Lecture 15a
Computational Gene Finding

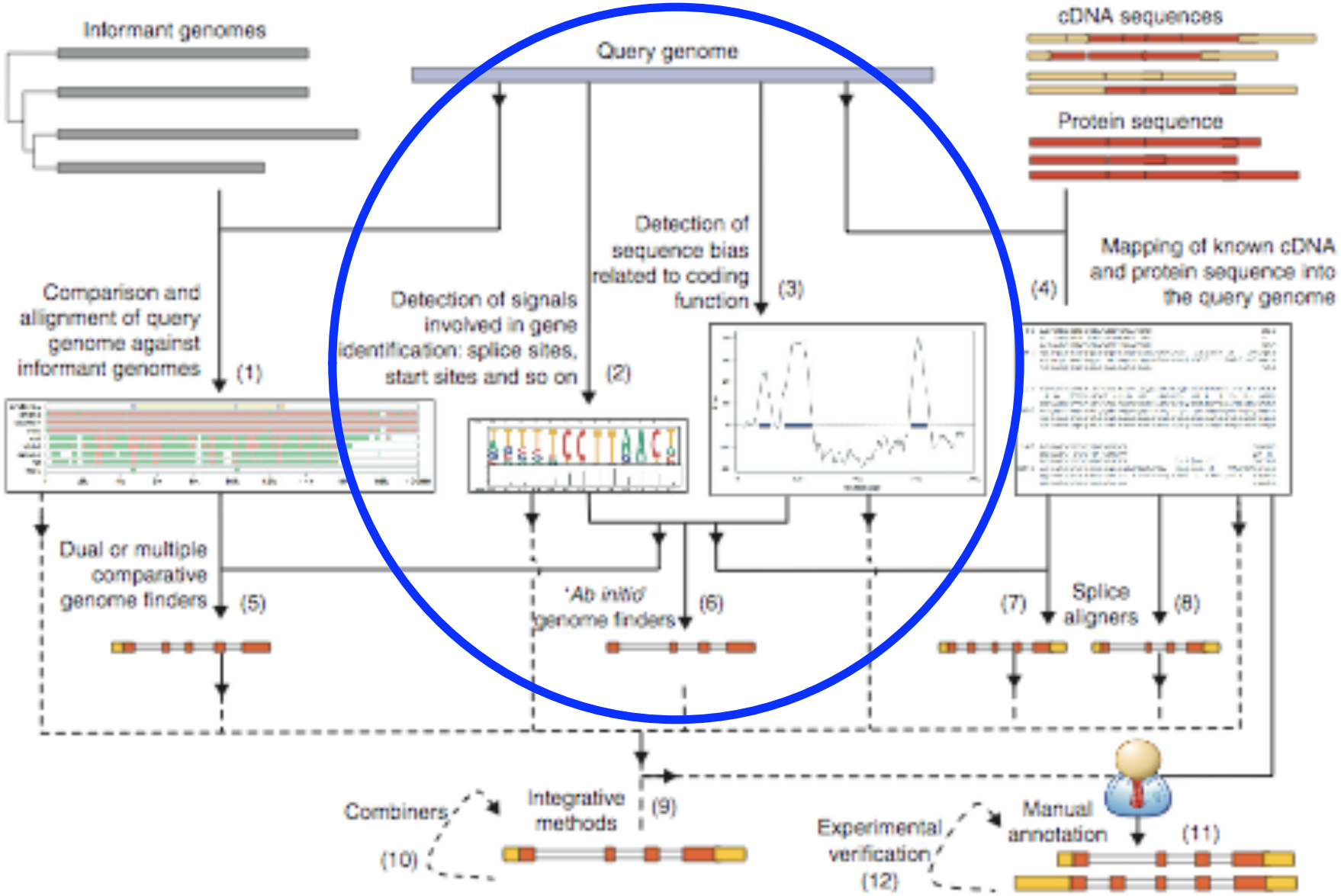
1 Minute Reflections

Good discussion of multiple sequence alignment. I never realized what a challenging problem this presents for computing.

I liked the multiple alignment material but some examples would be great.

I didn't really follow along with the details of progressive alignment.

Work up an approximate phylogenetic tree, aligning most similar groups first, more distant ones later. Sub-alignments are aligned with respect to each other but not internally re-adjusted. E.g., if you insert a gap somewhere, it goes in all rows for that subtree.


Informant genomes
Query genome
cDNA sequences
Protein sequence
Comparison and alignment of query genome against informant genomes
(1)
Detection of signals involved in gene identification: splice sites, start sites and so on
(2)
Detection of sequence bias related to coding function
(3)
(4)
Mapping of known cDNA and protein sequence into the query genome
Dual or multiple comparative genome finders
(5)
'Ab initio' genome finders
(6)
(7)
Splice aligners
(8)
Combiners
(10)
Integrative methods
(9)
Experimental verification
(12)
Manual annotation
(11)

Codons & The Genetic Code

First Base	Second Base: U	Second Base: C	Second Base: A	Second Base: G	Third Base
U	Phe	Ser	Tyr	Cys	U
	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

Idea #1: Find Long ORF's

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

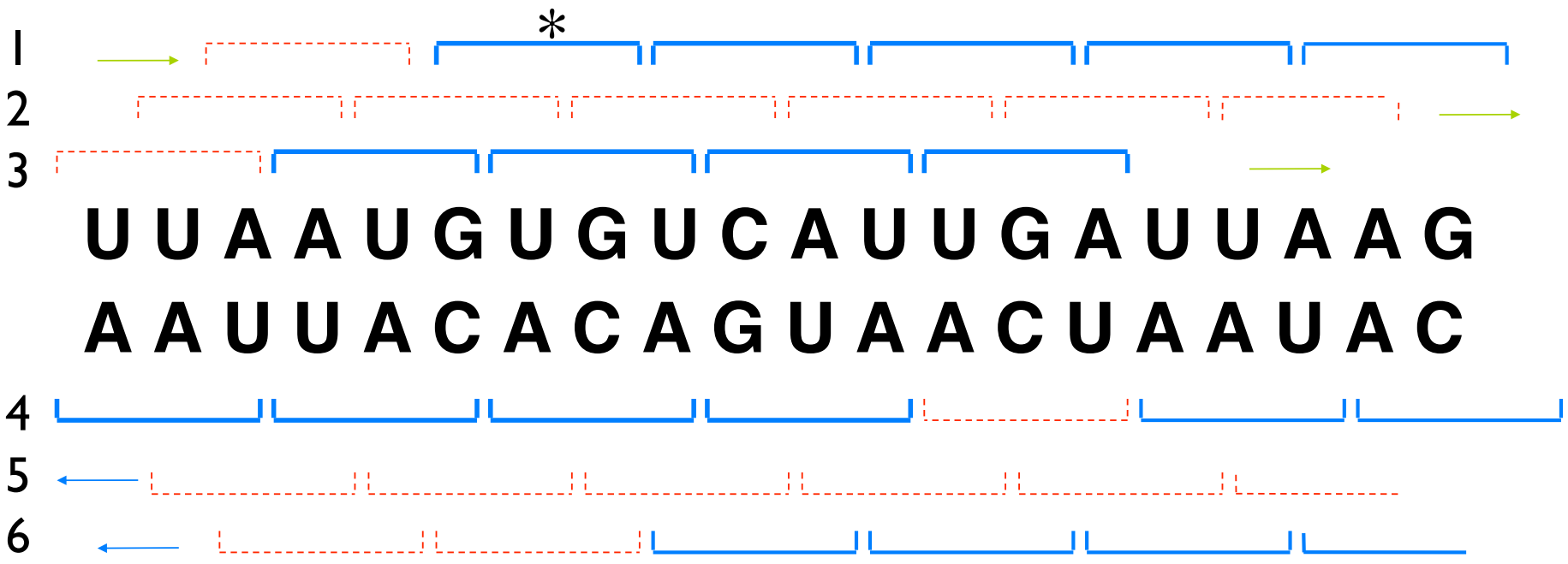
Open Reading Frame: No stop codons

In random DNA

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

But average protein ~ 1000bp

Scanning for ORFs



* 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, since significant influences spanning codons are detectable

Codon Usage in Φx174

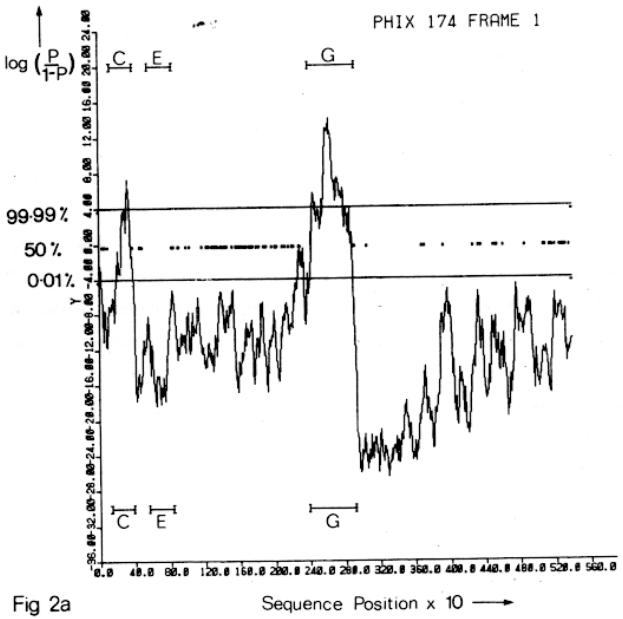


Fig 2a

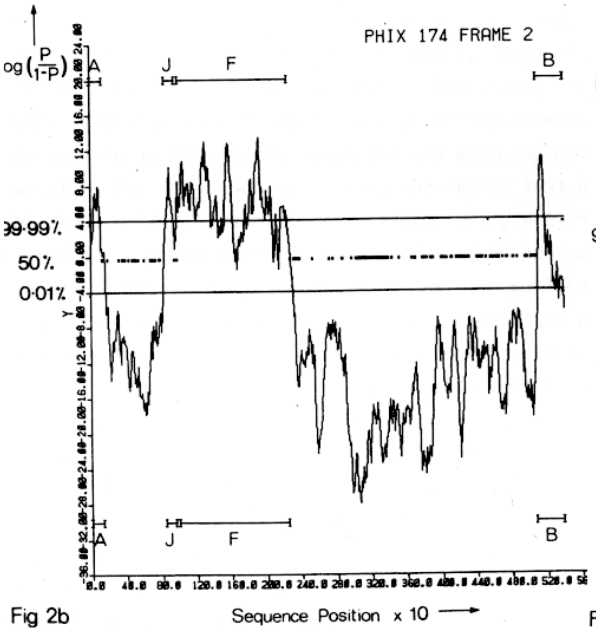


Fig 2b

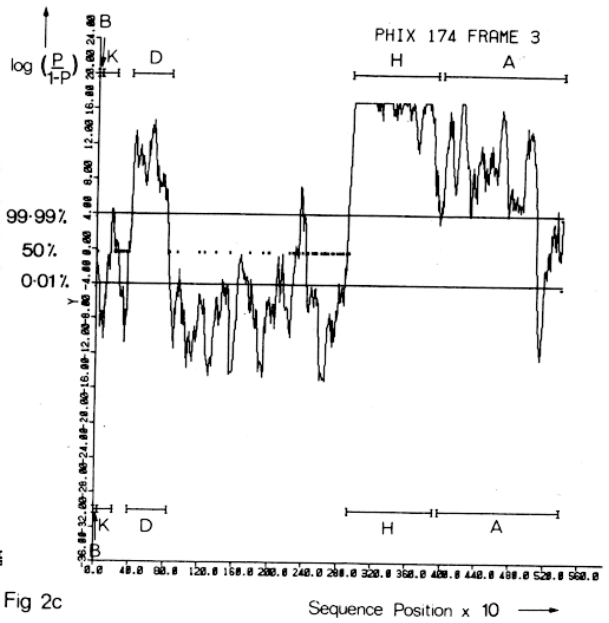


Fig 2c

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

Better: Markov Models

Can always represent a joint probability distribution

$$P(x) = P(x_1)\, P(x_2 \mid x_1)\, P(x_3 \mid x_1 x_2) \ldots P(x_n \mid x_1 x_2 \ldots x_{n-3} x_{n-2} x_{n-1})$$

If each letter only depends on the k previous ones, it's a "k-th order Markov model." E.g., k=3:

$$P(x) = P(x_1)\, P(x_2 \mid x_1)\, P(x_3 \mid x_1 x_2)\, P(x_4 \mid x_1 x_2 x_3)\, P(x_5 \mid x_2 x_3 x_4) \ldots P(x_n \mid x_{n-3} x_{n-2} x_{n-1})$$

Idea: distant influences fade

For “gene finding”

Given:

P(- | -) for known genes, vs

Q(- | -) for background,

again can look at likelihood ratio

P/Q (or log(P/Q))

that given sequence comes from the “gene” model vs the “background” model.

Report high scores.

Summary

Computational gene prediction relies on statistical properties observed in protein coding genes that differ from random DNA sequences, e.g.

- long ORFS
- codon-usage- or other baises

Often use k^{th}-order Markov models, $k \approx 6$

(Noncoding genes behave differently.)