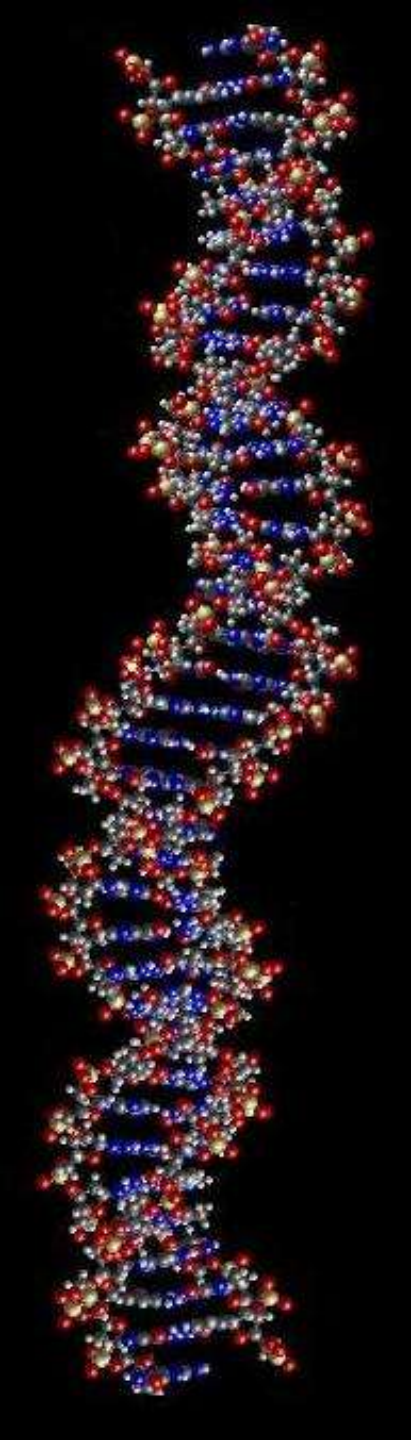
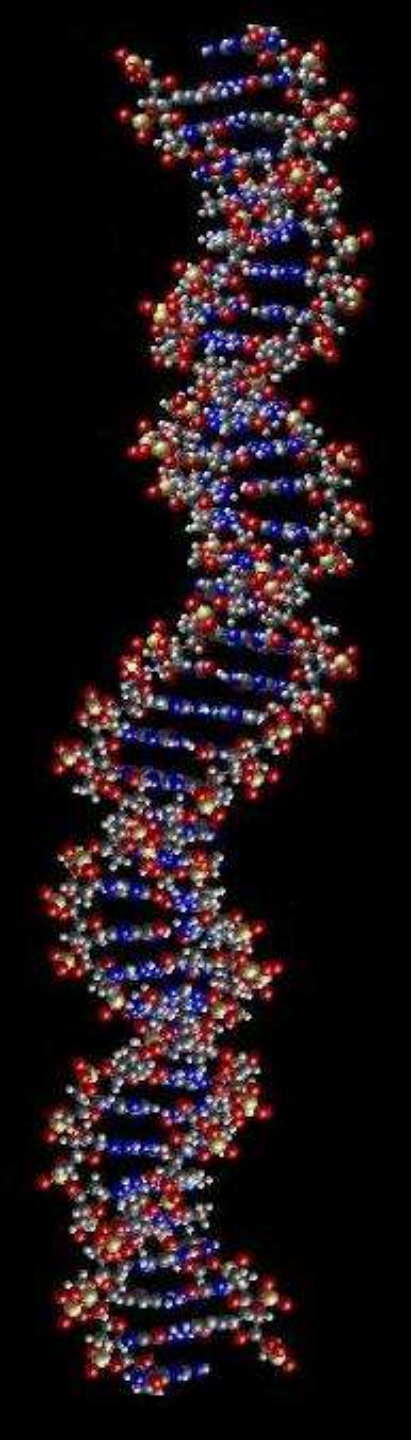


An Evaluation of Intron Significance Using Bioinformatics

Kelsey Byers

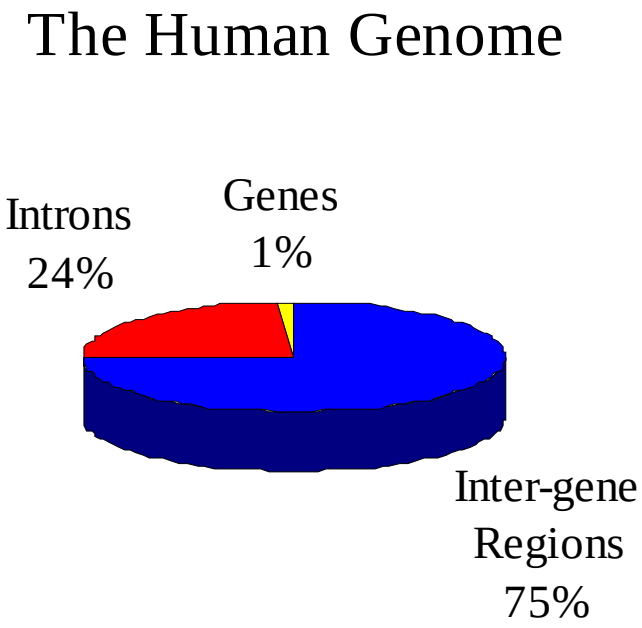
Biotechnology Academy

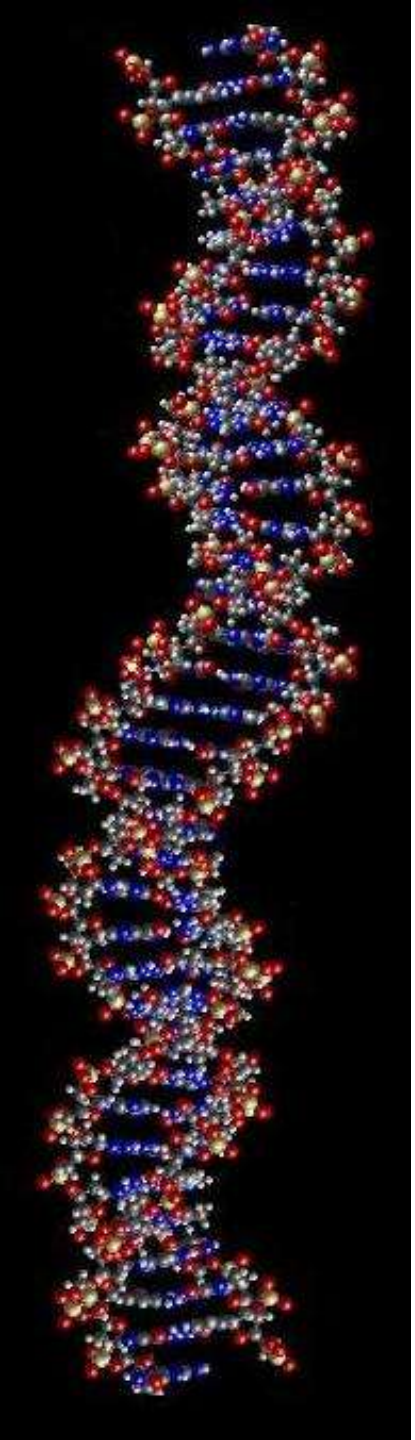
April 17, 2003



What Is “Junk DNA?”

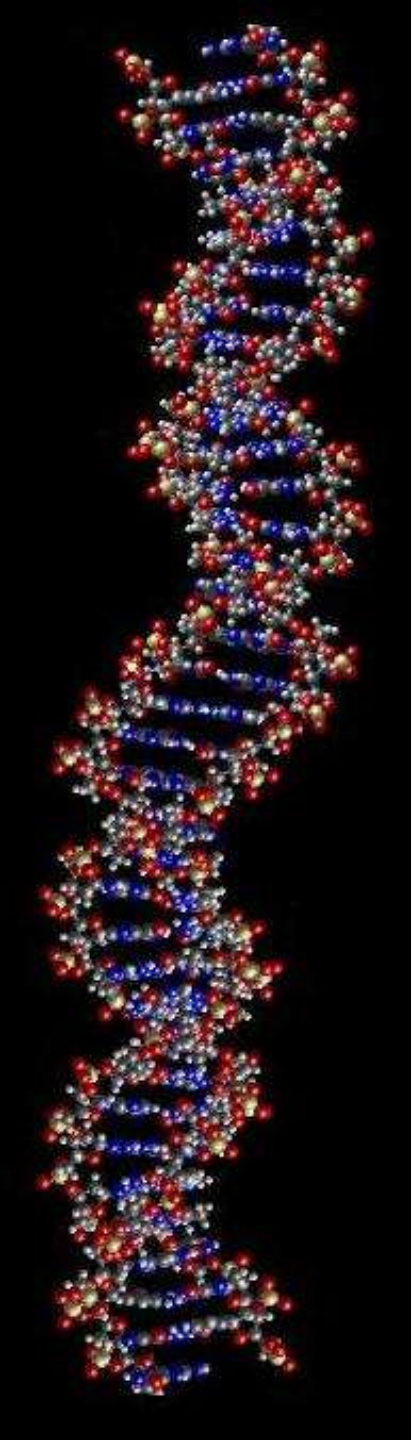
- DNA that doesn't get turned into proteins
- Between genes or inside genes
- Introns = “junk DNA” inside genes





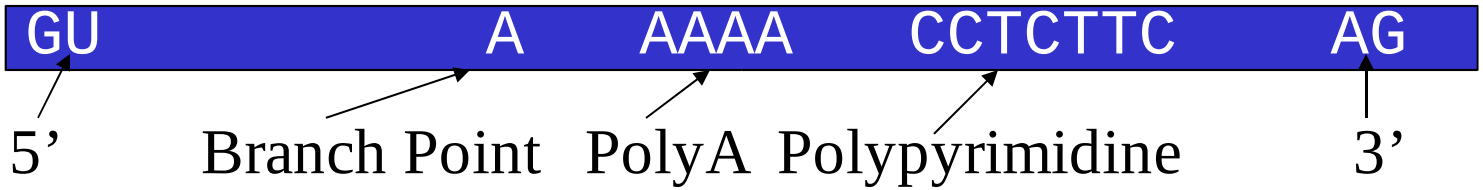
Introns: A Question of Conservation

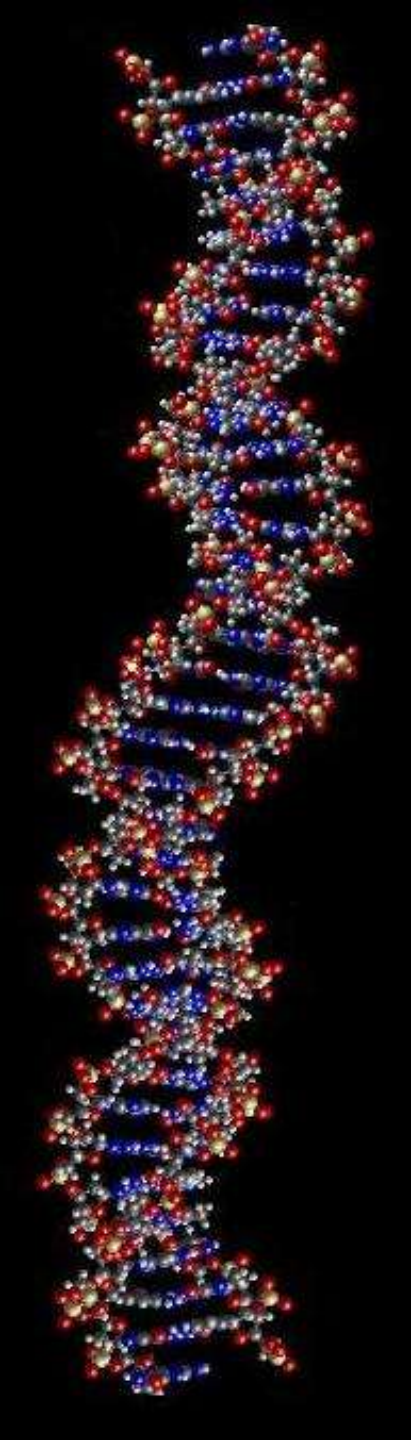
Introns Are Significant	**Introns Have No Function**
Play regulatory role in transcription; influence proteins	Spliced out of genes early in transcription
Found in all higher organisms	Not present in mRNA or proteins
If they're useless, why are they present?	Only known motifs relate to splicing process



Introns: Conserved Sequences

- Splice sites: 5’ (GU) and 3’ (AG)
- Polypyrimidine tract before 3’ end
- Branch point: Single A 20-50 bases upstream of 3’ splice site
- Some branch points contain polyAs
- Few other conserved sequences found

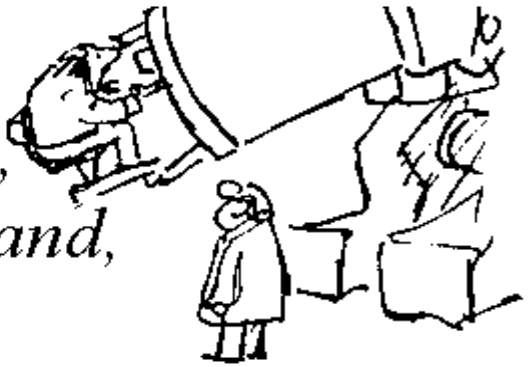




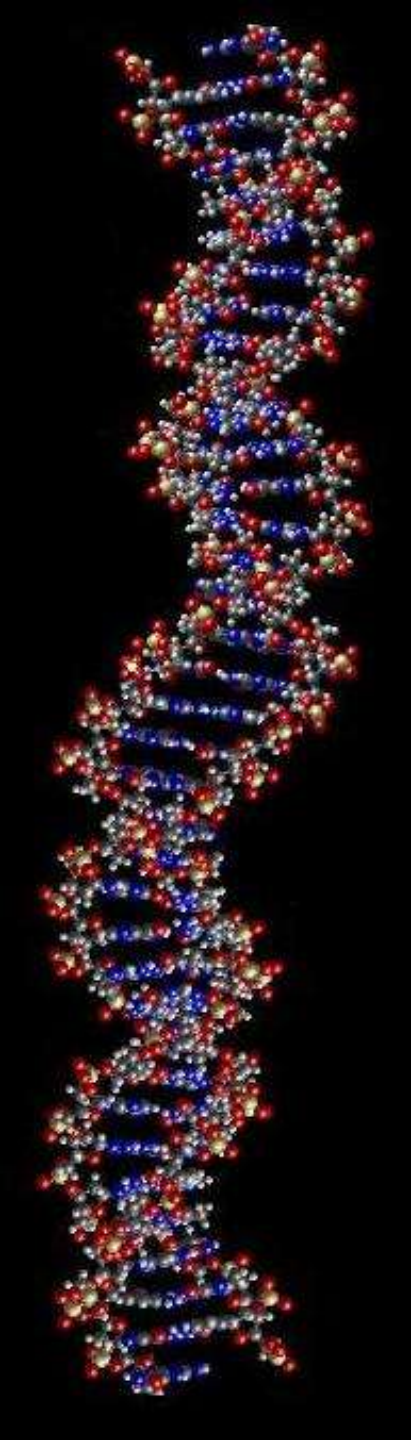
Why Use Bioinformatics?

- Human genome contains ~3 billion bases
- Too tedious to do by hand

"Let's see, now... picking up where we left off... one billion, sixty-two million, thirty thousand, four hundred and thirteen..."

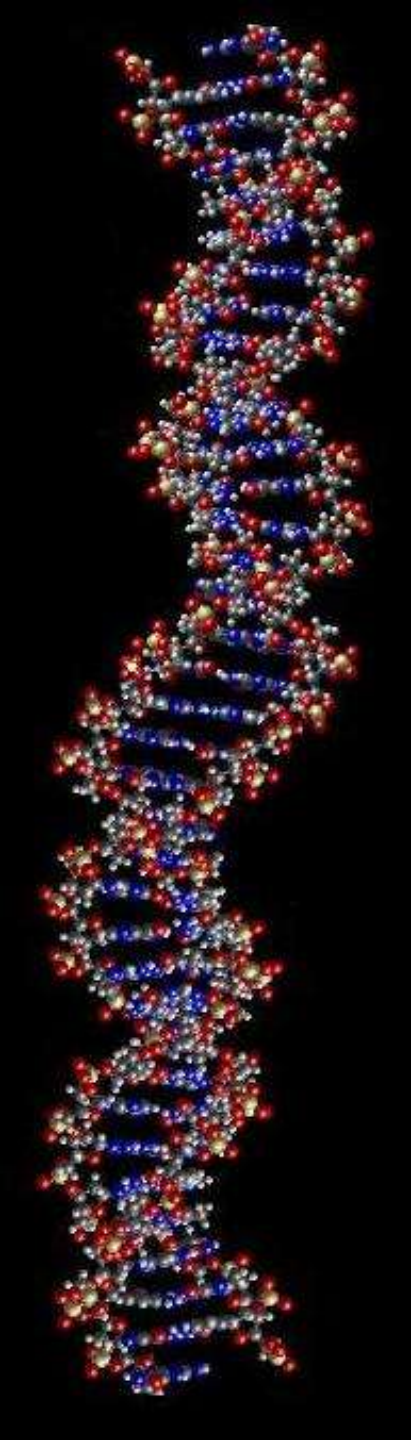


- Stored in computer database
- Quickly + automatically analyzed
- Highly accurate + versatile



Perl: The Major Bioinformatics Language

- Runs under Unix, Linux, some Win32
- Similar to C/C++ in style
- Interpreted language
- Not necessarily OO, but can be
- Major uses in bioinformatics:
  - Running processes/programs
  - Parsing/reformatting results/data
  - Managing files/data
  - Building analysis programs



LALIGN

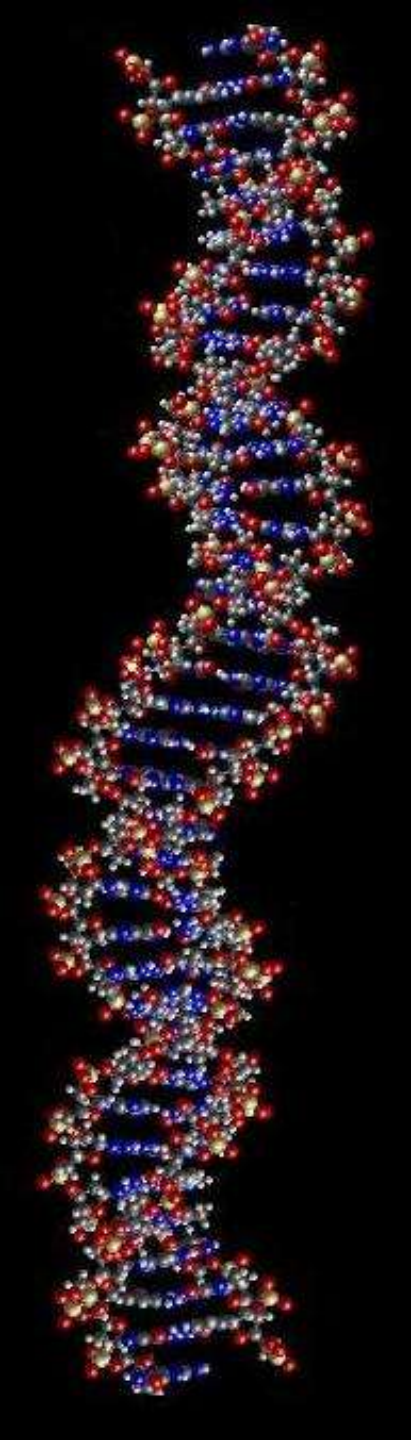
- Local alignment program
- Similar to Pairwise BLAST
- Takes 2 sequences, compares base pairs

```
 LALIGN finds the best local alignments between two sequences
 version 2.1u03 April 2000
Please cite:
 X. Huang and W. Miller (1991) Adv. Appl. Math. 12:373-381

alignments <  E(  0.05):score: 42 (50 max)
 Comparison of:
(A) @          SEQ1 sequence
                                        - 146 aa
(B) @          SEQ2 sequence
                                        - 146 aa
 using matrix file: BL50, gap penalties: -12/-2 E(limit)    0.05

 69.2% identity in 146 aa overlap (1-146:1-146); score:  701 E(10000): 3.2e-43

               10        20        30        40        50        60        70
SEQ1   VHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLTSFGDAV
       :: ::.::  ..:::::::: : :.:::.:::.:::::::::.:::.::. .:..::  :.::::::::.:::...
SEQ2   VHLTADEKAAVSGLWGKVNVDEVGGEALGRLLVVYPWTQRFFTSFGDLSNAAAVMGNSKVKAHGKKVLNSFGEGL
               10        20        30        40        50        60        70
```

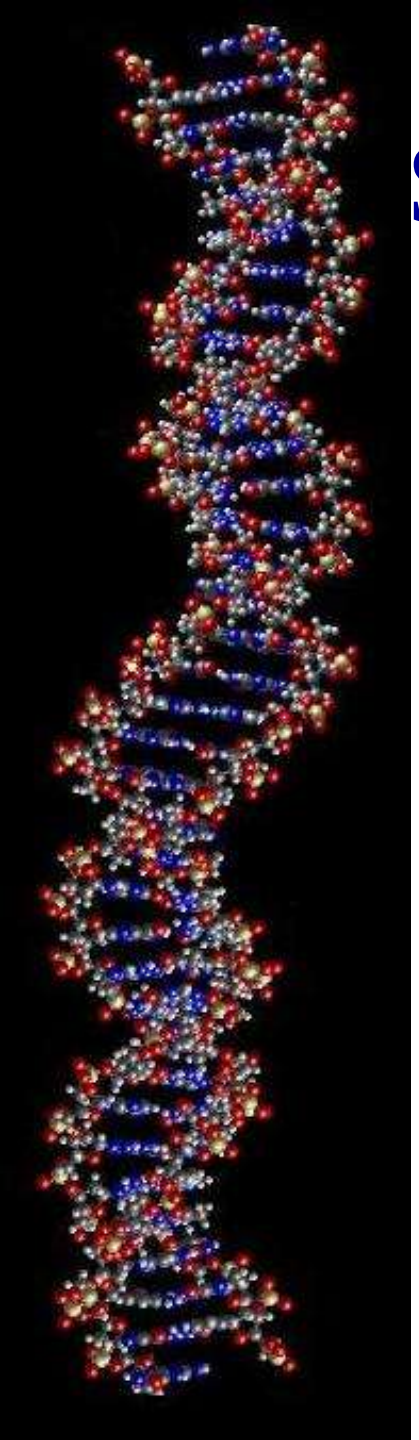


The Apollo Genome Browser

- Program for displaying genomic data
- Input data in Artemis flavor of GFF (Gene Finder Format)

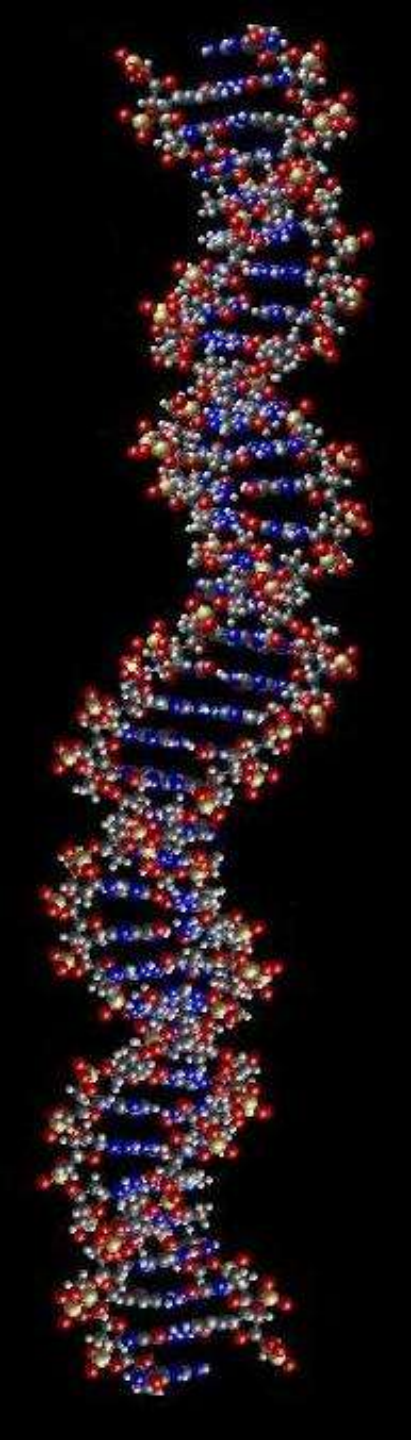
Seq_id	Type	Program	Start	Stop	Score	Strand	Frame
FMR1	exon	LALIGN	1	537	100.0	+	.

- Shows many types of results
- Can be customized



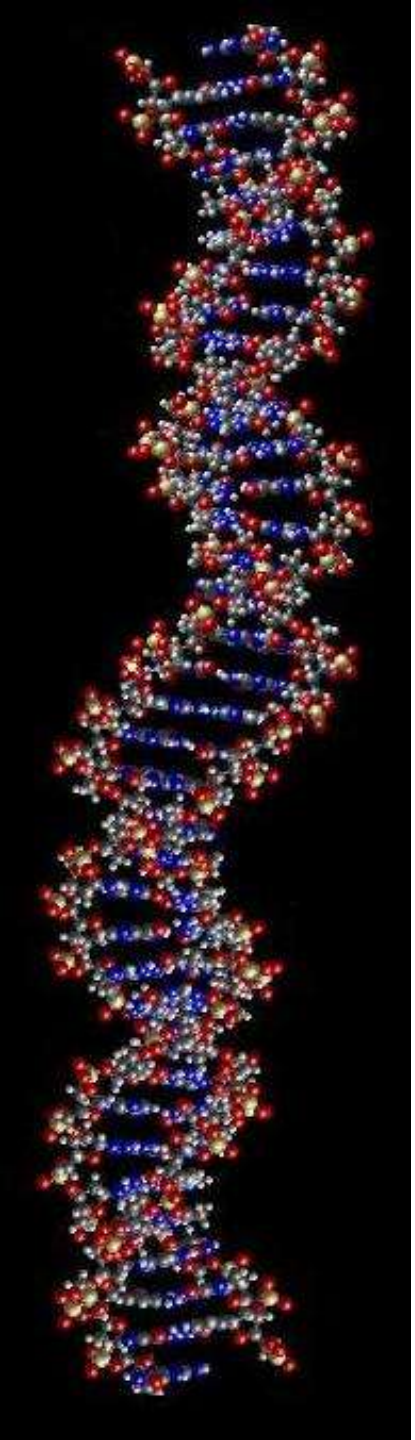
Seven Alternative Splice Genes

- AF088282: OGG1: Removes 8-oxoguanine (mutagenic byproduct of reactive oxygen exposure)
- AF110798: IL18BP: Inhibits IL18, a cytokine that causes inflammation
- AF135025: KLK12: Serine protease; possible cancer/ disease marker
- AF199339: PSIP1/PSIP2: PSIP1: Protects cells from cell death by apoptotic cleavage; PSIP2: Binds to involucrin promoters; regulates involucrin gene expression
- AY052369: PPP2R5C: Function unknown at this time
- L29074: FMR1: Mutations cause fragile X syndrome (esp. mental retardation)
- M10014: FGG: Component of fibrinogen, a precursor of fibrin (blood clotter)



Methodology: The Basics

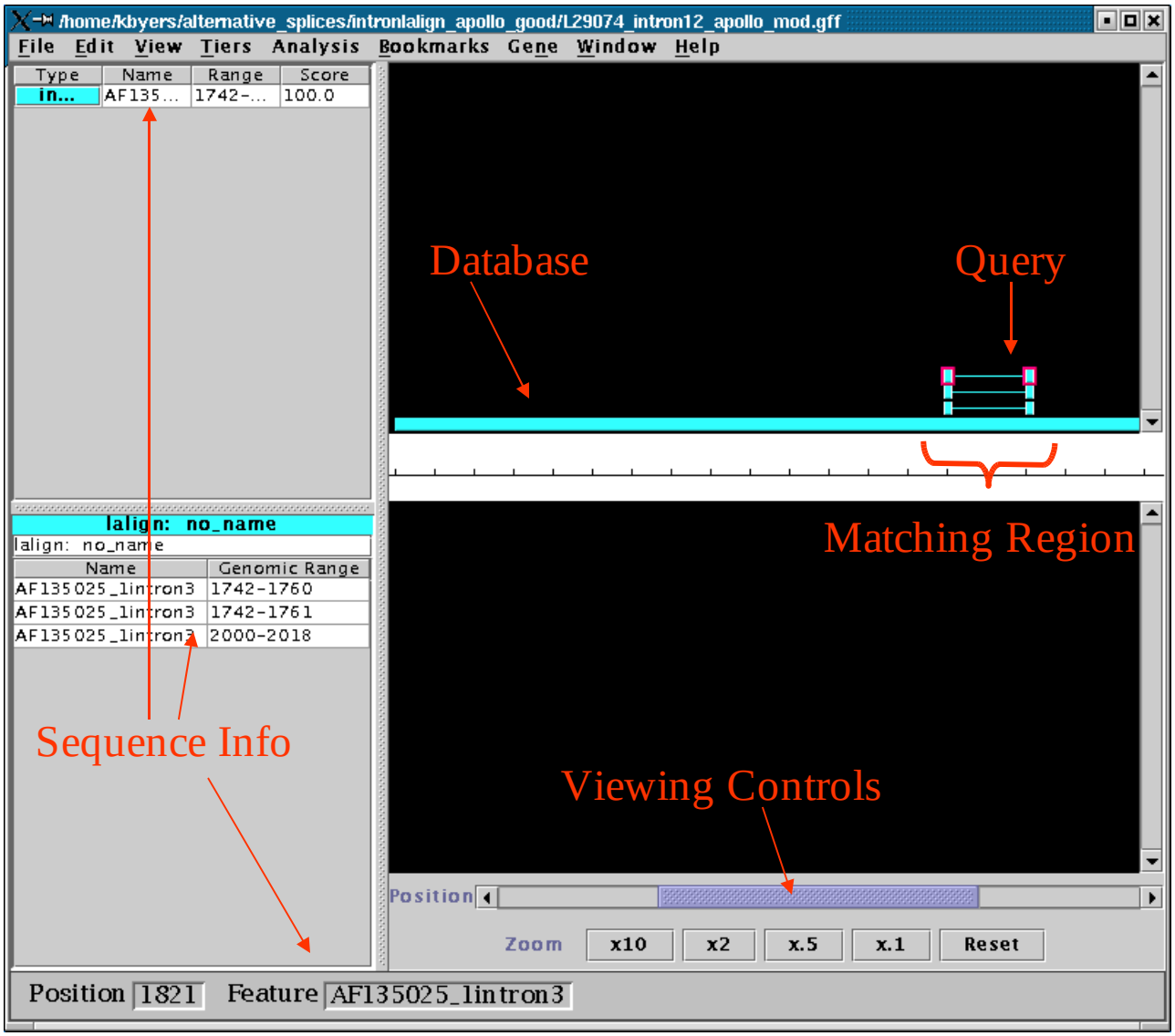
1. Found 7 genes that sometimes skip exons during transcription (alternative splice genes)
2. Got introns (102) from GenBank (removed by hand)
3. Compared introns using LALIGN
4. Wrote a Perl script to do this with one command instead of 10,404

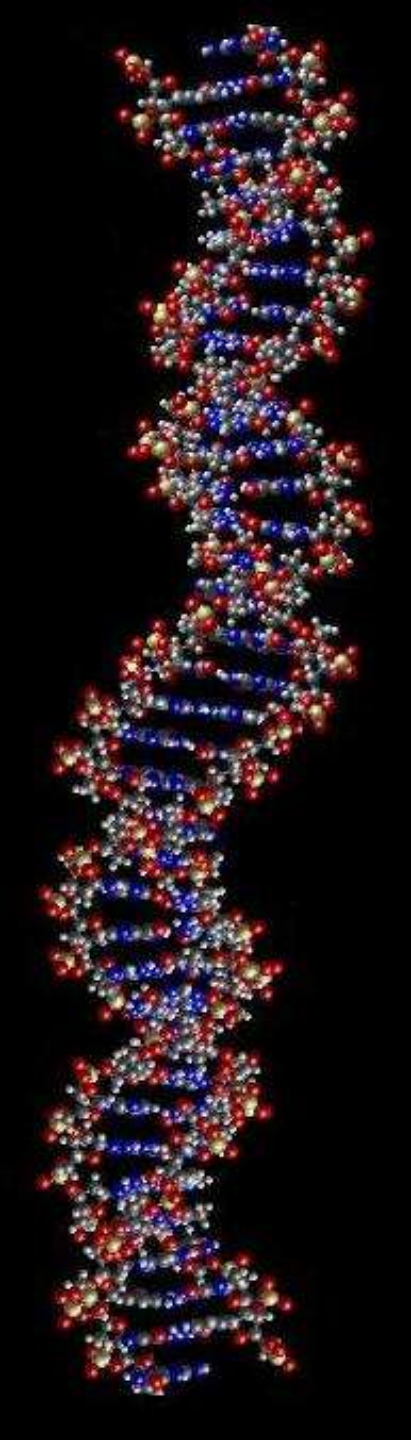


Methodology: Parsing

1. Parsed LALIGN files to remove all but coordinate/score info
2. Removed all hits below 100% similarity
3. Converted the results into GFF format
4. Removed conflicting/duplicated results
5. Took screenshots of Apollo displays using xv (graphics program)
6. Found regions of similarity and removed them

Example of Results





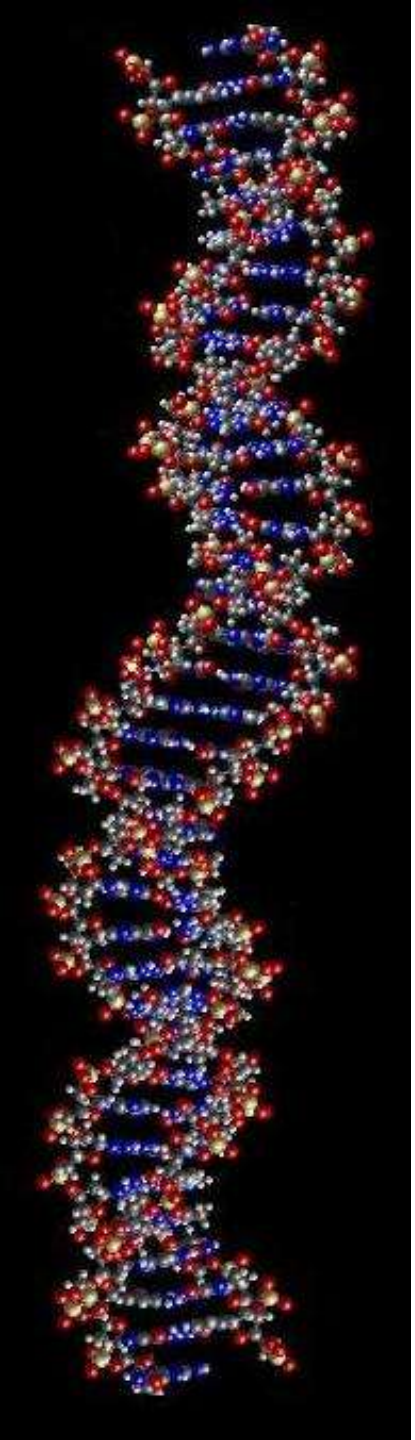
Results Summarized

- Found 16 common sequences
- All polyAs or polyTs

AGTGCAAAAAAAAAAAAAAAAAAAAAAACTGCTACTG

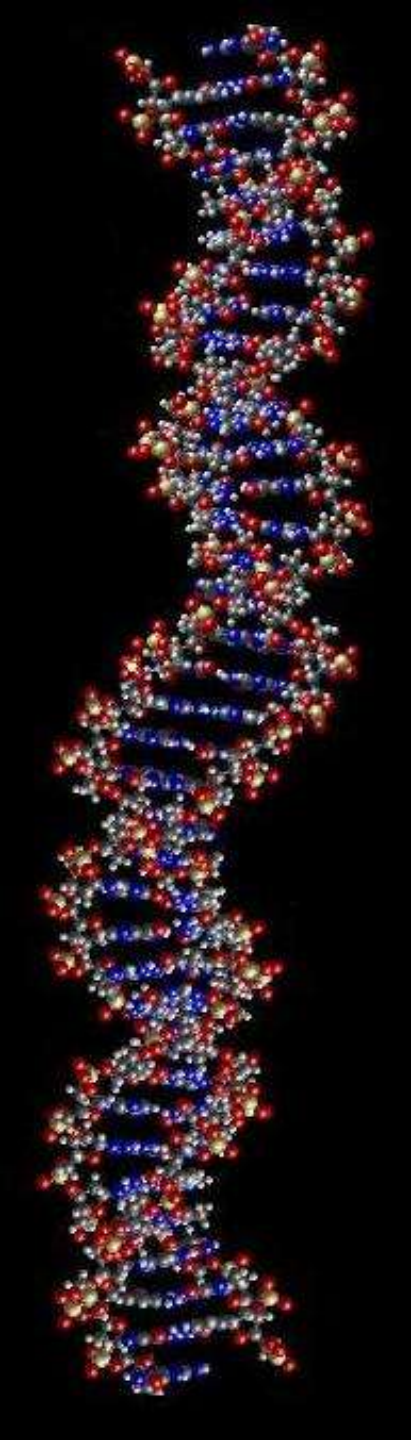
AGTGCTTTTTTTTTTTTTTTTTTTTTTTTTTCTGCTACTG

- Data on sizes of sequences:
  - Mean: 20.9375 base pairs
  - Median: 20 base pairs
  - Mode: 20 or 24 base pairs



Conclusions

- PolyAs common but NOT where found
- PolyTs not discussed in introns at all
- Similar Sequences = Conservation
- Conservation = Energy
- Energy expenditure must have a reason
- “Junk DNA” may not be “junk” at all
- Possible function for introns?



Thank You

Thanks especially to Dr. David Form, my advisor and bioinformatics mentor.

Questions?