

Computational Biology

Biology 23
August 7, 2001

DNA and Protein Databases

- GenBank - run by US National Library of Medicine
- EMBL - European Molecular Biology Laboratory
- DDBJ - DNA Database of Japan
- PIR and SwissProt - protein databases
- GenProt - translated sequences from GenBank

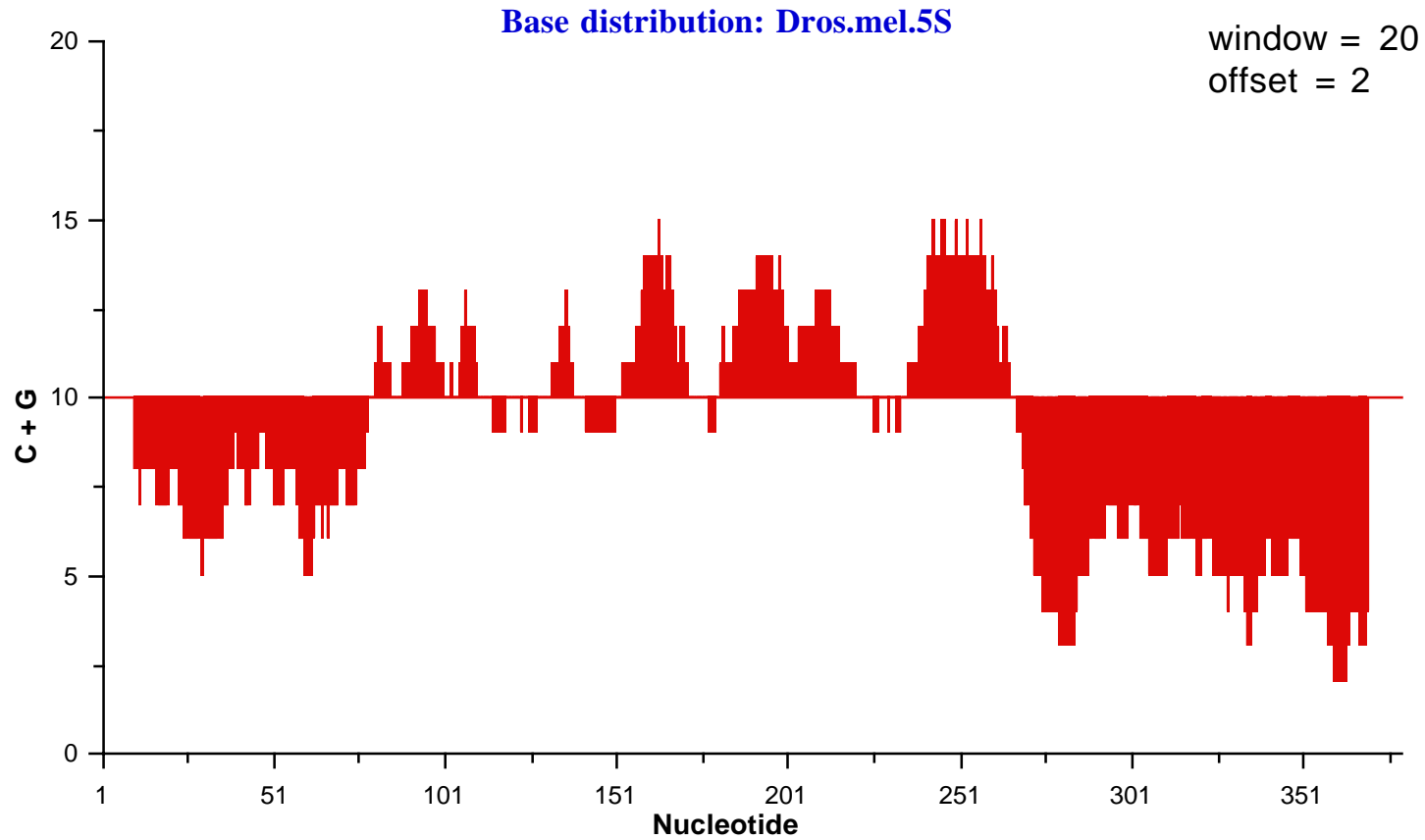
Other Kinds of Databases

- Brookhaven National Labs structure database
- Entré́z - DNA, protein sequences and associated literature
- TFD - transcription factor database (TFs, motifs, and binding sites on DNA)
- Prosite - protein motifs
- Flybase - Drosophila genes, products, protocols, literature
- Chromosome/Genome databases

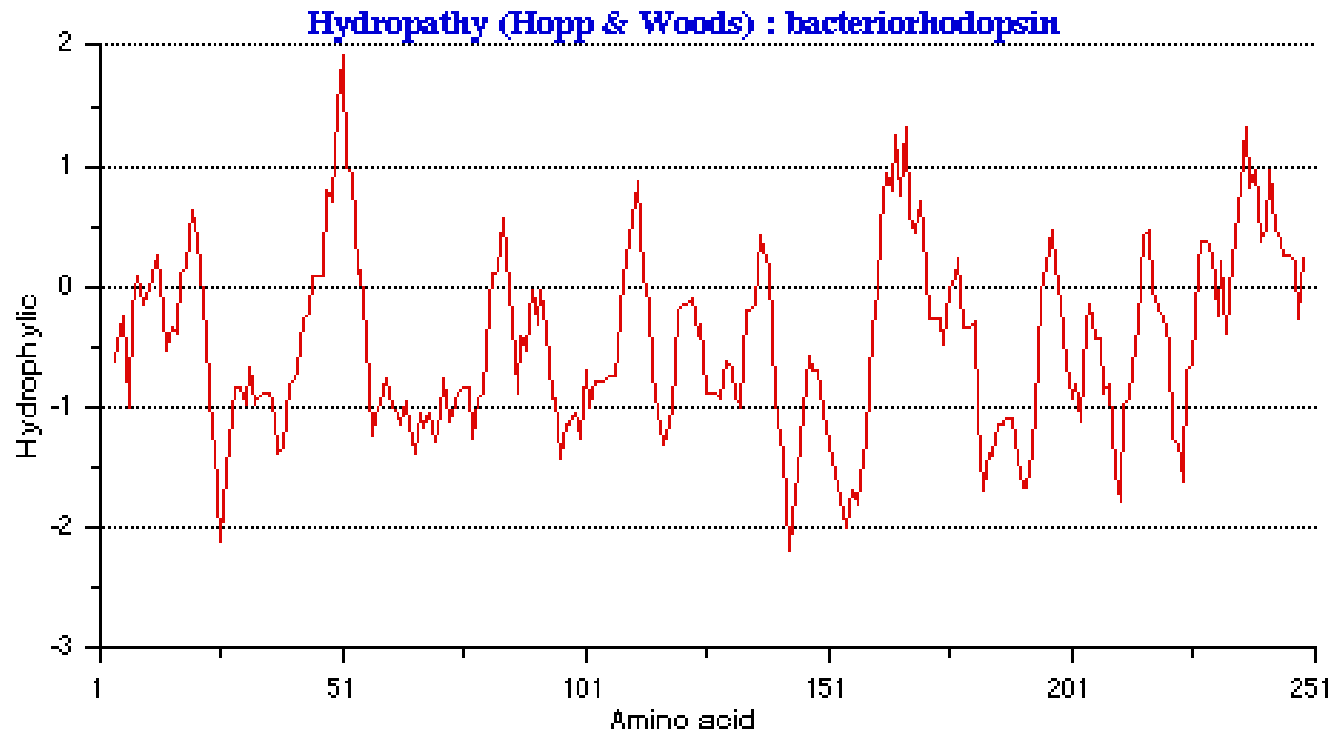
Algorithm

a finite sequence of well-defined actions whose purpose is to accomplish a given task

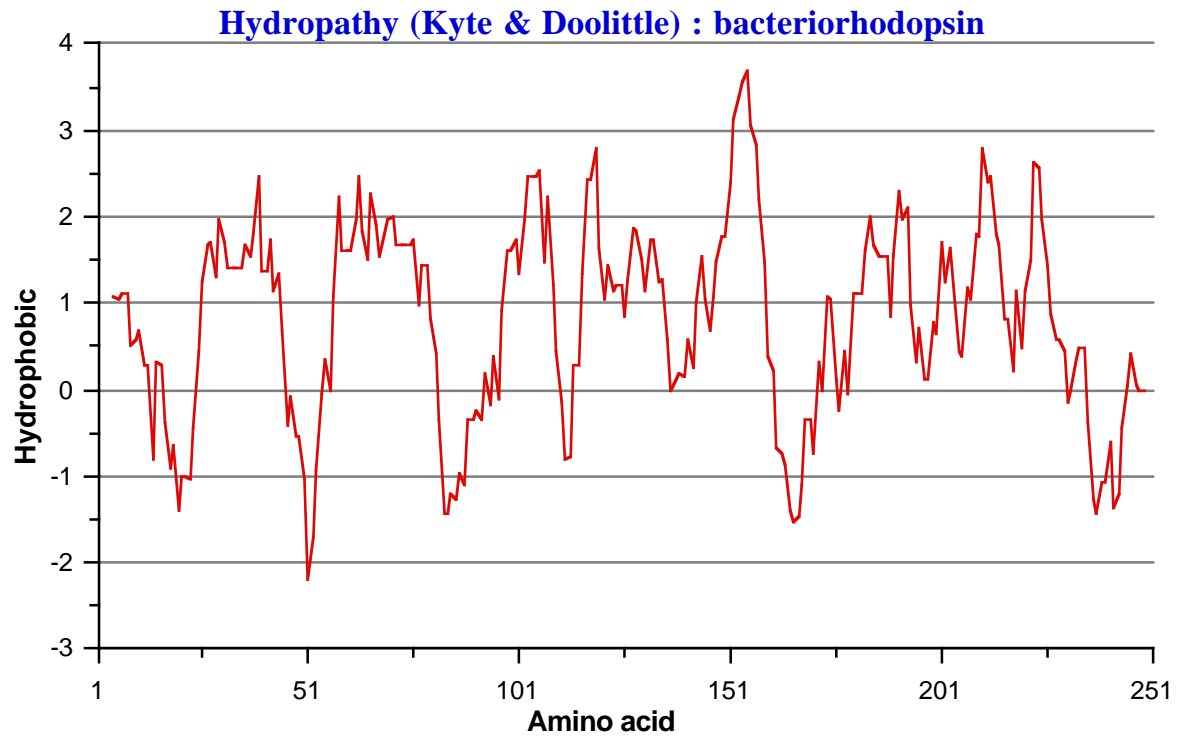
Sliding Window Base Composition



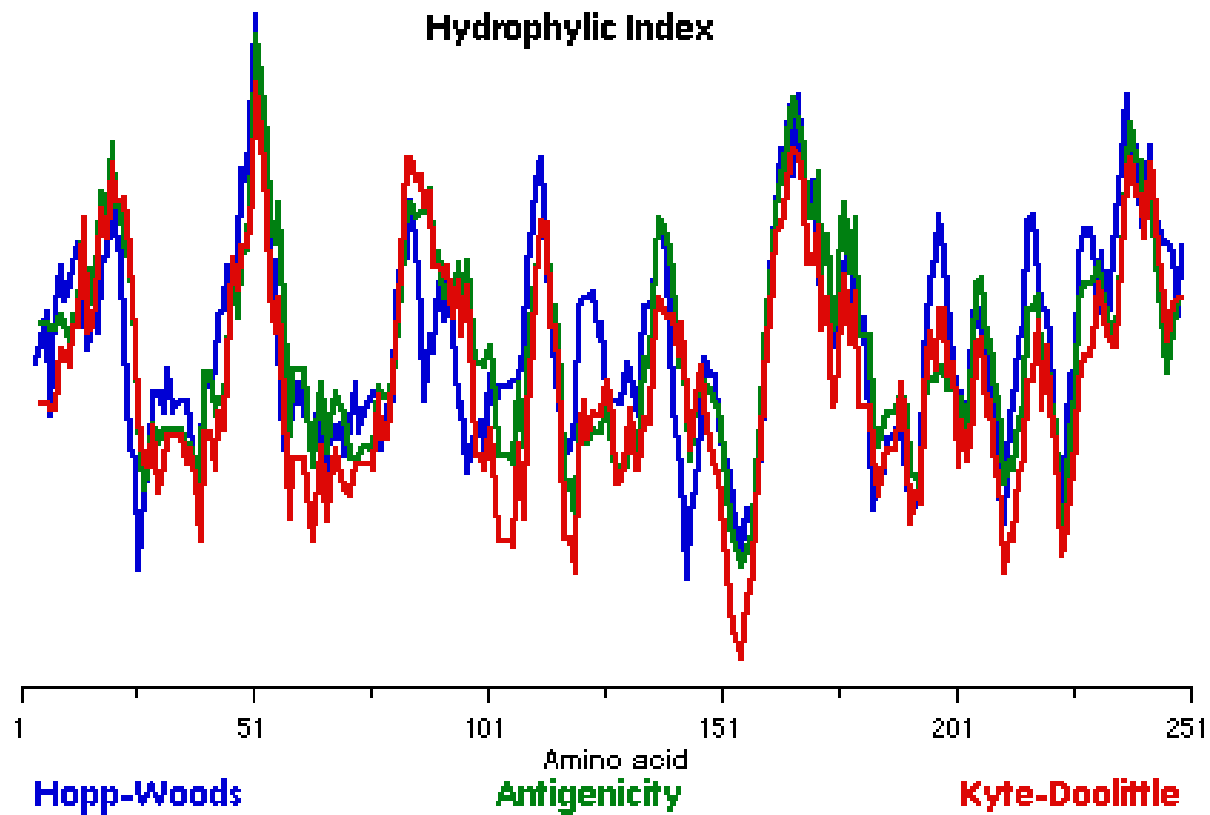
Hydropathy (Hopp-Woods)



Hyrdopathy (Kyte-Doolittle)

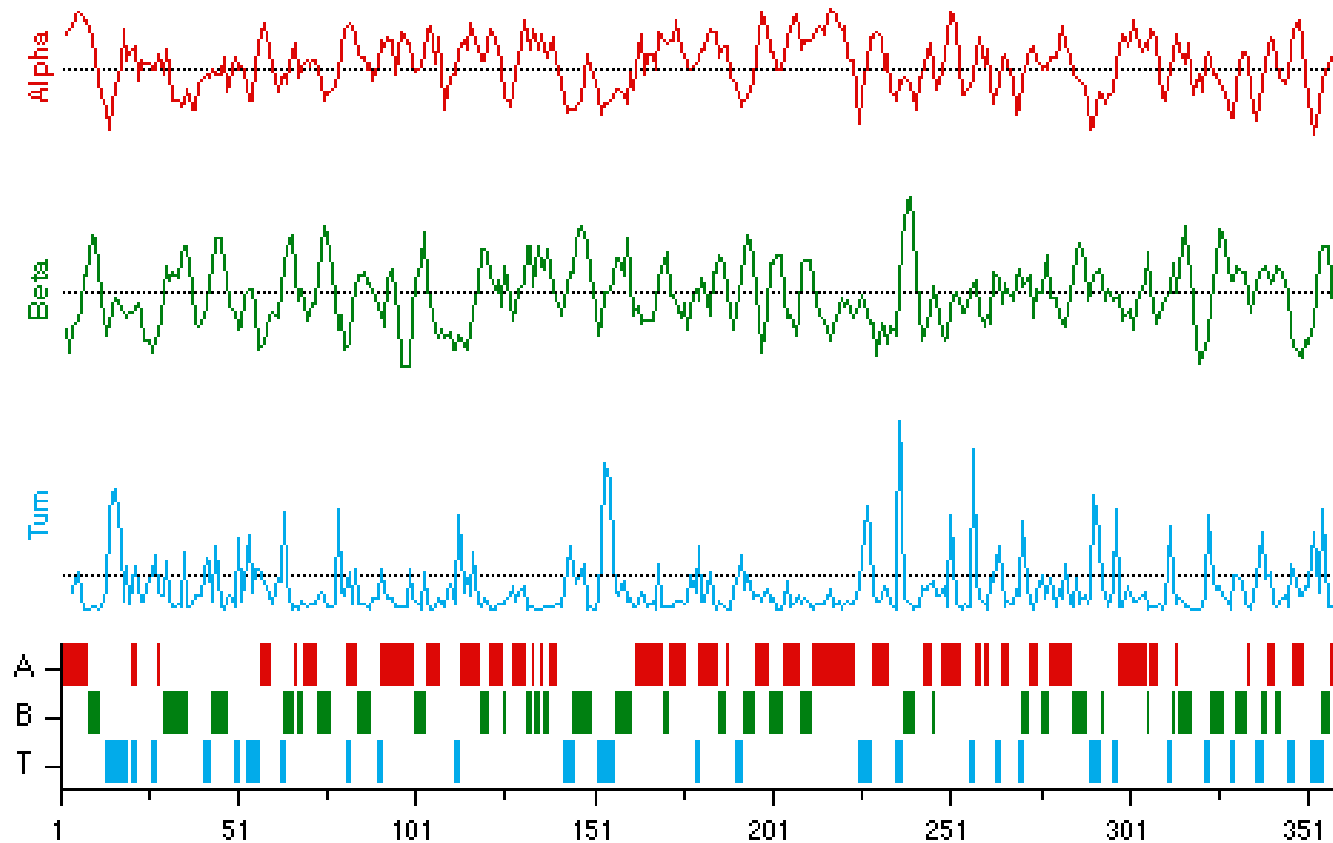


Comparison of Hydropathies



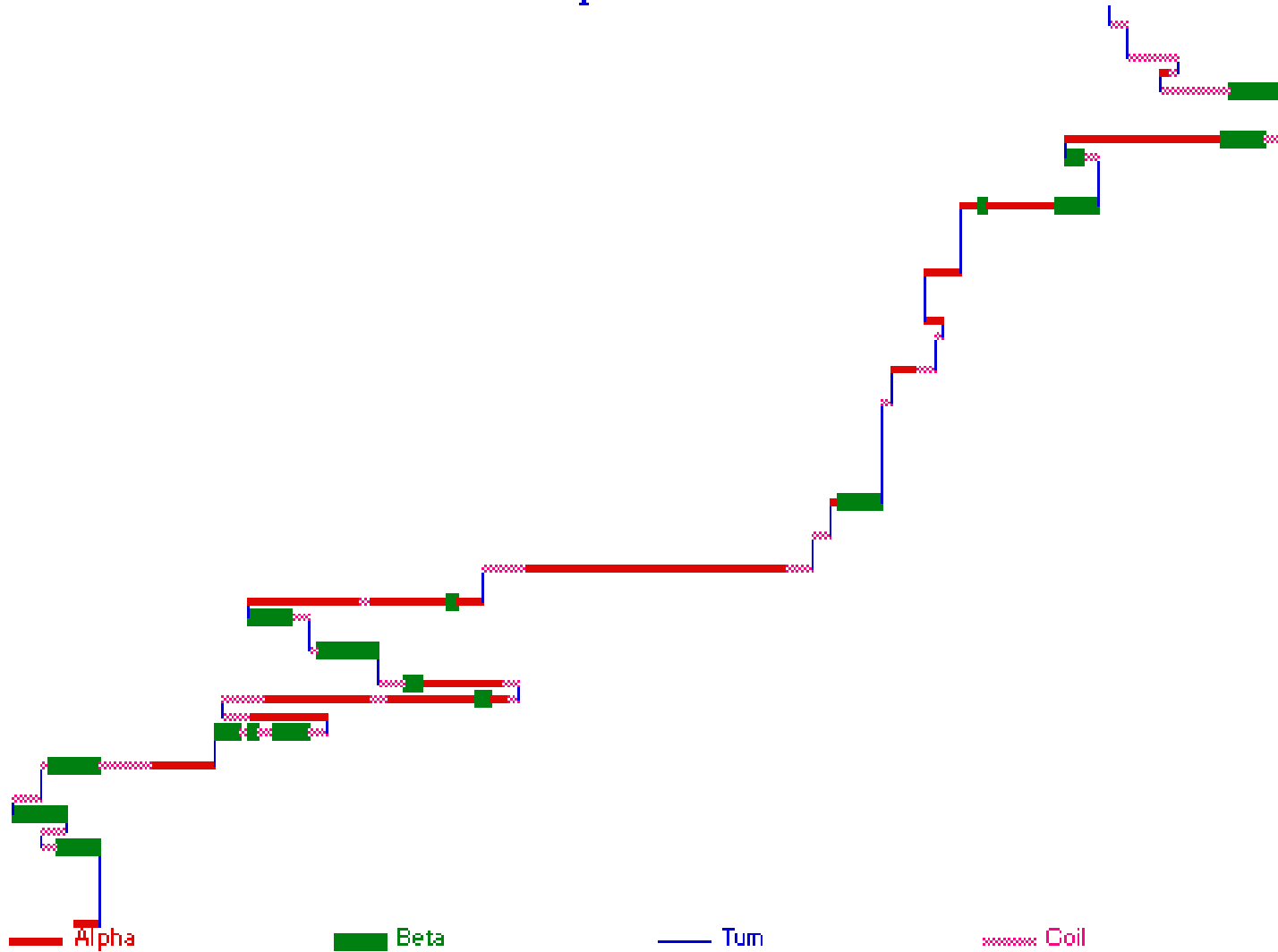
CF Structure Prediction (graph)

CF structure prediction: Dros actin 5C

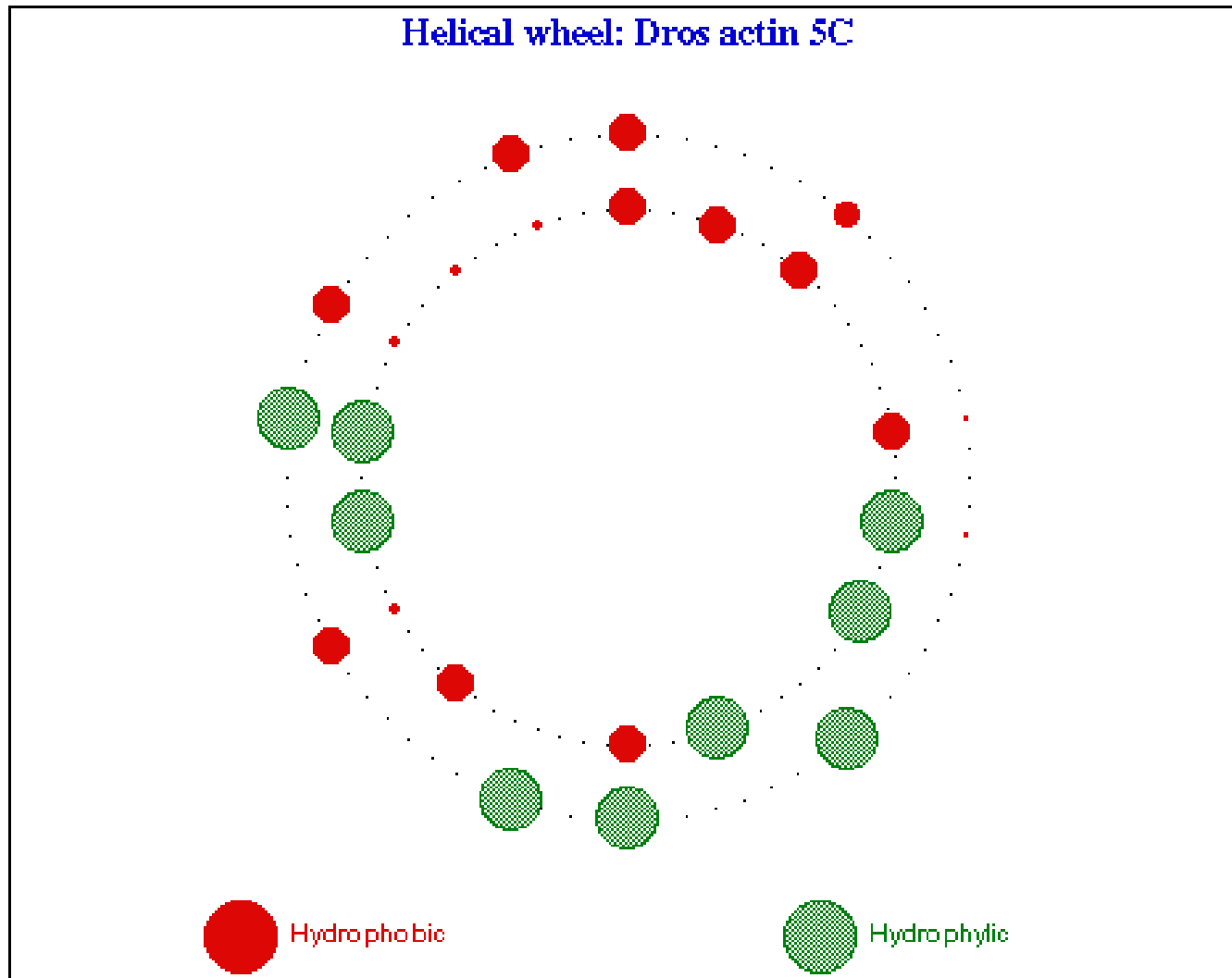


GOR Structure Prediction (Squiggles)

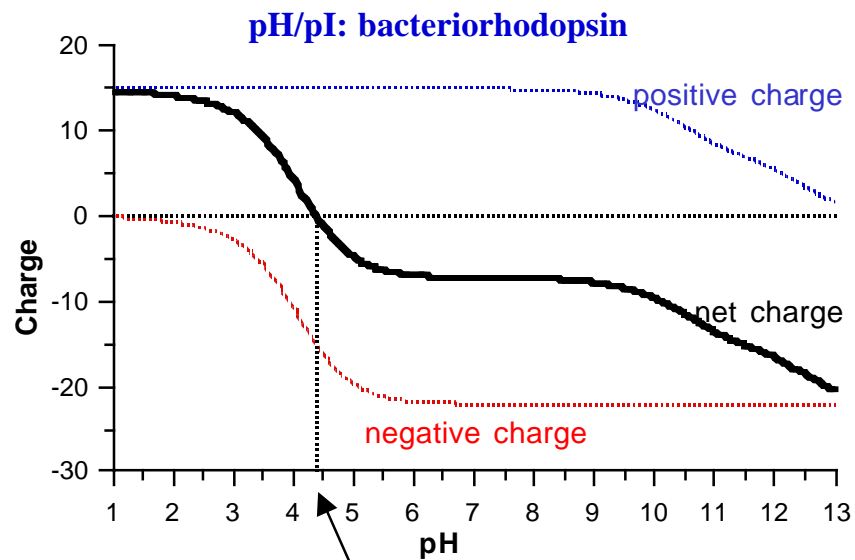
GOR structure prediction: Dros actin 5C



Helical Wheel



Physical Characteristics



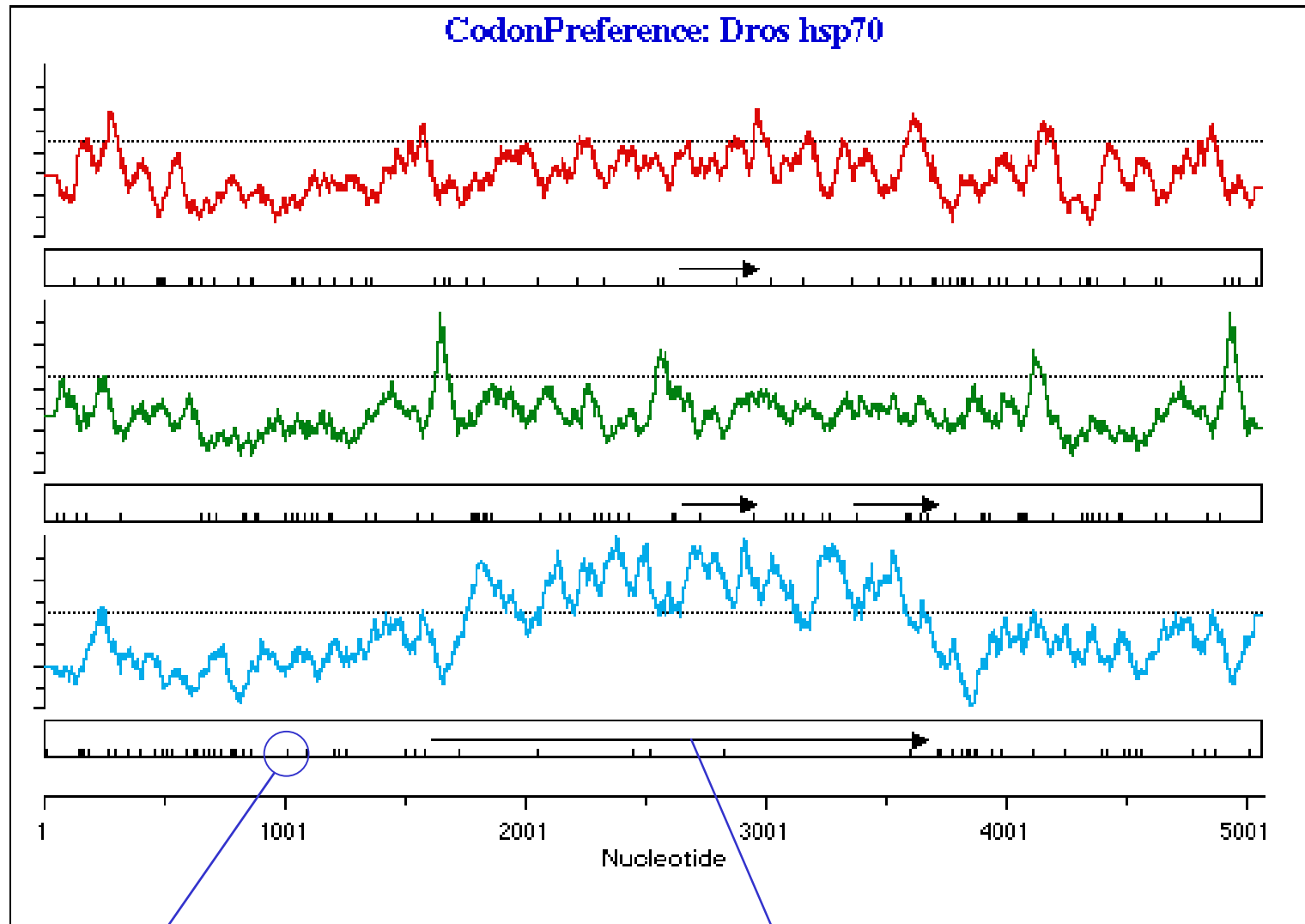
Physical characteristics: bacteriorhodopsin	
	value
molar extinction coefficient	44472.0 A(280 nm)/mole
length	251 amino acids
molecular weight	26462.4 daltons
isoelectric point	4.38
net charge at pH = 7	-6.98
1 microgram =	37.8 picomoles
1 A(280 nm) =	0.6 mg/ml

Codon Usage Tables

Table 1: Some Codon Preference Numbers

Amino acid	codon	E. coli	Arabidopsis	Drosophila	human
Gly	GGG	.13	.10	.09	.23
Gly	GGA	.09	.37	.30	.26
Gly	GGU	.38	.39	.22	.18
Gly	GGC	.40	.14	.40	.33
Pro	CCG	.55	.16	.28	.11
Pro	CCA	.20	.33	.26	.27
Pro	CCU	.16	.34	.15	.29
Pro	CCC	.10	.16	.30	.33

Codon Preference Analysis

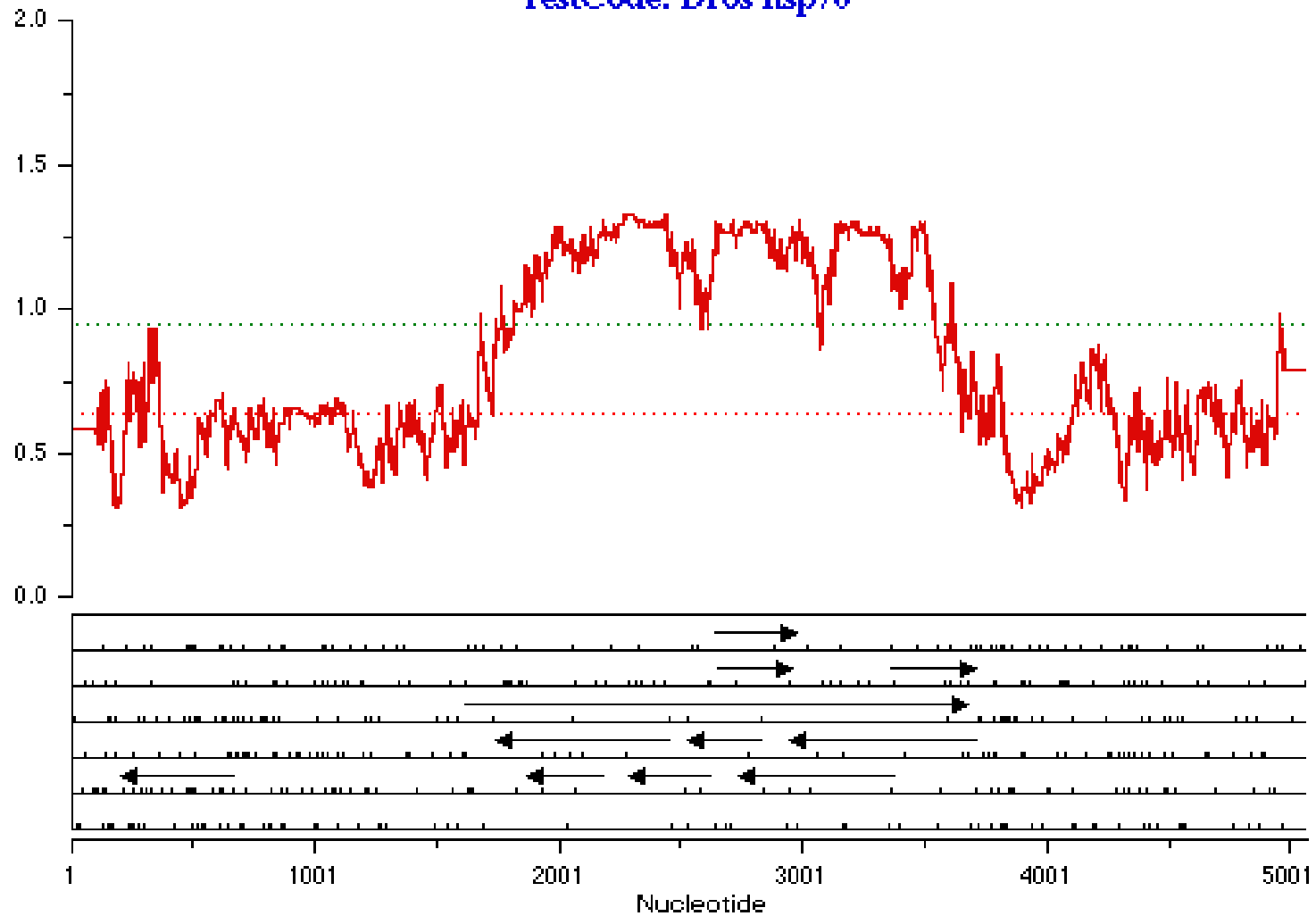


rare codons

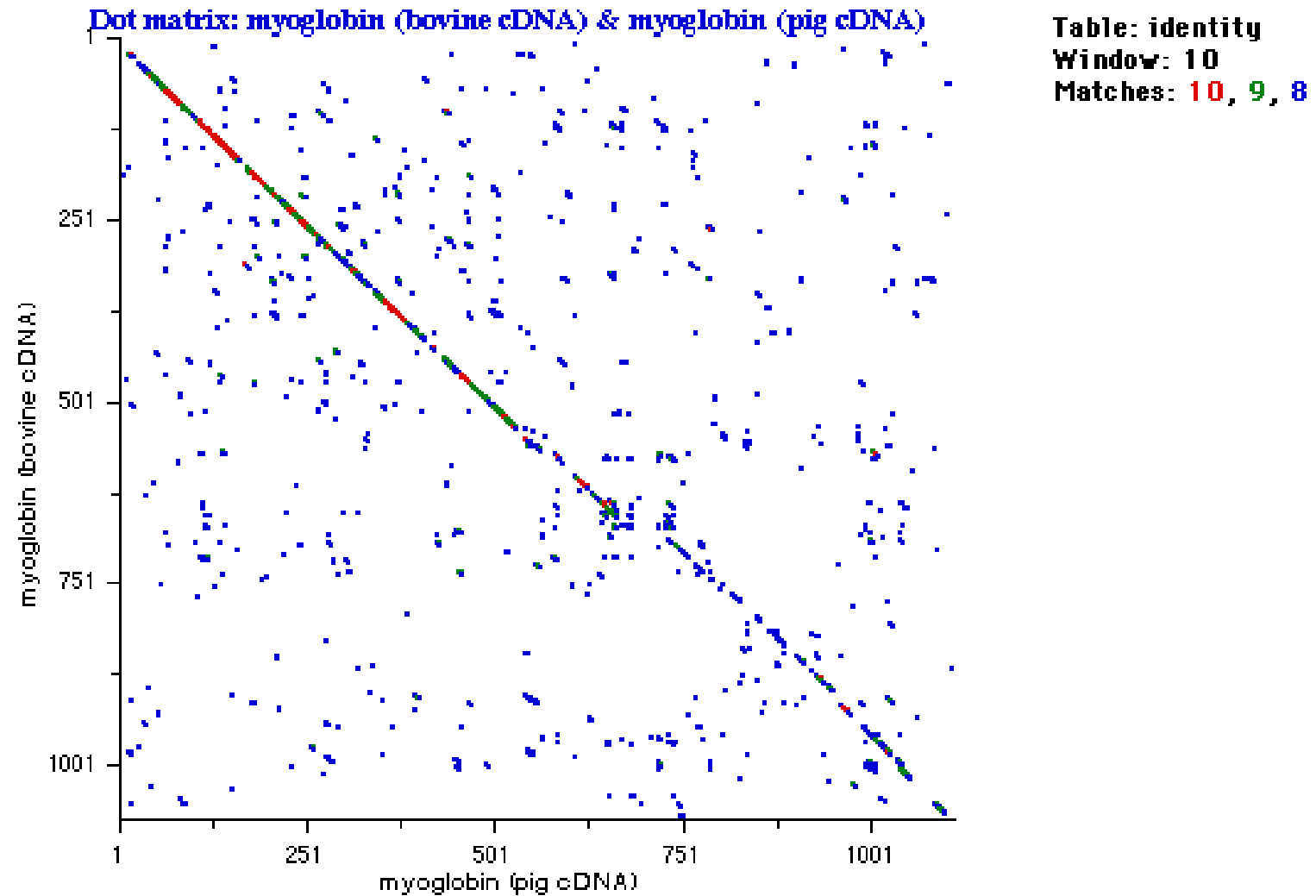
open reading frames (ORFs)

TestCode Analysis

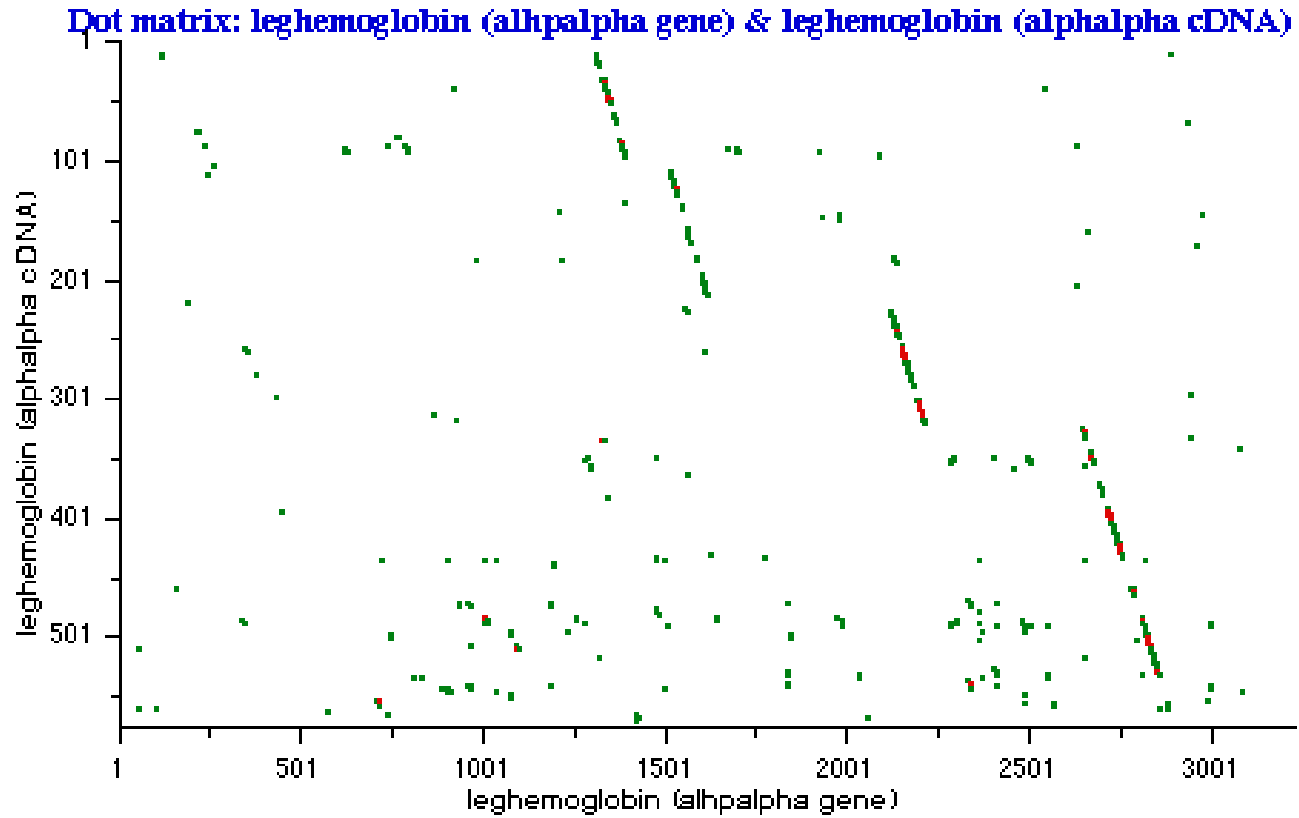
TestCode: Dros hsp70



Dot Matrix (DNA - Identity Table)

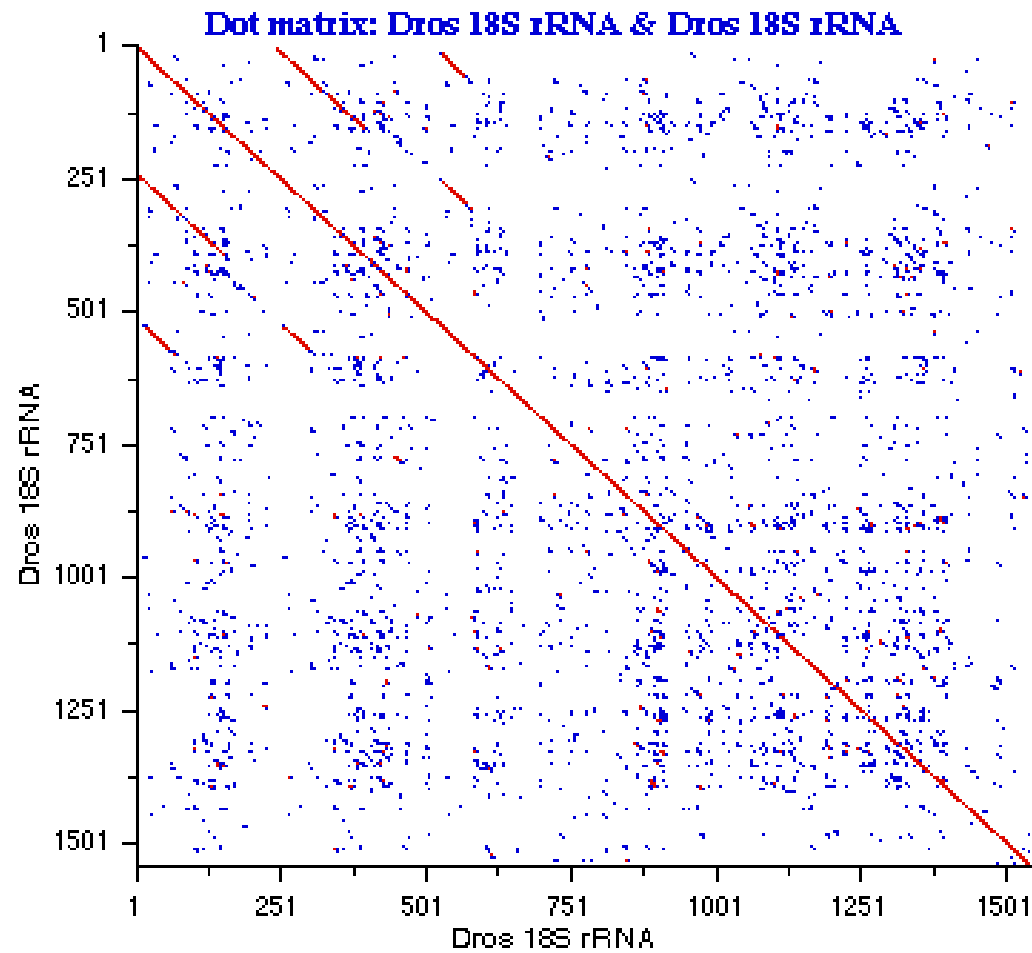


Dot Matrix - cDNA vs Genomic DNA



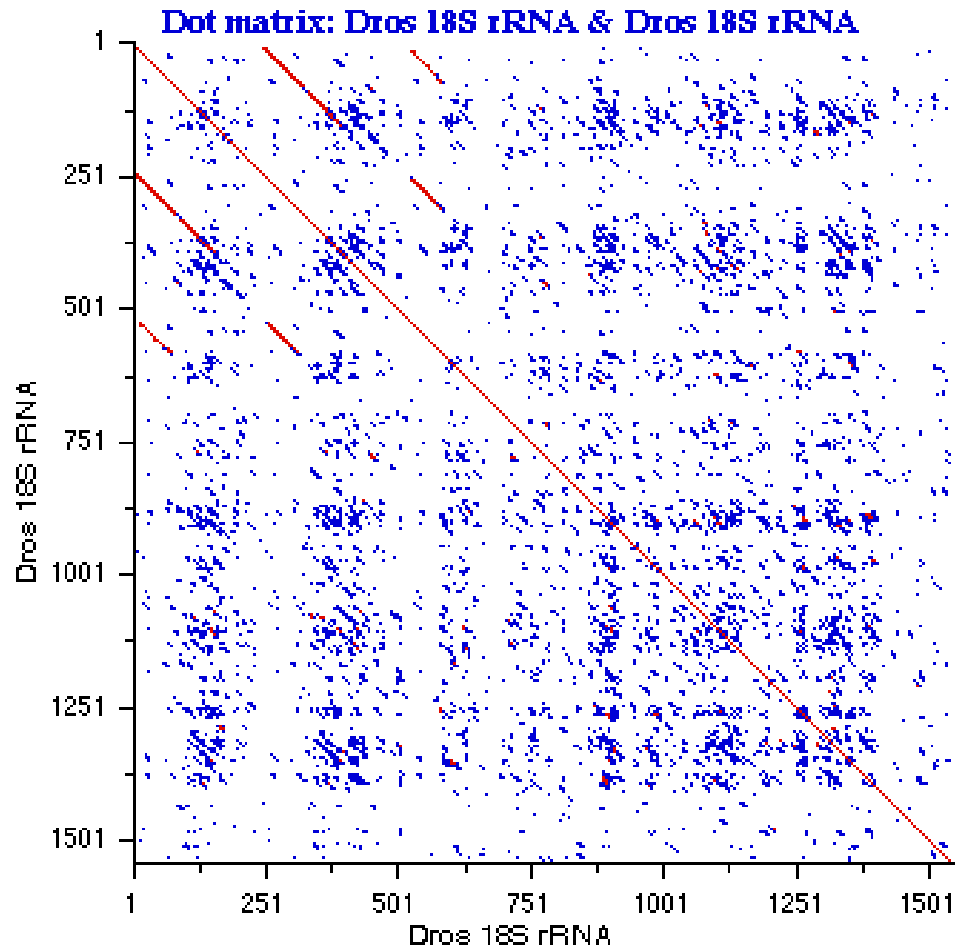
Ribosomal DNA Dot Matrix#1

Window = 10
 ≤ 1 mismatch
2 mismatches

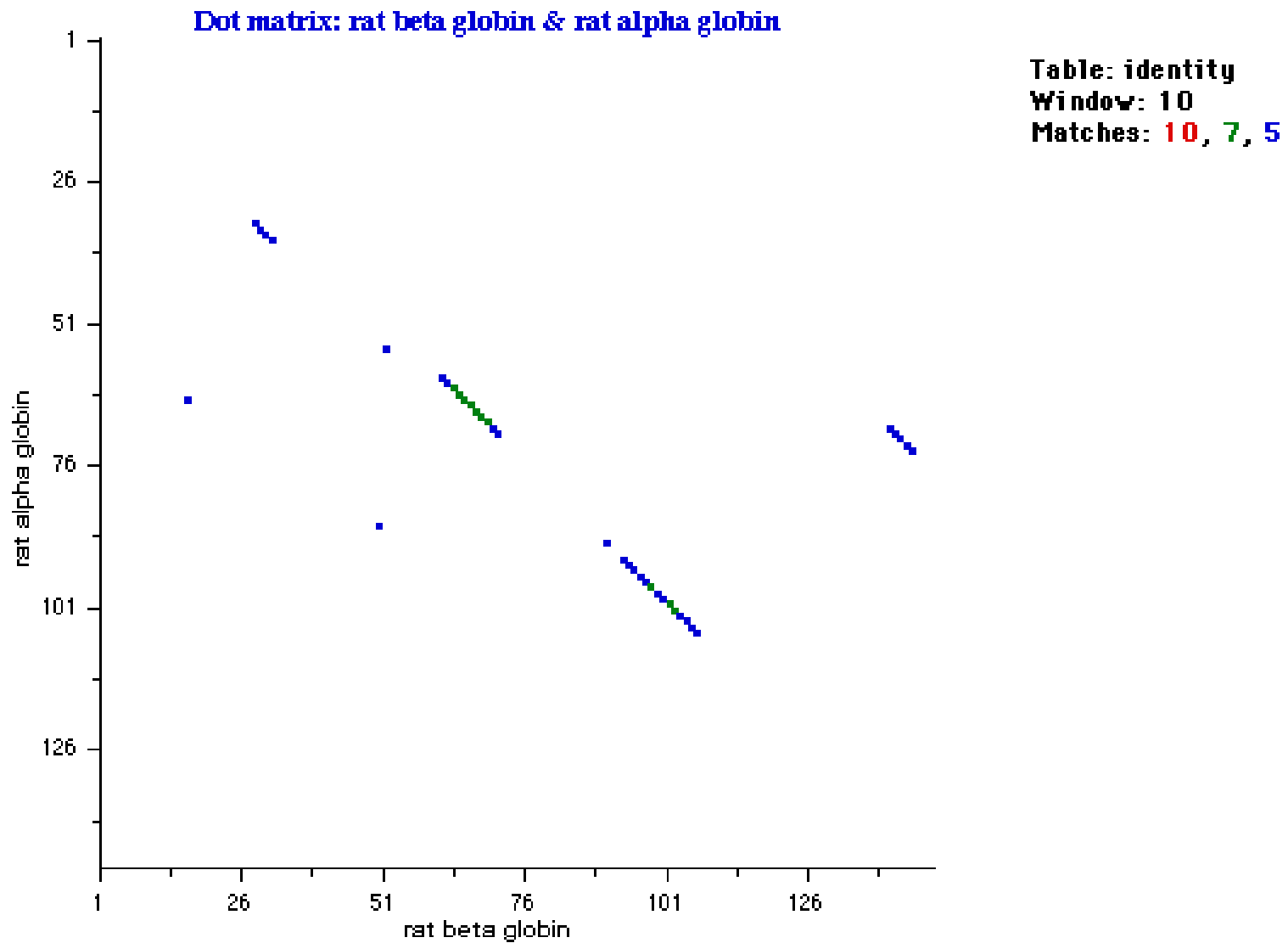


Ribosomal DNA Dot Matrix#2

Window = 15
3 mismatches
5 mismatches



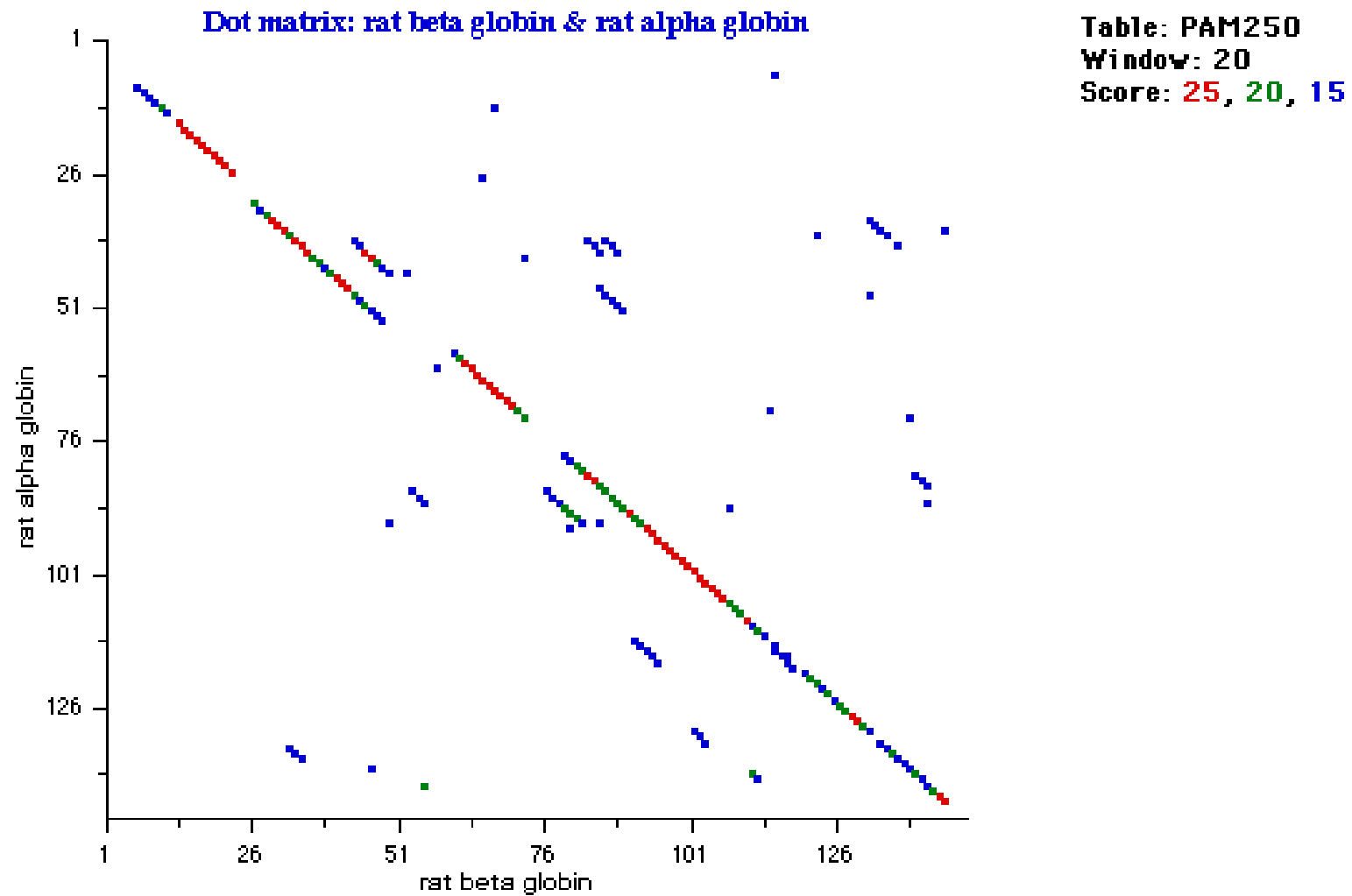
Dot Matrix (Protein - Identity Table)



PAM250 Table

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	X
Ala	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	0
Arg	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-1
Asn	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	0
Asp	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	-1
Cys	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-3
Gln	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	-1
Glu	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	-1
Gly	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	-1
His	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	-1
Ile	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-1
Leu	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-1
Lys	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	-1
Met	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-1
Phe	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-2
Pro	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	-1
Ser	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	0
Thr	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0
Trp	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-4
Tyr	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-2
Val	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-1
X	0	-1	0	-1	-3	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	0	0	-4	-2	-1	0

Dot Matrix (Protein - PAM250 Table)



Consider These Sequences

ACGGTCGAAT (a)

ACGGACGAAT (b)

ACGGTTCGAAT (c)

ACGGTTTCGAAT (d)

CGCGACGGTCGAATAT (e)

Alignment of **a** vs **a**

ACGGTCGAAT (a)
| | | | | | | |
ACGGTCGAAT (a)

what is the score?

Alignment of **a** vs **b**

ACGGTCGAAT	(a)
ACGG <u>A</u> CGAAT	(b)

subtract for mismatch? or just not add?

Alignment of **a** vs **c**

ACGGTCGAAT (a)

| | | | | |

ACGGTTCGAAT (c)

Gapped Alignment of **a** vs **c**

ACGGT-CGAAT (a)

| | | | | | | |

ACGGTTCGAAT (c)

*gap penalty? subtract?
relative to mismatch penalty?*

Gapped Alignment of **a** vs **d**

ACGGT--CGAAT (a)

| | | | | | | |

ACGGTTTCGAAT (d)

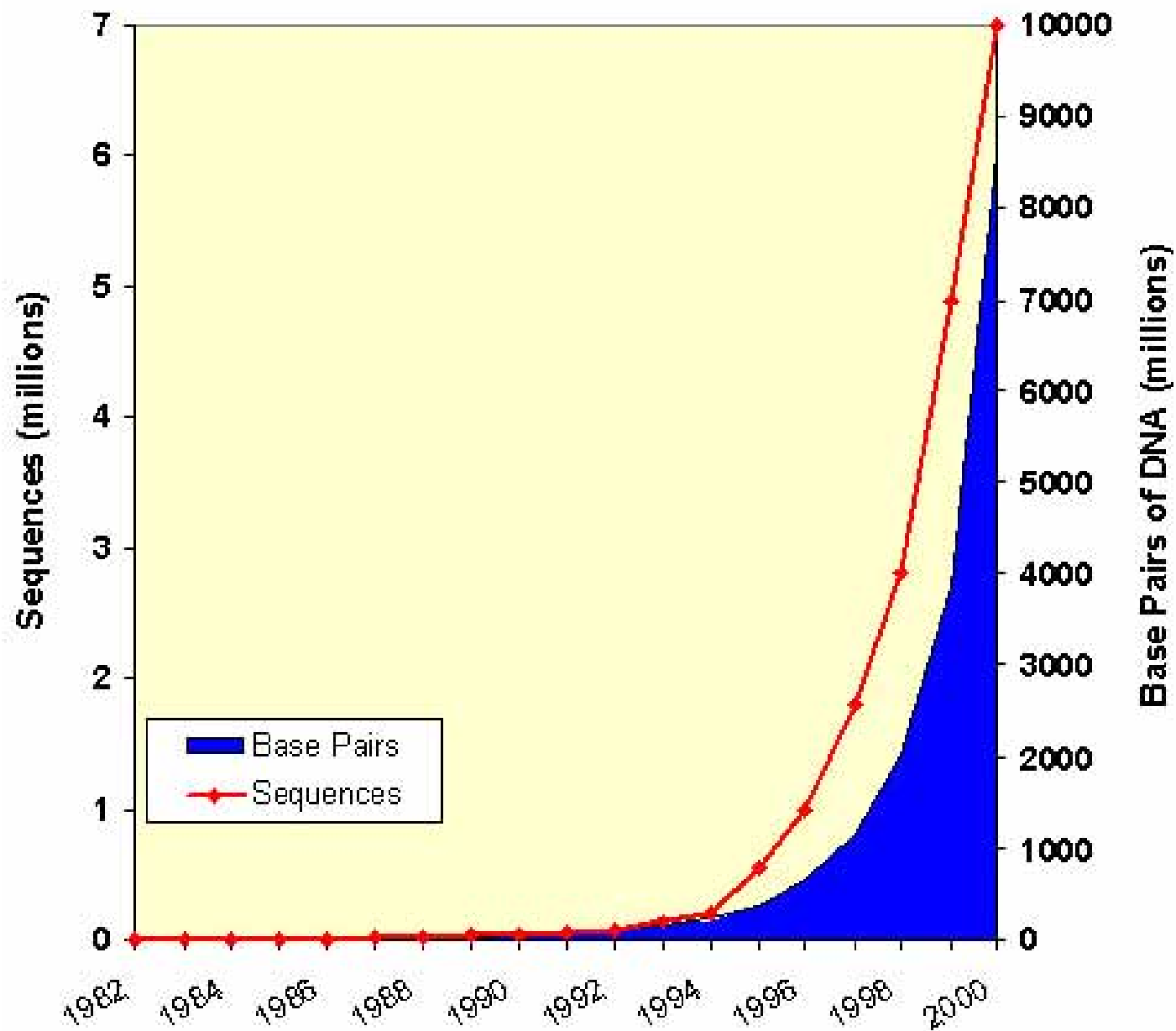
gap length penalty?

Gapped Alignment of **a** vs **e**

-----ACGGTCGAAT-- (a)
 | | | | | | | |
CGCGACGGTCGAATAT (e)

*end gaps?
same as gap=6 internally?*

GenBank Database Growth



BLAST Search Using HOX3

[Distribution of 512 Blast Hits on the Query Sequence](#)

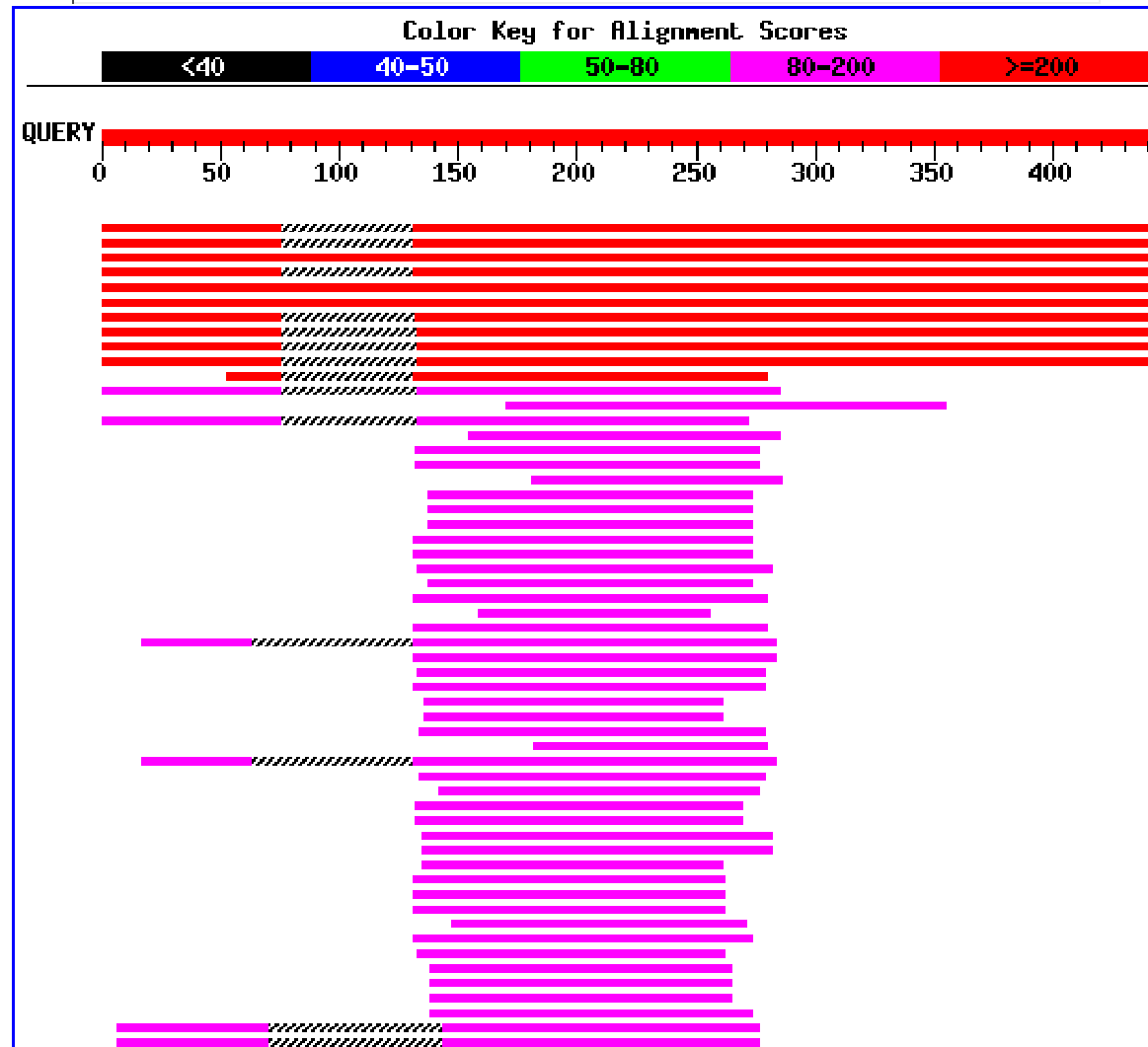
Mouse-over to show defline and scores. Click to show alignments



PSI-BLAST Search with HOX3

[Distribution of 114 Blast Hits on the Query Sequence](#)

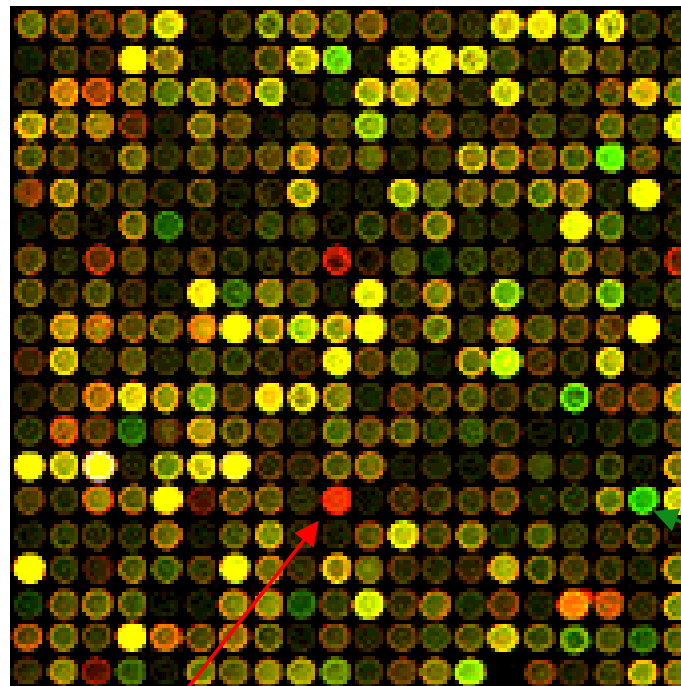
Mouse-over to show defline and scores. Click to show alignments



Finding Genes

- **A number of different algorithms have been developed**
- **All depend on identified physical characteristics of gene sequences or on other properties categorized in Markov models**
- **Some important properties**
 - **Splice junction sequences**
 - **Branch point sequences**
 - **Promoter sequences**
 - **Codon usage in exons**
 - **5' capping sequence**
 - **Poly(A) sites**
 - **Similarity to other known genes**

DNA MicroArray Expression Data



treated

untreated

Effects of Centering and Normalizing Data Manipulation on Clustering Results

