along a genome tell us about:
  - Horizontal gene transfer
  - Possible coding regions
  - isochores

Data Formats

- For analysis, sequences are stored in text files, which are read by programs
- Programs are very finicky about what they read
- A number of standard formats exist
- 1 letter<--> 1 base (or amino-acid)
- Information in addition to the sequence:
  - Source organism, keywords, unique database identifiers, sequence features...

FASTA Format

- Pros: simple.
- Cons: does not contain information about the sequence



See for yourself

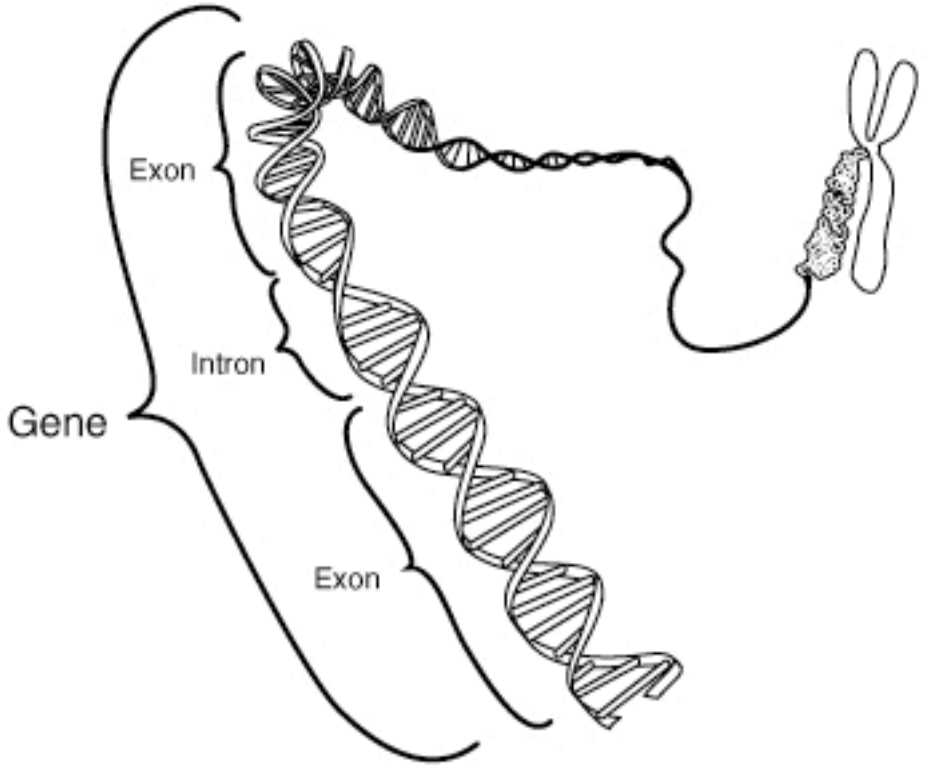
- Steroidogenic Acute Regulatory Protein (StAR)
- http://www.ncbi.nlm.nih.gov/protein/71152974
- http://www.uniprot.org/uniprot/P49675
- http://www.uniprot.org/uniprot/P49675.txt
- http://www.uniprot.org/uniprot/P49675.fasta
- http://www.uniprot.org/uniprot/P49675.gff

Finding Genes

What is a Gene?*

What is a Gene?*

- Basic unit of heredity in an organism
- Do not confuse gene with allele!
- Genes have *products*: protein or RNA



ORF Finding

- ORF: Open Reading Frame
- Also called gene finding, gene calling
- Uses the conservative definition of a gene: i.e. a contiguous coding unit
- Hey, we have to start *somewhere*!

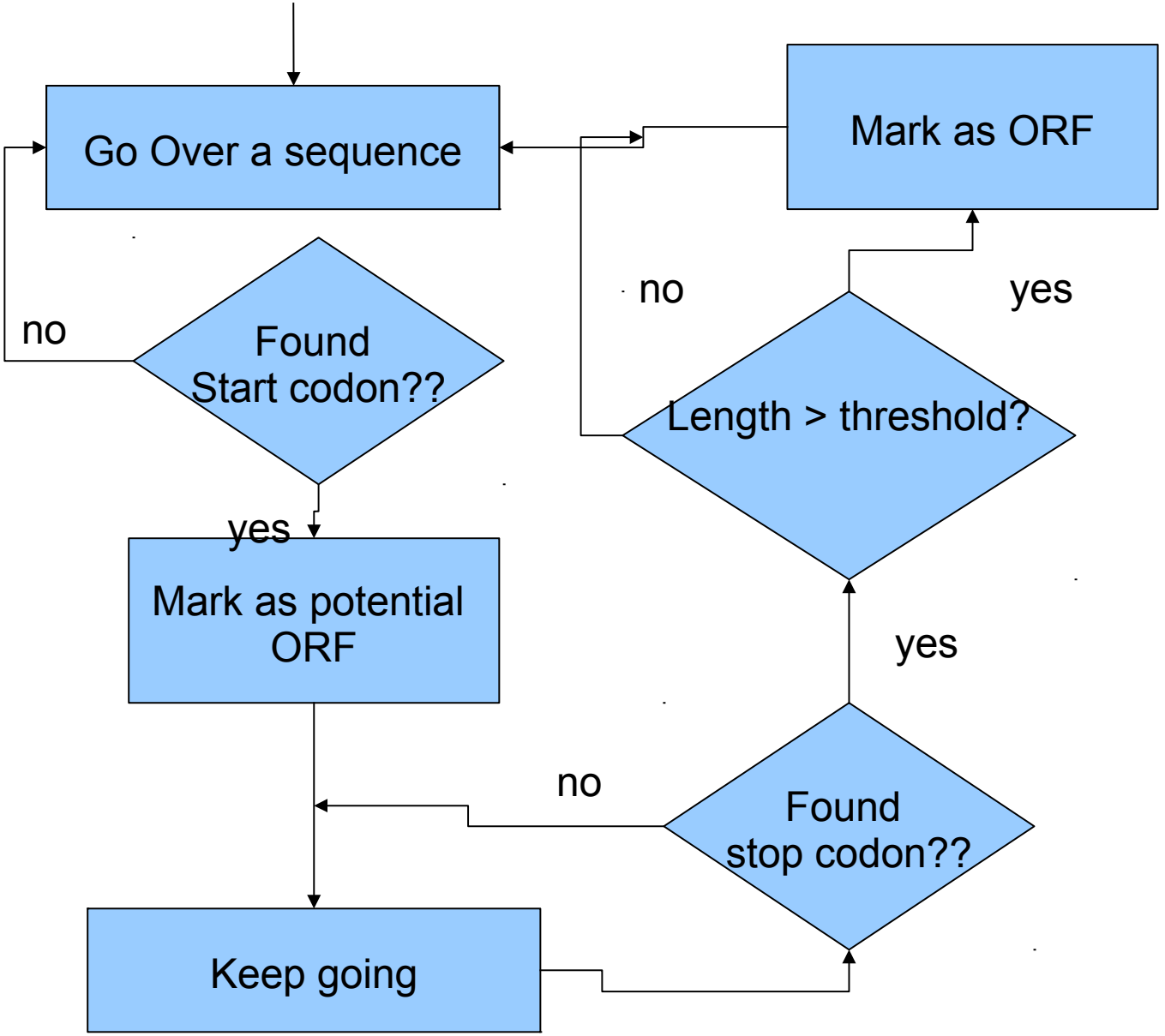
To ORF or not to ORF?

- Today's limits: protein coding genes
- Prokaryotes (no messy introns)

ORF finding algorithm*

?

ORF finding algorithm 1*



Pros and Cons of ORF Finder 1*

Pros and Cons of ORF Finder 1*

- Pros:
  - Simple
- Cons:
  - How do we set the length threshold?
  - What if a true ORF is shorter than the length threshold? Or a false ORF longer than the the length threshold?
  - Should length be the only determinant of being a true ORF?

Hypothesis testing (ifs and buts)

- Ho: A potential ORF does not code for a gene
- H1: A potential ORF codes for a gene

if / but	**Ho is true**	**Ho is false**
Ho is rejected	Type I error (FALSE POSITIVE)	No error
Ho is not rejected	No error	Type II error (FALSE NEGATIVE)

Hypothesis testing

- Ho: A potential ORF does not code for a gene
- H1: A potential ORF a true ORF

	If Ho is true	**If Ho is false**
If Ho is rejected	Type I error (FP)	No error
If Ho is not rejected	No error	Type II error (FN)

Keep it *known* and *low*
$\alpha < 0.05$

Type I error:saying that some sequence is an ORF, when it is not (FALSE POSITIVE)

More big words to confuse you

- **Significance level:** the probability of committing a Type I error. Also called α
- **Test statistic**: what we wish to test Ho against. In our case, ORF length
- ***p*-value**: the probability of finding the observed or more extreme test statistic when Ho is true
  - if p-value < significance level the result is significant: Ho is rejected.
- **Significant result**: not necessarily a true result.

Back to ORF Finding

- Informal: How long does an ORF have to be to be considered “significant”?
- Formal: what is the probability of an ORF of *k* or more codons arising by chance?
- More formal: what is the threshold value of *k* so that 95% of random ORFs are < *k?*

First stab at finding a good ORF length

- Assume codons are uniformly distributed in the genome
- P(consecutive run of *k* non-stop-codons) $=(61/64)^k$

?

First stab at finding a good ORF length

- Assume codons are uniformly distributed in the genome
- P(consecutive run of *k* non-stop-codons) $=(61/64)^k$
- Setting $\alpha=0.05$ we get $k=62$
- $(61/64)^{62}$
- 62*3=186bp

?=62

Refining finding a good ORF length

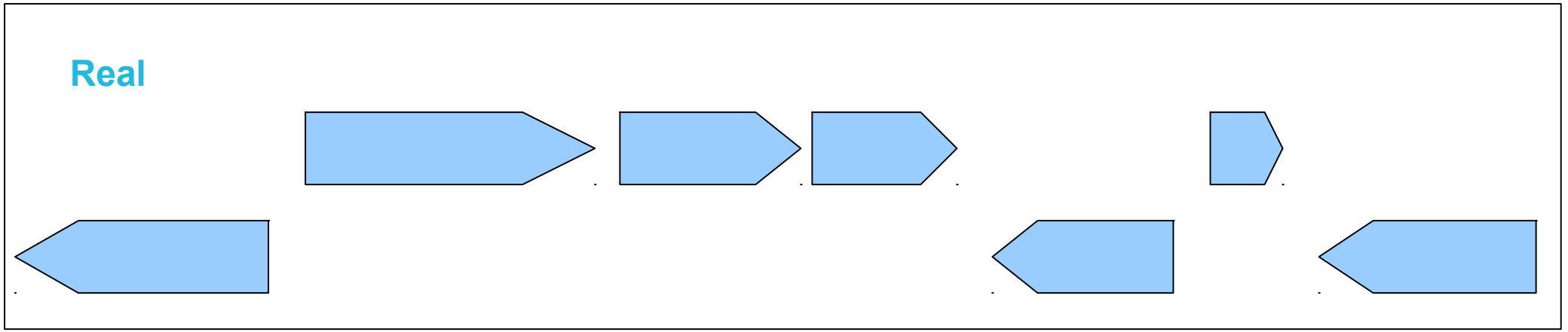
- Codons are NOT uniformly distributed!
- In fact, their distribution varies between genomes (and even within a genome).
- We need to refine the *null model* on a per-genome basis

?=62

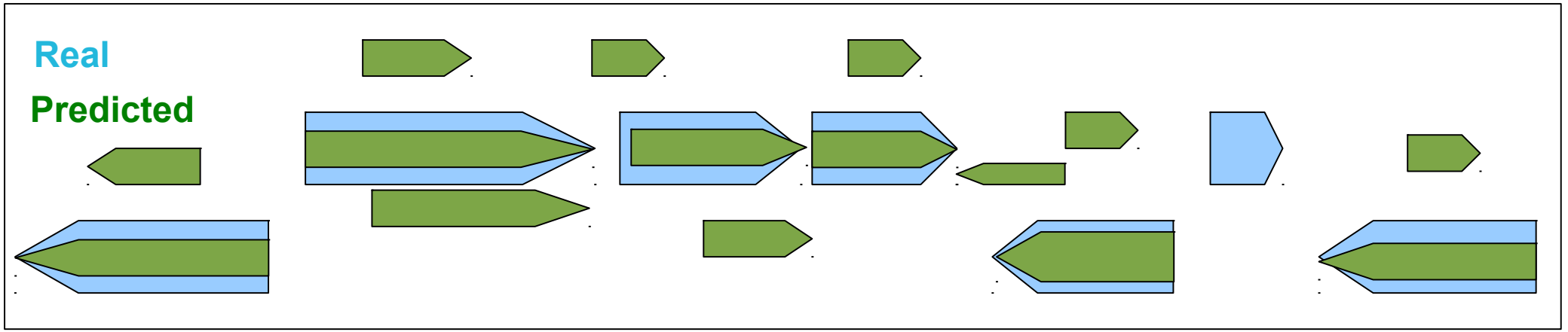
Randomizing Genomes

- We can *permute* (shuffle) the genome for a genome-specific null-model.

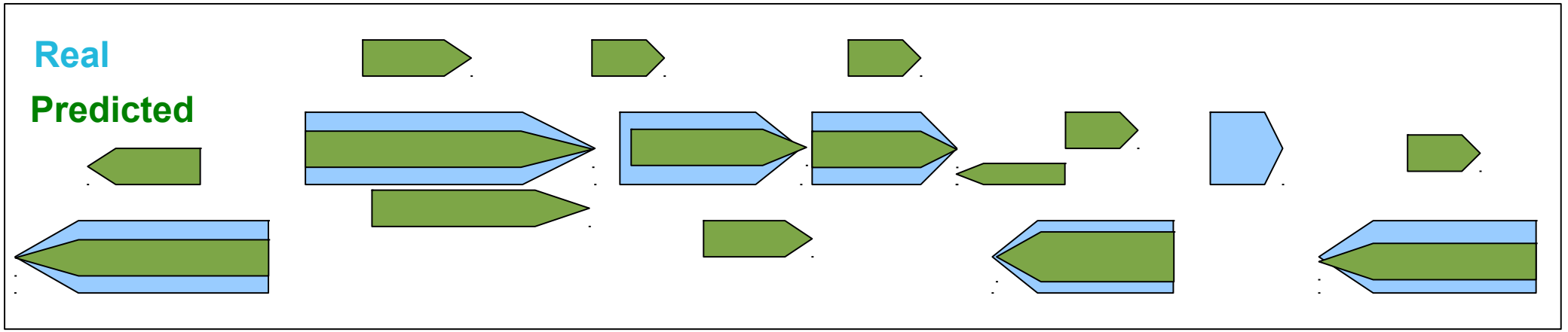
Casino Genome

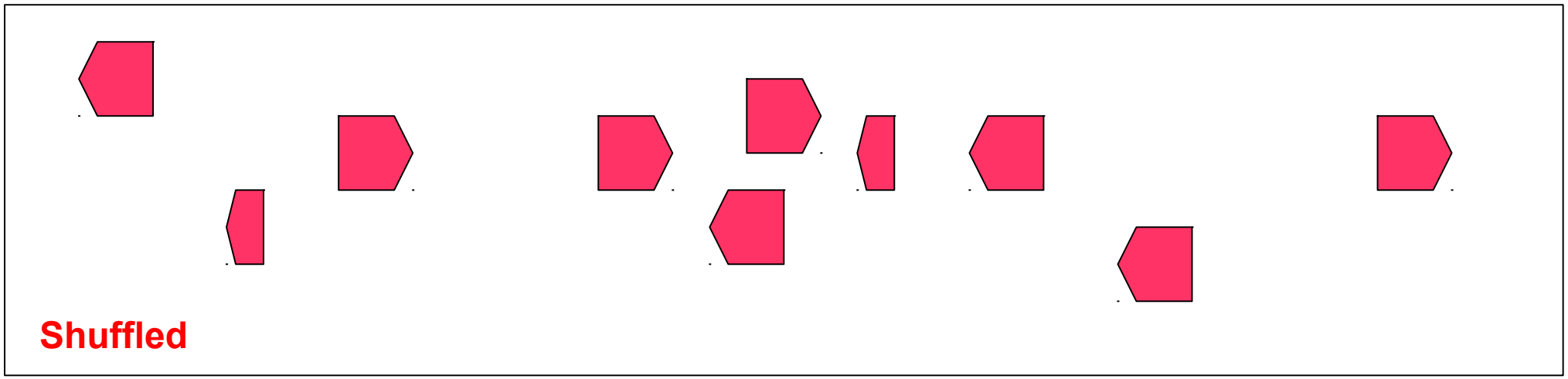


Casino Genome

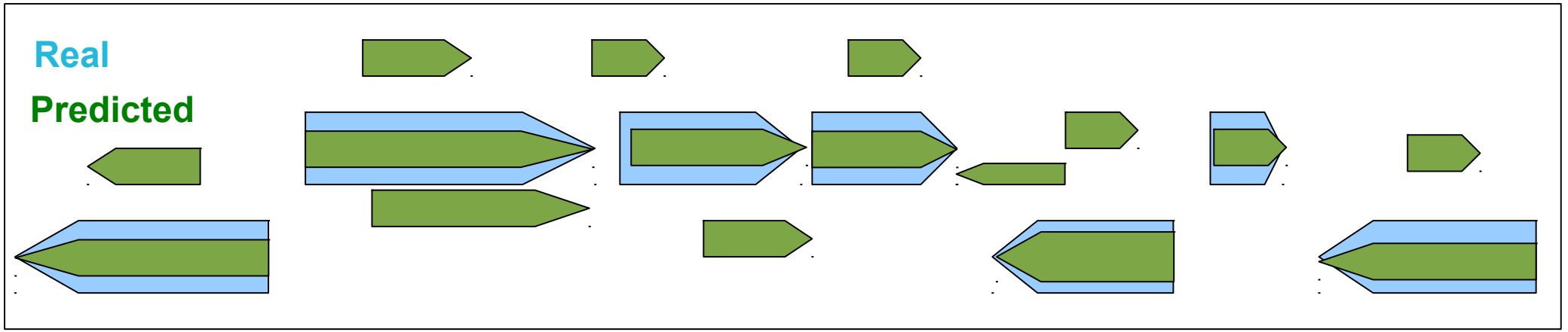


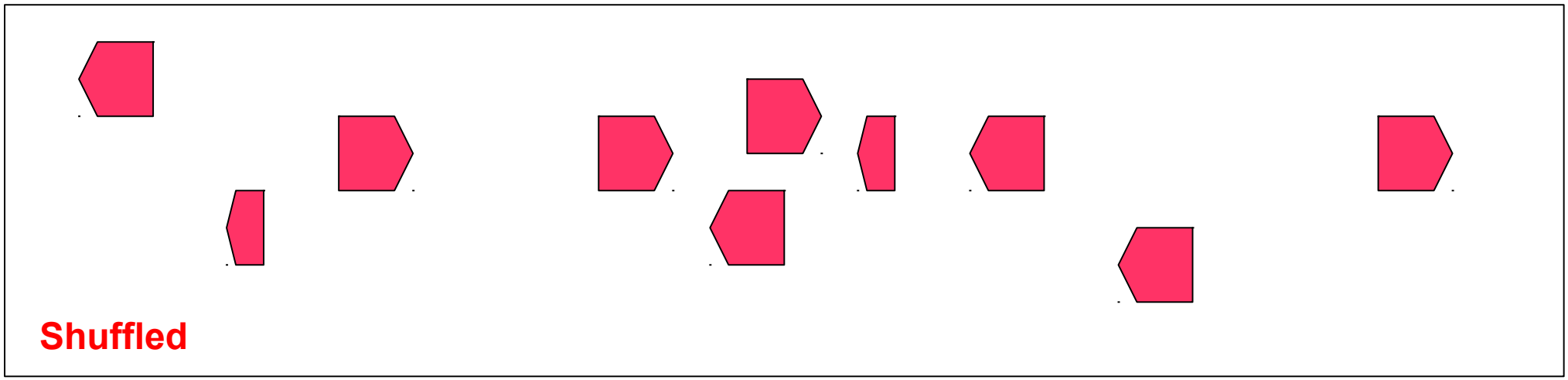
Casino Genome



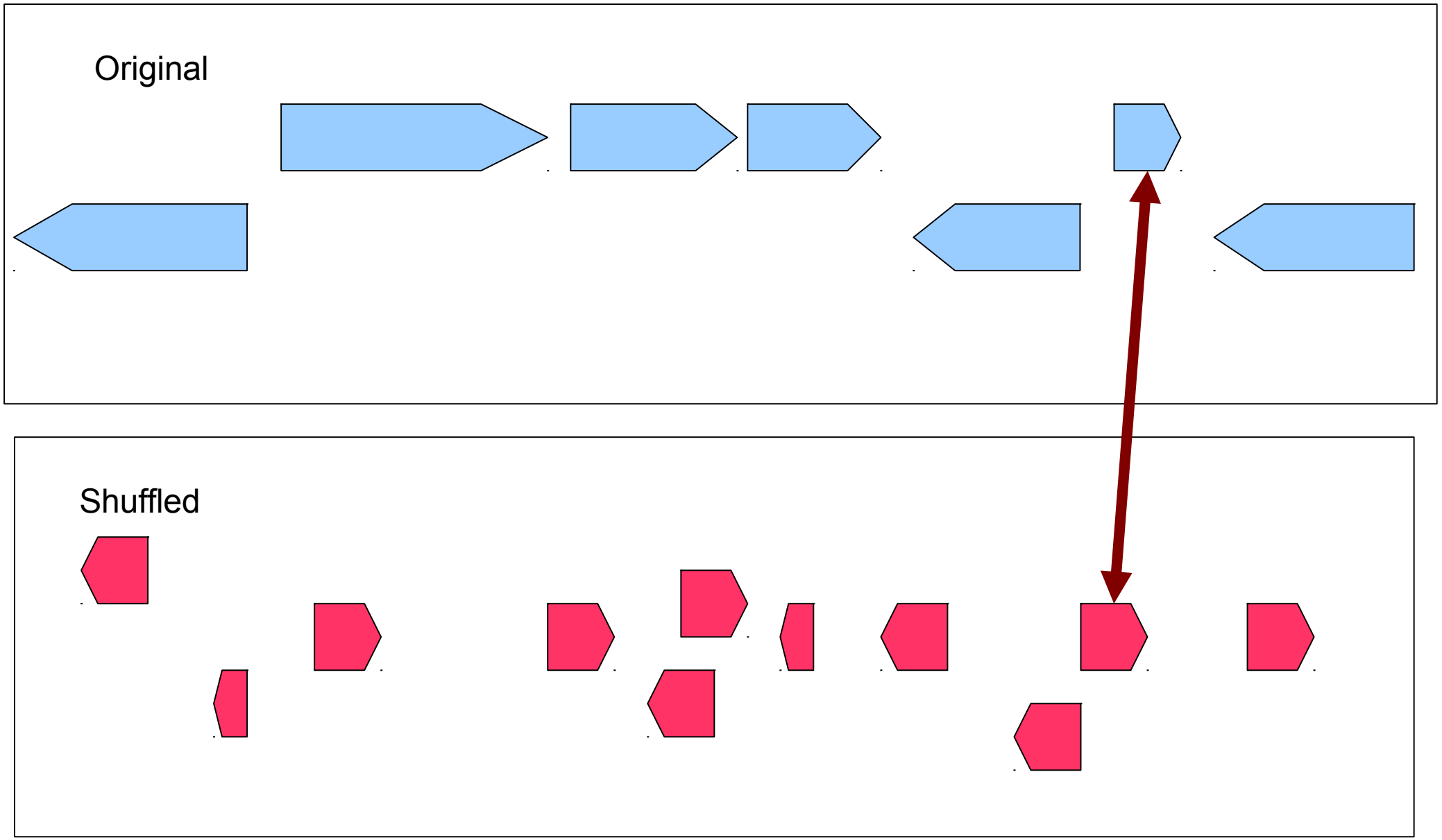


Casino Genome

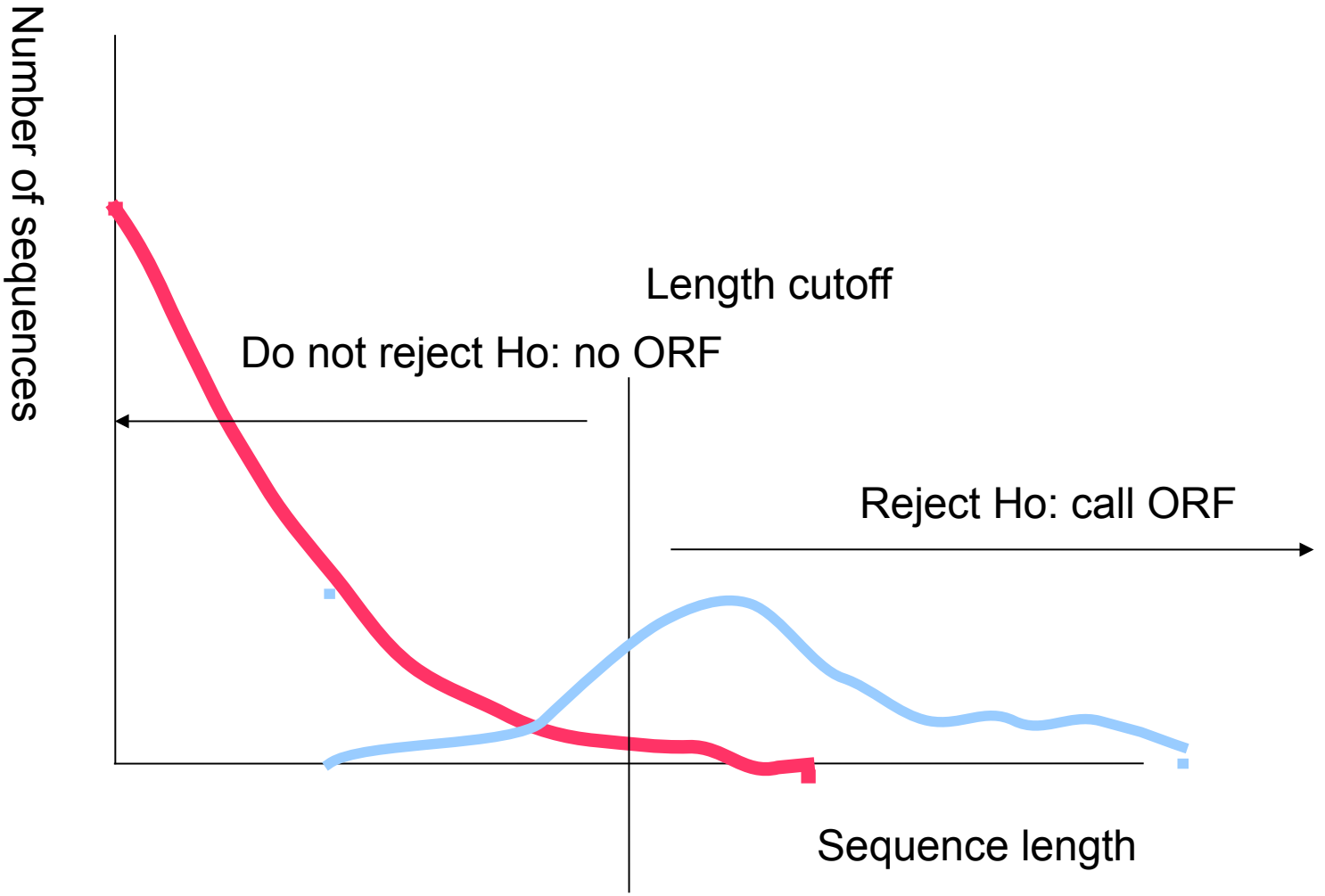




Casino Genome



Choosing a threshold



Choosing a threshold