CSE 427
Computational Biology

Gene Prediction

A statistical interlude: Fair or biased?

H H H H T H H T T H

More likely fair or biased?

H H H H T H H T T H

More likely H0 or H1?

H H H H T H H T T H

- H0: .5 – .5
- H1: .9 – .1

Quantify likelihood: H_0 vs H_1

H H H H T H H T T H

H0: .5 – .5 .5^10

H1: .9 – .1 .9^7 * .1^3

Likelihood ratio: (.5^10)/(.9^7 * .1^3) = .4898

(I.e., odds favor “biased” by about 2:1)

Gene Finding: Motivation

Sequence data flooding into Genbank

What does it mean?

protein genes, RNA genes, mitochondria, chloroplast, regulation, replication, structure, repeats, transposons, unknown stuff, …

Protein Coding Nuclear DNA

Focus of this lecture

Goal: Automated annotation of new sequence data

State of the Art:

- In Eukaryotes:
  - predictions ~ 60% similar to real proteins
  - ~80% if database similarity used
- Prokaryotes
  - better, but still imperfect
- lab verification still needed, still expensive

Biological Basics

Central Dogma:

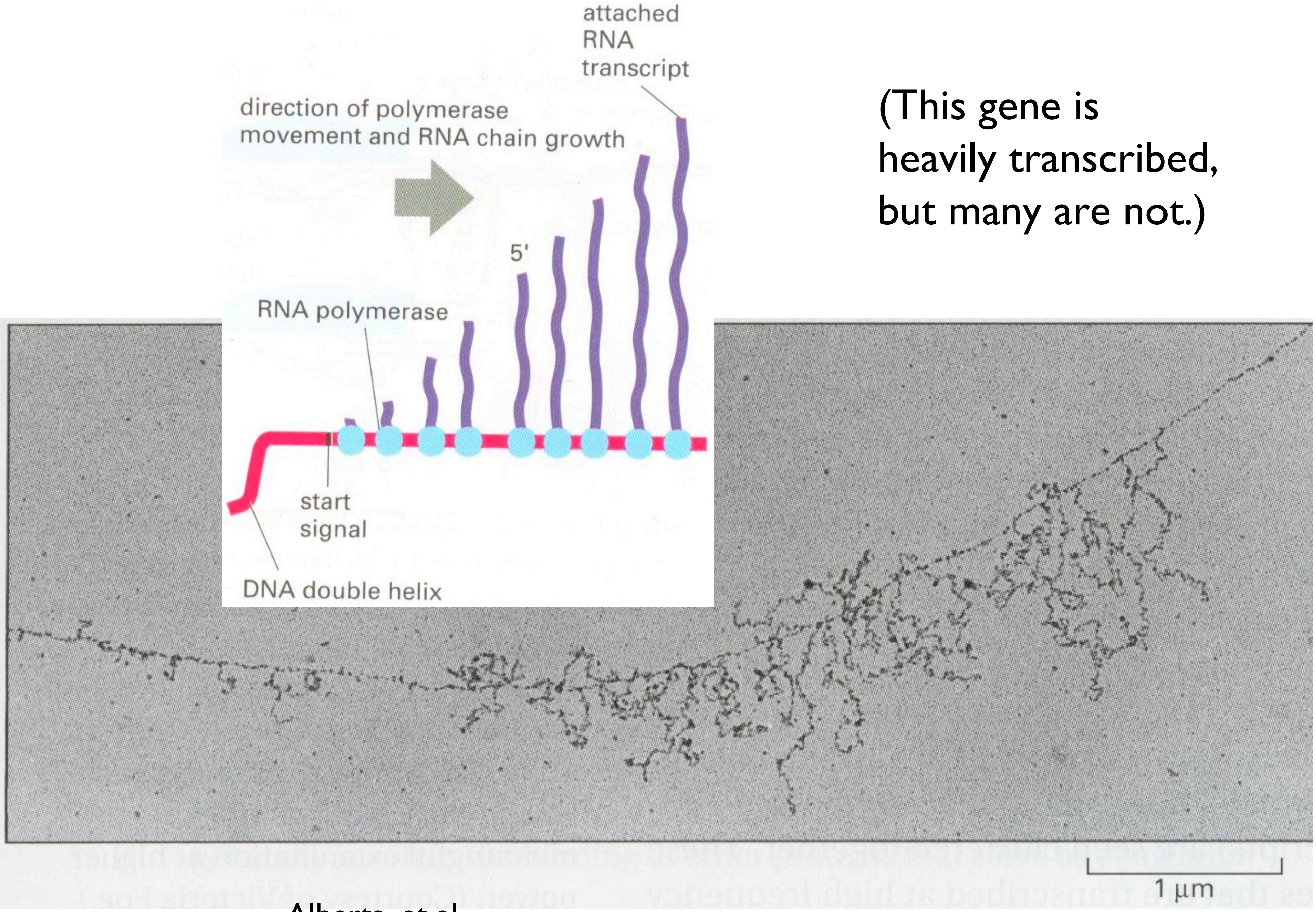
DNA $\xrightarrow{\text{transcription}}$ RNA $\xrightarrow{\text{translation}}$ Protein

Codons: 3 bases code one amino acid

Start codon

Stop codons

3’, 5’ Untranslated Regions (UTR’s)



(This gene is heavily transcribed, but many are not.)

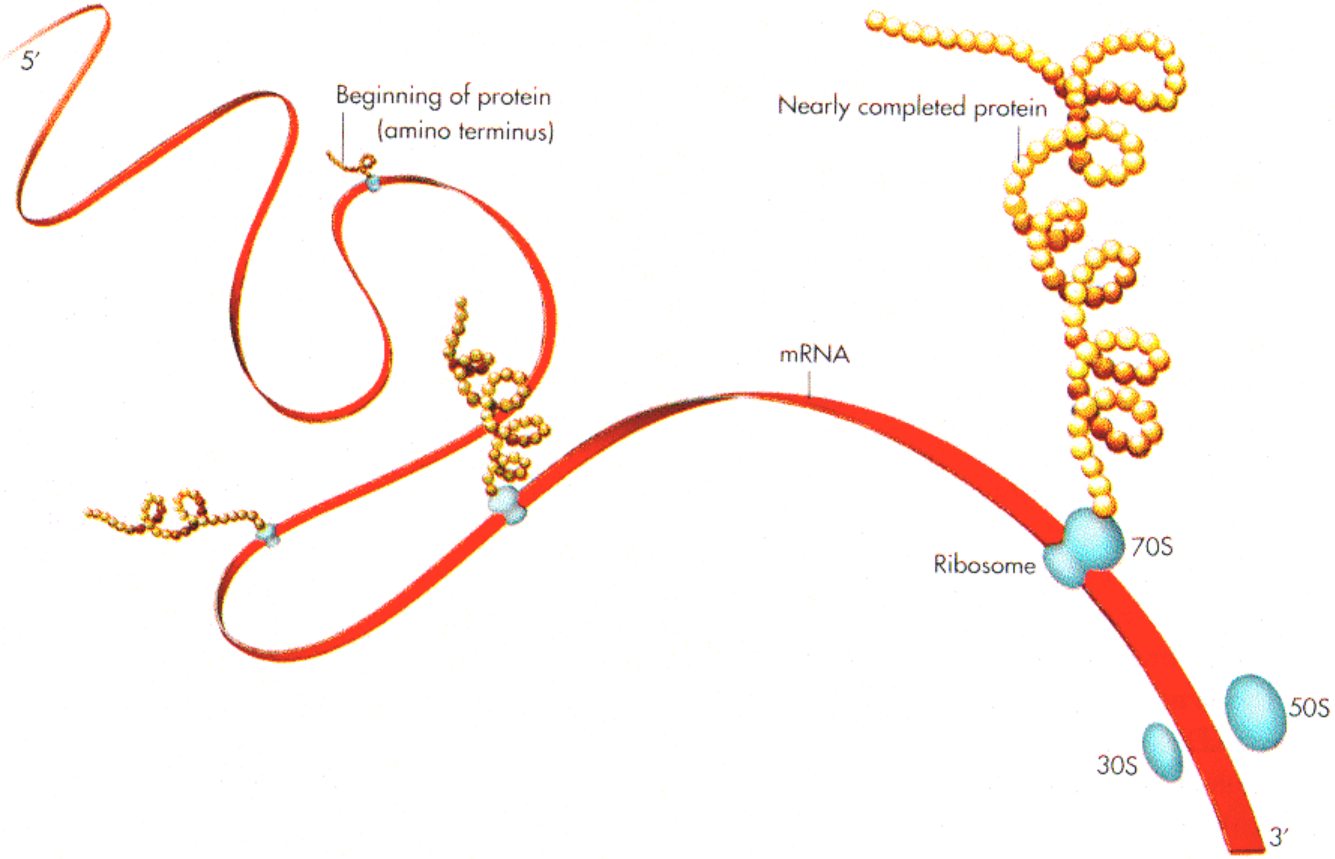
Alberts, et al.

Codons & The Genetic Code

		Second Base					
		U	C	A	G		
First Base	U	Phe	Ser	Tyr	Cys	U	Third Base
		Phe	Ser	Tyr	Cys	C	
		Leu	Ser	Stop	Stop	A	
		Leu	Ser	Stop	Trp	G	
	C	Leu	Pro	His	Arg	U	
		Leu	Pro	His	Arg	C	
		Leu	Pro	Gln	Arg	A	
		Leu	Pro	Gln	Arg	G	
	A	Ile	Thr	Asn	Ser	U	
		Ile	Thr	Asn	Ser	C	
		Ile	Thr	Lys	Arg	A	
		Met/Start	Thr	Lys	Arg	G	
	G	Val	Ala	Asp	Gly	U	
		Val	Ala	Asp	Gly	C	
		Val	Ala	Glu	Gly	A	
		Val	Ala	Glu	Gly	G	

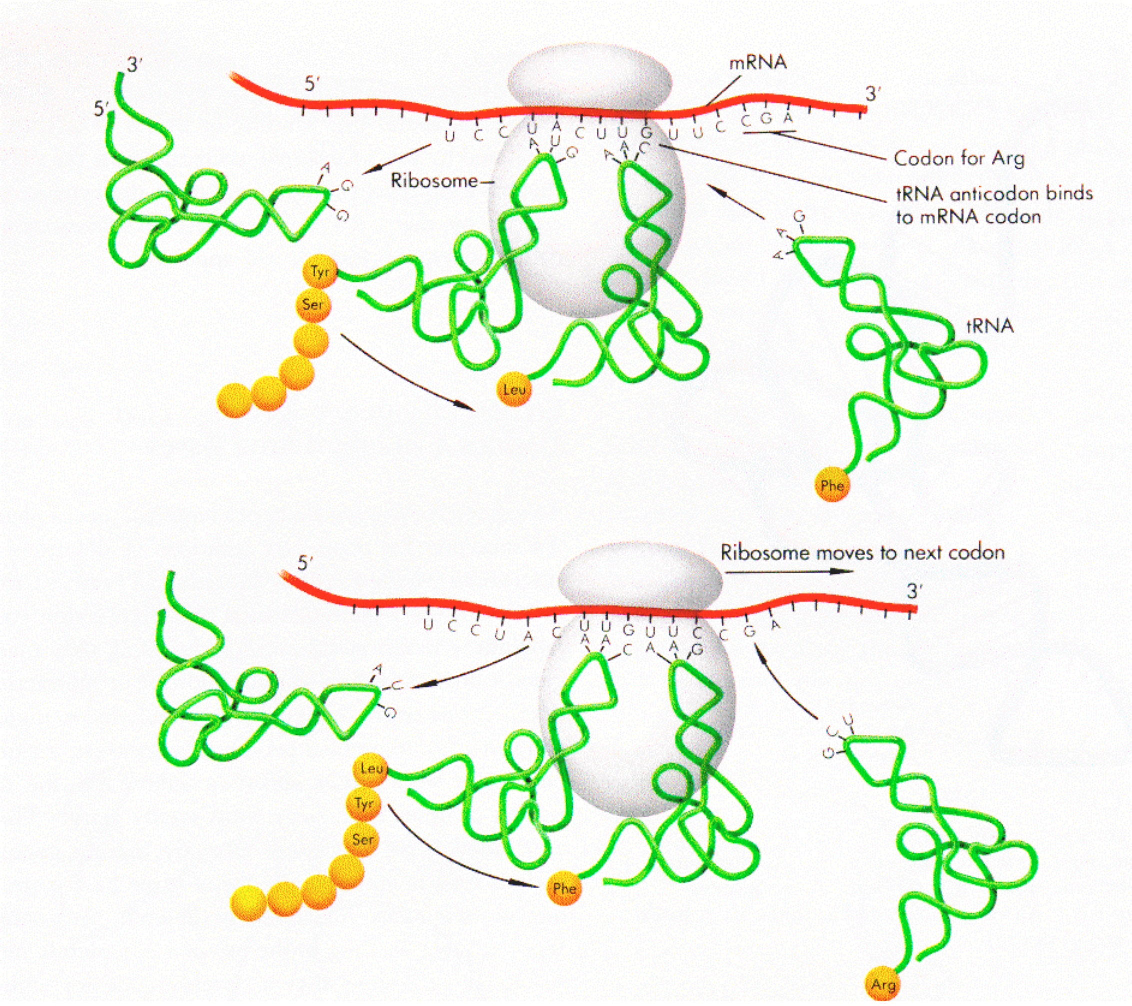
Ala	:	Alanine
Arg	:	Arginine
Asn	:	Asparagine
Asp	:	Aspartic acid
Cys	:	Cysteine
Gln	:	Glutamine
Glu	:	Glutamic acid
Gly	:	Glycine
His	:	Histidine
Ile	:	Isoleucine
Leu	:	Leucine
Lys	:	Lysine
Met	:	Methionine
Phe	:	Phenylalanine
Pro	:	Proline
Ser	:	Serine
Thr	:	Threonine
Trp	:	Tryptophane
Tyr	:	Tyrosine
Val	:	Valine

Translation: mRNA → Protein



Watson, Gilman, Witkowski, & Zoller, 1992

Ribosomes



Watson, Gilman, Witkowski, & Zoller, 1992

Idea #1: Find Long ORF’s

Reading frame: which of the 3 possible sequences of triples does the ribosome read?

Open Reading Frame: No stop codons

In random DNA

- average ORF = 64/3 = 21 triplets
- 300bp ORF once per 36kbp per strand

But average protein ~ 1000bp

A Simple ORF finder

start at left end

scan triplet-by-non-overlapping triplet for AUG

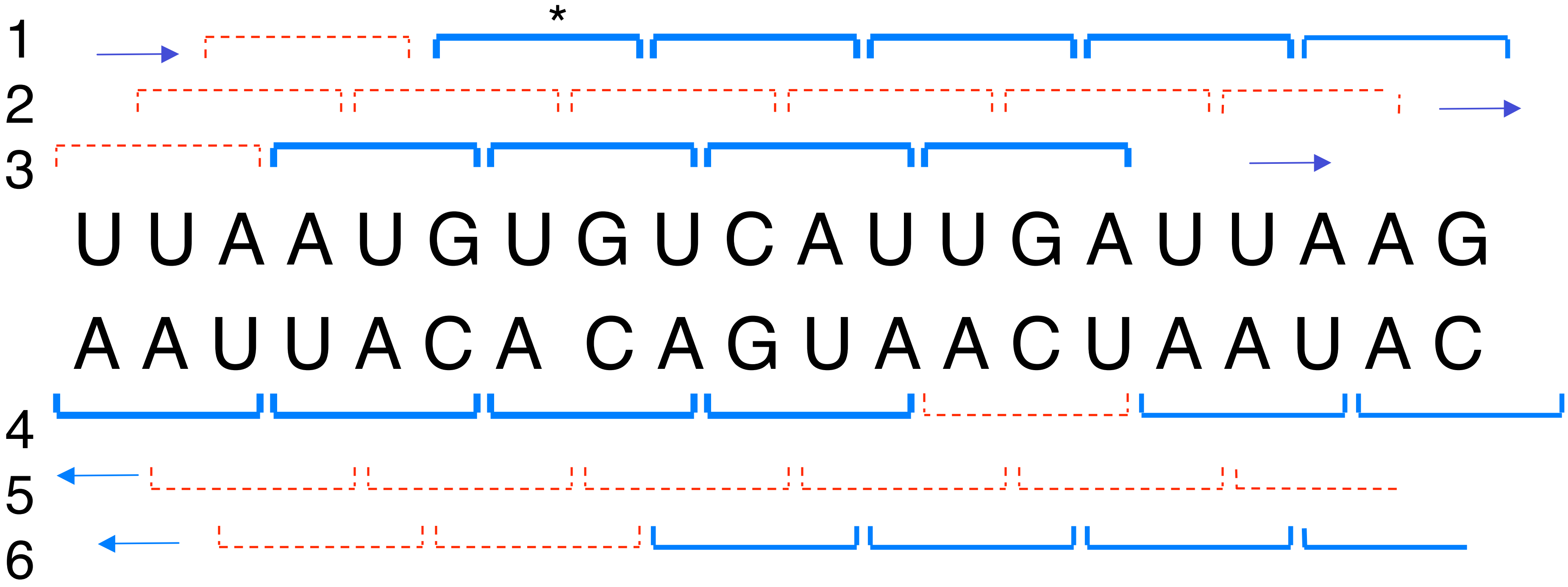
then continue scan for STOP

repeat until right end

repeat all starting at offset 1

repeat all starting at offset 2

Scanning for ORFs



1

2

3

*

U U A A U G U G U C A U U G A U U A A G

A A U U A C A C A G U A A C U A A U A C

4

5

6

* In bacteria, GUG is sometimes a start codon…

Idea #2: Codon Frequency

In random DNA

- Leucine : Alanine : Tryptophan = 6 : 4 : 1

But in real protein, ratios ~ 6.9 : 6.5 : 1

So, coding DNA is not random

Even more: synonym usage is biased (in a species dependant way)

- examples known with 90% AT 3rd base
  - Why? E.g. efficiency, histone, enhancer, splice interactions

Recognizing Codon Bias

Assume

- Codon usage i.i.d.; abc with freq. f(abc)
- $a_1a_2a_3a_4\ldots a_{3n+2}$ is coding, unknown frame

Calculate

- $p_1 = f(a_1a_2a_3)f(a_4a_5a_6)\ldots f(a_{3n-2}a_{3n-1}a_{3n})$
- $p_2 = f(a_2a_3a_4)f(a_5a_6a_7)\ldots f(a_{3n-1}a_{3n}\,a_{3n+1})$
- $p_3 = f(a_3a_4a_5)f(a_6a_7a_8)\ldots f(a_{3n}\,a_{3n+1}a_{3n+2})$
- $P_i = p_i / (p_1+p_1+p_3)$

More generally: k-th order Markov model

- k=5 or 6 is typical (next lecture)

Codon Usage in Φx174

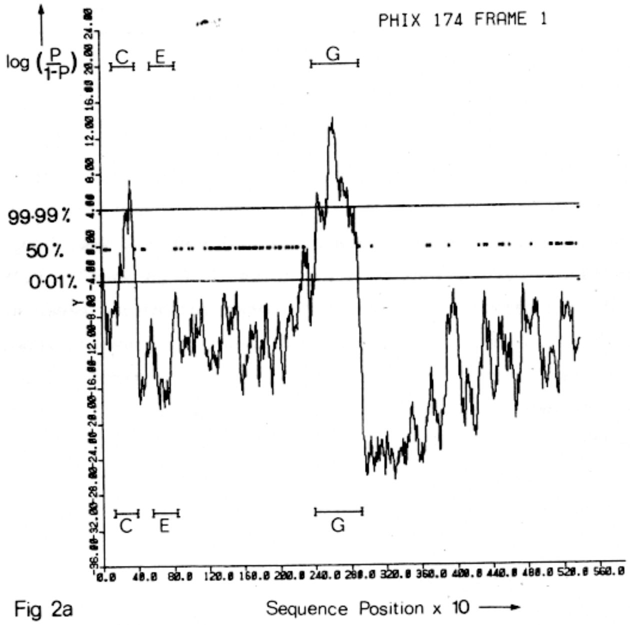


Fig 2a

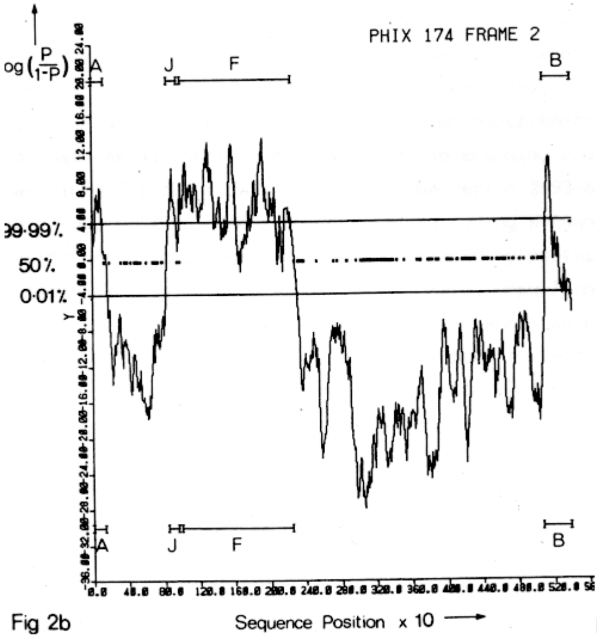


Fig 2b

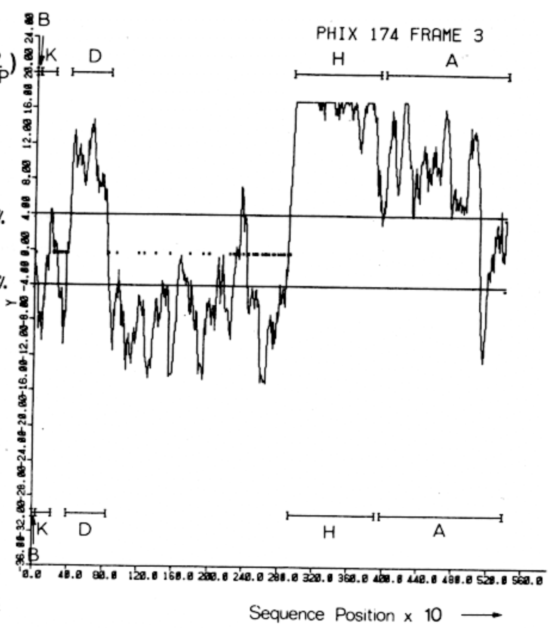


Fig 2c

Staden & McLachlan, NAR 10, 1 1982, 141-156