

CMSC 423: Bioinformatics

Introduction

What does biology have to do with computers?

- Huge amount of data: too much to analyze by hand, lots of mystery left about how life works.
- Requires clever algorithms to:
 - find interesting patterns
 - store / search / compare
 - visualize vast collections of data
 - predict missing or hard-to-observe features (like protein structure or evolutionary relationships)
- Nearly all molecular biology is now “computational biology”: biologists depend on computer scientists every day.

Course work & Evaluation

- One midterm exam (25%):
 - **Oct. 16th**
 - The date is fixed.
- Comprehensive final (35%):
 - **Thursday, Dec 13, 2012**
 - **8-10am.**
 - Covers everything in the class
- Several homework assignments.
 - 10% of your grade
 - Neatness counts
- Programming project:
 - 30% of your grade.
 - Details soon.
 - Almost certainly in small groups.

Administrative Details

- Homeworks are due at the start of class on their due dates. **No late homeworks will be accepted.**
- You can **discuss** homeworks with your classmates.
- You **must list** the names of the people with whom you collaborated at the top of the homework.
- You must write up homeworks solutions **on your own.**
- Late **programming** assignments will lose 10% per day, up to 5 days, after which they will not be accepted.
- TA: Milad Gholami
- TA office hours: MW 9-11am.
- Instructor office hours: Tuesdays 12:30-2pm, or by appointment, in CBCB 3113.
- Grades will be posted on: <http://grades.cs.umd.edu>
- More details on syllabus handout.

Computational Successes:

1. Genome assembly
2. Gene discovery
3. Understanding the origin of swine flu
4. Predicting protein structure

Genome assembly



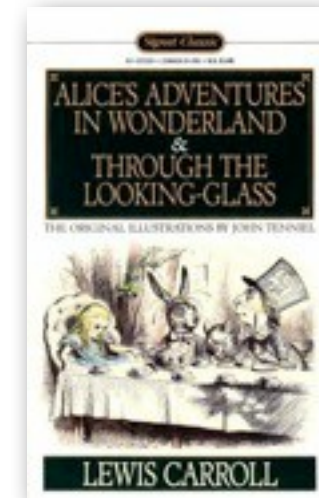
Genome of the Cow

a sequence of 2.86 billion
letters

enough letters to fill a
million pages of a typical
book.

```
TATGGAGCCAGGTGCCTGGGGCAACAAGACTGTGGTCACTGAATTCATCCTTCTTGGTCTAACAGAGAACATAG  
AACTGCAATCCATCCTTTTTTGCCATCTTCCTCTTTGCCTATGTGATCACAGTCGGGGGCAACTTGAGTATCCTG  
GCCGCCATCTTTGTGGAGCCCAAACCTCCACACCCCCATGTACTACTTCCTGGGGAACCTTTCTCTGCTGGACAT  
TGGGTGCATCACTGTCACCATTCCTCCCATGCTGGCCTGTCTCCTGACCCACCAATGCCGGGTTCCTTATGCAG  
CCTGCATCTCACAGCTCTTCTTTTTTCCACCTCCTGGCTGGAGTGGACTGTCACCTCCTGACAGCCATGGCCTAC  
GACCGCTACCTGGCCATTTGCCAGCCCCCTCACCTATAGCATCCGCATGAGCCGTGACGTCCAGGGAGCCCTGGT  
GGCCGTCTGCTGCTCCATCTCCTTCATCAATGCTCTGACCCACACAGTGGCTGTGTCTGTGCTGGACTTCTGCG  
GCCCTAACGTGGTCAACCACTTCTACTGTGACCTCCCGCCCCCTTTTCCAGCTCTCCTGCTCCAGCATCCACCTC  
AACGGGCAGCTACTTTTCGTGGGGGCCACCTTCATGGGGGTGGTCCCCATGGTCTTCATCTCGGTATCCTATGC  
CCACGTGGCAGCCGCAGTCCTGCGGATCCGCTCGGCAGAGGGCAGGAAGAAAGCCTTCTCCACGTGTGGCTCCC  
ACCTCACCGTGGTCTGCATCTTTTATGGAACCGGCTTCTTCAGCTACATGCGCCTGGGCTCCGTGTCCGCCTCA  
GACAAGGACAAGGGCATTGGCATCCTCAACACTGTCATCAGCCCCATGCTGAACCCACTCATCTACAGCCTCCG  
GAACCCTGATGTGCAGGGCGCCCTGAAGAGGTTGCTGACAGGGAAGCGGCCCCCGGAGTG ...
```

We can only read ~ 1000 characters at a time from a random place:



good-natured, she thought: still
when it saw Alice. It looked
ought to be treated
good-natured, she thought, still
Cat only
a greet many
It looked good-
The Cat only grinned when it saw Alice.
be treated with respect.
still it had very long claws
claws and a great many teeth, so she
so she felt that it ought

Algorithms are needed to piece the story together.

The Cat only grinned when it saw Alice.
Cat only when it saw Alice. It looked
It looked good-

good-natured, she thought: still
good-natured, she thought, still
still it had very long claws

claws and a great many teeth, so she
a greet many so she felt that it ought

ought to be treated
be treated with respect.

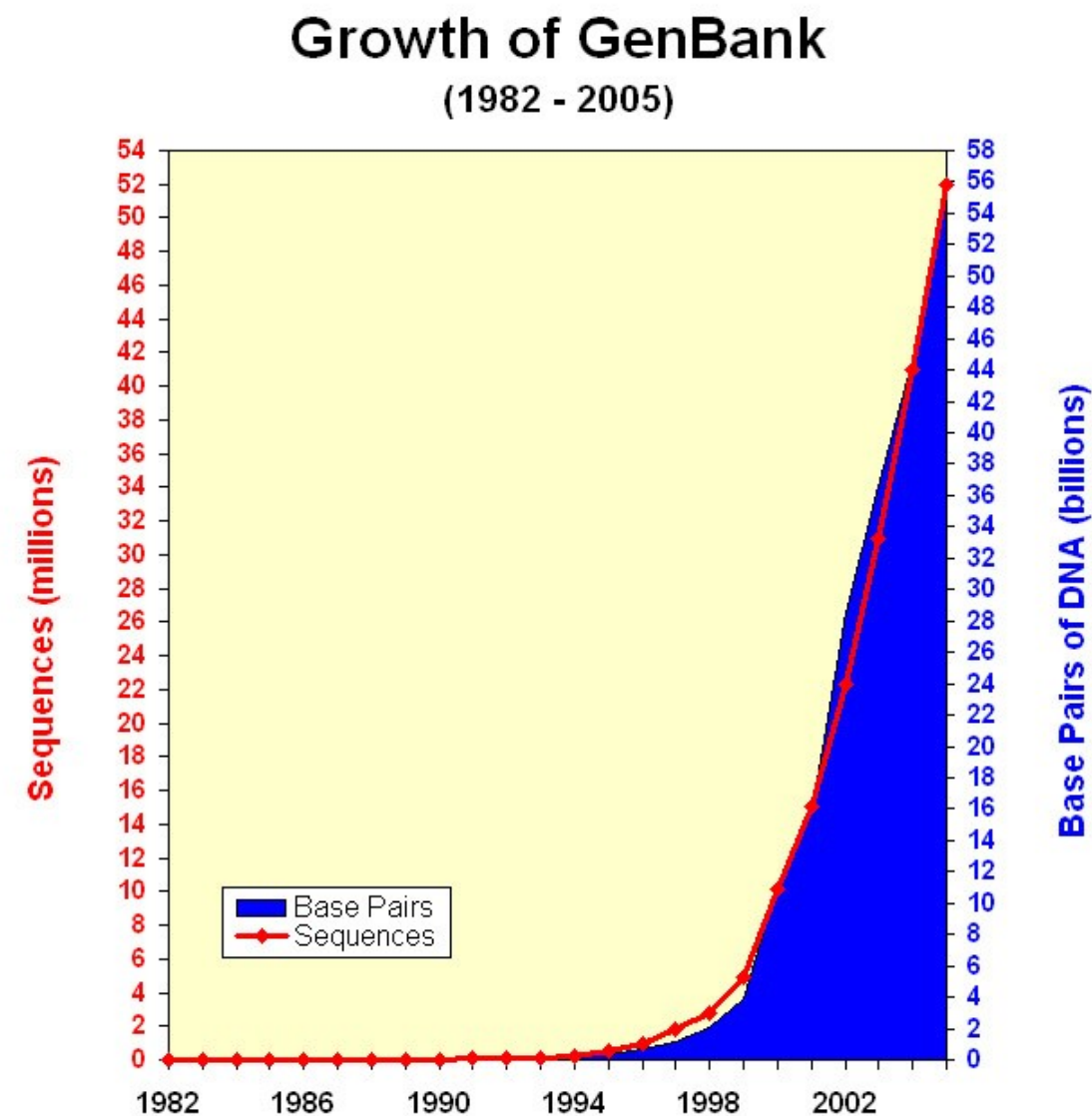
It's a jigsaw
puzzle ...



...except with 35
million pieces

Recent Genomics (DNA)

- First genome sequenced in 1995 (the bacteria *H. influenzae* with a genome of 1,830,140 letters).
- 1st draft of human genome finished in 2001 (~ 3 billion letters)
- Now: Over 1100 bacterial genomes
- Hundreds of higher-order genomes done or in progress.
- Several complete human genomes finished.





Researchers at many institutions are putting together the genomes of many animals.



Help understand how to make animals and plants more hardy, resistant to disease, and understand their biology.



New technologies and larger genomes require new algorithms and faster computers.

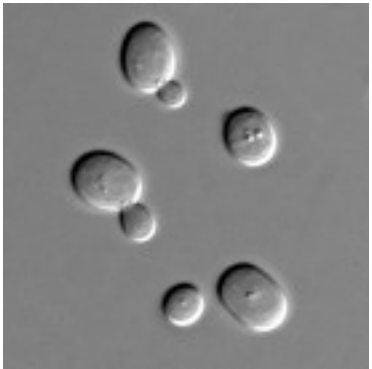
Other Sequenced Genomes



Arabidopsis thailiana



Callithrix jacchus
(marmoset)



Saccharomyces cerevisiae
(baker's yeast)



Canis lupus familiaris
(dog)



Apis mellifera
(honey bee)



Drosophila melanogaster
(fruit fly)



Bos torus
(cow)



Equus caballus
(horse)

Sequenced Eukaryotic Chromosomes

and many
more...

Gene discovery

```
121 tgggtgtaccc cccctccccc tatgtatatc gtgcattaat ggcgtgcccc atgcatataa
181 gcatgtacat actgtgcttg gctttacatg aggatactca ttacaagaac ttattttcaag
241 cगतगुतुता तगुतगुतुता ततुतगुतुता ततुतगुतुता ततुतगुतुता ततुतगुतुता
301 accagaatc cttgcgagta cgtgtacctc ttctcgctcc gggcccataa tttgtggggg
361 tttctatact gaaactatac ctggcatctg gttcttacct cagggccatg tttagcgtaa
421 ctcaatccta ctaacccttc aaatgggaca tctcgatgga ctaatgacta atcagcccat
481 gatcacacat aactgtggtg tcatgcatth ggtatthttt aatthttagg ggggaacttg
541 ctatgactca gctatgacg taaaggtctc gtcgcagtc aatcagctgt agctgggctt
601 attcatcttt cgaggtcct catggacacc cataaggtgc aattcagtc atggtcacag
661 gacataacac tatagatcac ccgactggc gttacgtgta cgtacgtgta cgtacgtgta
721 cgcacgtgta cgtacgtgta cgcacgtgta cgtacgtgta cgcacgtgta cgcacgtgta
781 cgtacgtgta cgcacgtgta cgtacgtgta cgtacgtgta cgcacgtgta cgcacgtgta
841 cgtacgtgta cgcacgtgta cgtacgtgta cgcacgtgta cgcacgtgta cgcacgtgta
901 cgtacgtgta cgcacgtgta cgtacgtgta cgcacgtgta cgcacgtgta cgtacgtgta
961 cgcacgtgta cgcacgtgta cgcacgtgta cgcacgtgta cgcacgtgta cgtacgtgta
1021 cgcgtacgta ttttagatac taagttagct tagacaaaacc ccccttacc cccgtaactt
1081 caagaagctt acatatactt atggatgtcc tgccaaaccc caaaaaaag actaaatata
1141 tgcgcaaaaca tgaagtcaact tacacctaata cccatataat taagctaacc cccagccaa
1201 tgttgcaaca actacggaca tgggactcta aattttaatt tatctataga tatttttctt
1261 ttactgtgtc tccccagcat tgatttttta attatcatta ttccacacca ccaatttcca
1321 ttgagctatt tcacatgagt tccaaatcaa ttatgttcat gtagcttaac gaataaagca
1381 aggtactgaa aatgcctaga tgggtcacgc taccocatag acataaaggt ttggtcctag
1441 ctttctatt agccattaac aagattacac atgtaagtct ccacgctcca gtgaaaatgc
1501 cccttaagtc ctcttagacg acctaagga gcgggtatca agcacacctt atggtagctc
1561 acaacgcctt gcttagcac accccacgg gaaacagca tgataaaaat taagctatga
1621 acgaaagttc gactaagcta tgtaataact aggttggtg aatctcgtgc cagccaccgc
1681 ggtcatacga ttaactcgag ttaataggcc tacggcgtaa agcgtgtaaa agaaaaaatc
1741 tccttacta aagttaaagt atgattaagc tgtaaaaagc taccattaat actaaataa
1801 actacgaaag tgacttttaa atttctgatt acacgatagc tagggcccaa actgggatta
1861 gatacccccac tatgcctagc tctaaacata gatattttac taaacaaaac tattcgccag
1921 agaactacta gcaacagctt aaaactcaaa ggacttggcg gtgctttata tccccctaga
1981 ggagcctggt ctgtaatcga taaacccga tagacctcac catcccttgc taatcagtt
2041 tatataccgc catcttcagc aaacccttaa aaggaaaaaa agtaagcata actaccctac
2101 ataaaaaagt taggtcaagg tgtaacctat gggctgggaa gaaatgggct acattttcta
2161 ttcaagaaca acttctacga aaacttttat gaaactaaa gctaaaggcg gatttagtag
2221 taaattaaga atagagagct taattgaaca gggcaatgaa gcacgcacac accgccgctc
2281 accctcctcg agtgatataa ttaattata acctatttaa actaagcaaa gcataagagg
2341 agacaagtcg taacaaggta agcatactgg aaagtgtgct tggatgagcc aaagtgtagc
2401 ttaacaaaag cgtctggctt acatccagaa gatttcatta atatatgact actttgaacc
2461 caaagctagc ccaagcaaca atgactagta aaaccattat gaaacattca acaaaaacat
2521 ttagtagcat gactagagta taggagatag aaatttttaa ctggagctat agagagagta
2581 ccgcaaggga atgatgaaag attacctaaa gtgataaaca gcaaagattg ccccttctac
2641 cttttgtata atgagttagc tagaaataac ttaacaaaga gaacttaagc taagtcccc
2701 gaaaccagac gagctaccta tgaacaatcc actgggatga actcatctat gttcaaaaat
2761 agtgagaaga tccataggta gaggtgaaag gcctaacgag cctggtgata gctggtgccc
2821 cagaatagaa ttttagttcg actttaaac tgctacaaa actaataatt ctaatgcaga
2881 tttaaaatat attctaaaaa ggtacagctt tttagagtta aggatacagc cttacttaga
2941 gagtaatat ttatataagc catagtaggc ctagaggcag ccatcaatta agaaagcgtt
3001 aaagctcaac atctctatta acttaatacc aagaatattt aatcaactcc taatgtatta
3061 ctgggtcaat ctattttaat atagaagtga taatgcta atgagtaaca agaaatattt
3121 ctcccattgca taagcttata acagcaacgg ataaccactg atagttaaca acaacataga
3181 aataacctaa tgataaaaaca cctattaaat caattgttag tccaacacag gcatgcaatc
3241 agggaaaagt taaaagaagt gaaaggaaact cggcaaatac aaaccccgcc tgtttaccaa
3301 aaacatcacc tccagcattt ccagtttgg aggcactgccc tgcccgggtga catcagttaa
3361 acggcccgcg tattctgacc gtgcaaaagg agcataatca tttgttctct aaataaggac
3421 ttgtatgaat ggccacacga gggtttaact gtctcttact tccaatcagt gaaattgacc
3481 tccccgtgaa gaggcgggga taagacaata agacgagaag accctatgga gctttaatta
3541 actaattcaa aaagaaacta ctaacgaccc aacaggaata atatatctct tttatgaatt
3601 agcaatttag gttggggcga cctcggagga caaaatagcc tccgagtgat tataaatcta
3661 gacttaccag tcaaaatgct taatcactta ttgatccaaa aattcttttg atcaacggaa
3721 caagttaccc taggataaac agcgcaatcc tatccgagag tccatcga caatagggtt
3781 tacgacctcg atgttgatc aggacatcct aatggtgcag cagctattaa aggttcgttt
3841 gttcaacgat taaagtcccta cgtgatctga gttcagaccg gagcaatcca ggtcgggttc
3901 tatctattca aataatttct cccagtacga aaggacaaga gaaataaggc ctacttctct
3961 gaagcgctt aagaccaata gatgaattta tctaaatcta gtaaatctaa ctccaatatt
4021 gcccaagaga cagggttttg ttagggtggc agagcccggt aattgtgcaa aacttaaaact
4081 cttgtgtcca gaggttcaat tcctctccct agcatatggt tataattaac atcttctcac
4141 taattgtacc tattcttctt gctgtagcct ttctgactct agtagaacga aaagtactag
```

Example Genomic Sequence

□ Giant Panda (*Ailuropoda melanoleuca*) mitochondrion
sequence [Peng et al, Gene 397:76-83 (2007)]



Obviously, computers are needed to understand what this means.

Where are the genes encoded in this sequence?

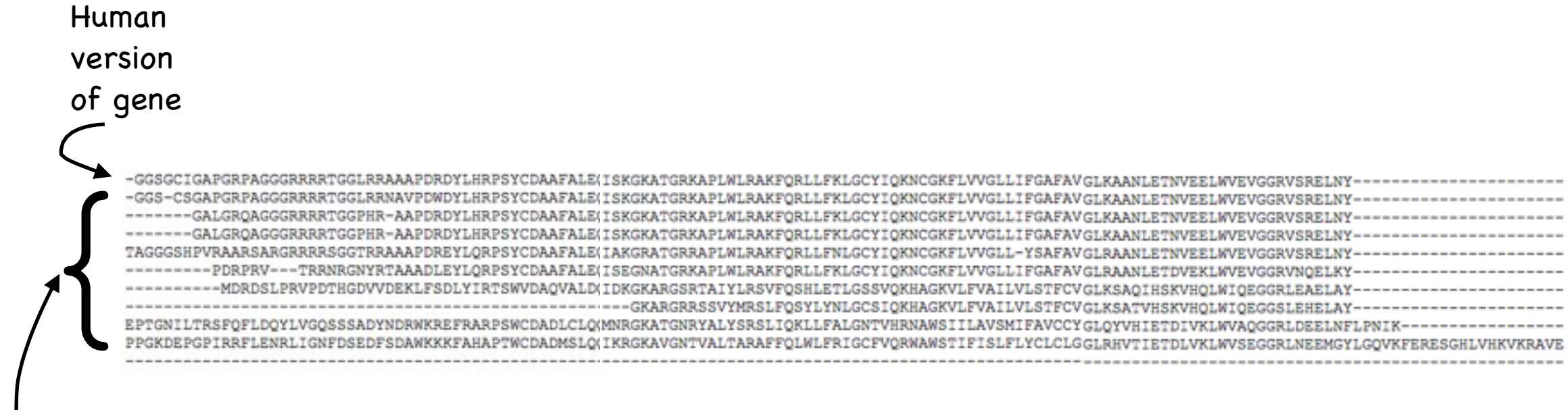
What causes each gene to be turned on or off?

How does the genome produce observed traits?

Two ways to find genes..

1. Search for sequence of DNA similar to known gene:

Human version of gene



The diagram shows a human gene sequence (top) and a set of similar sequences from other organisms (bottom). A bracket on the left groups the similar sequences, and an arrow points from the label 'Similar sequences in other organisms' to this bracket. The sequences are aligned, showing high similarity between the human version and the other organisms.

Similar sequences in other organisms

need to search for strings really fast (allowing for errors)

2. Search the genome sequence for patterns of letters that “look like genes”.

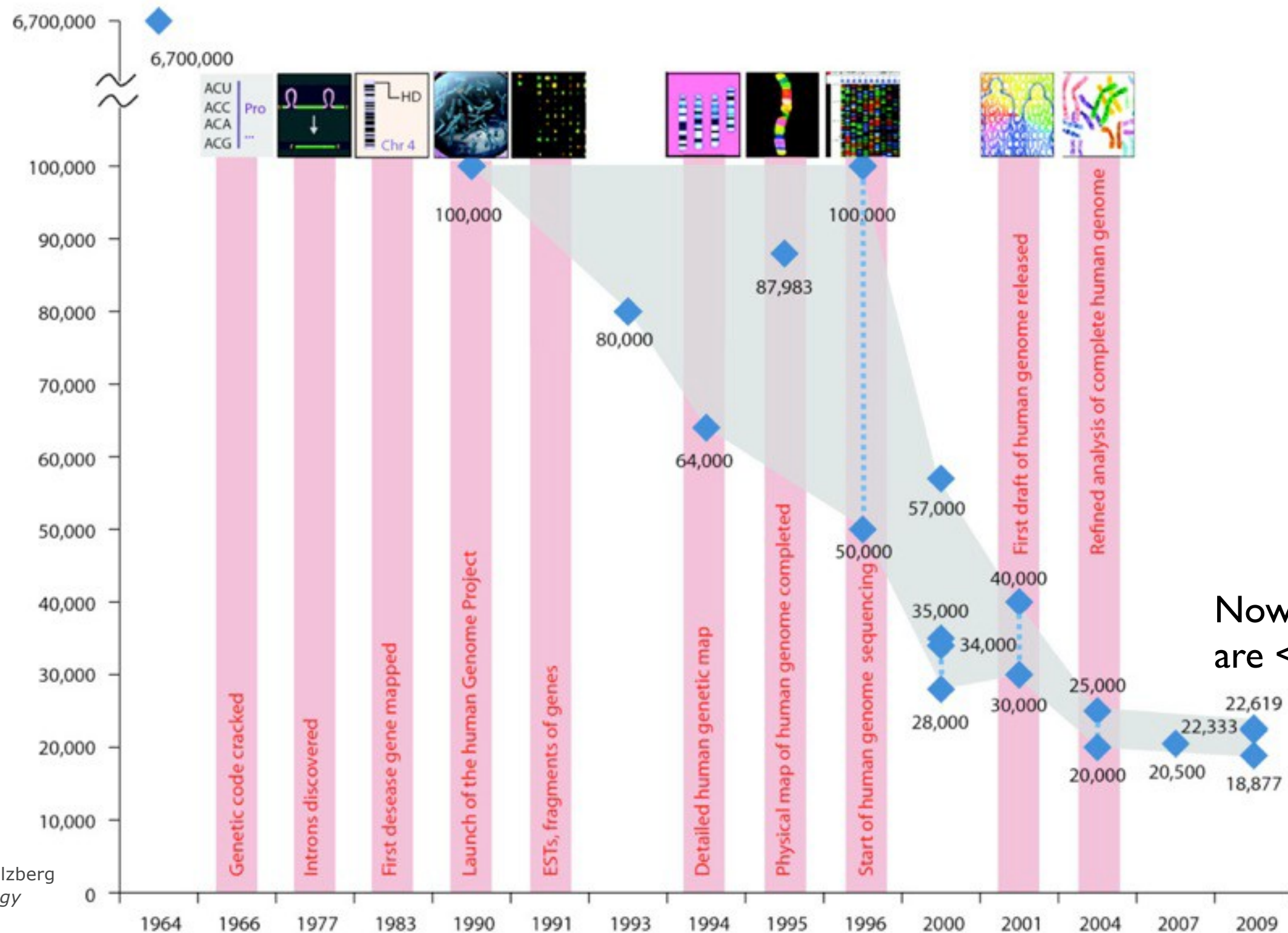
Some Human Genes



http://useast.ensembl.org/Homo_sapiens/Location/View?db=core;g=ENSG00000139618;r=13:32889611-32973805

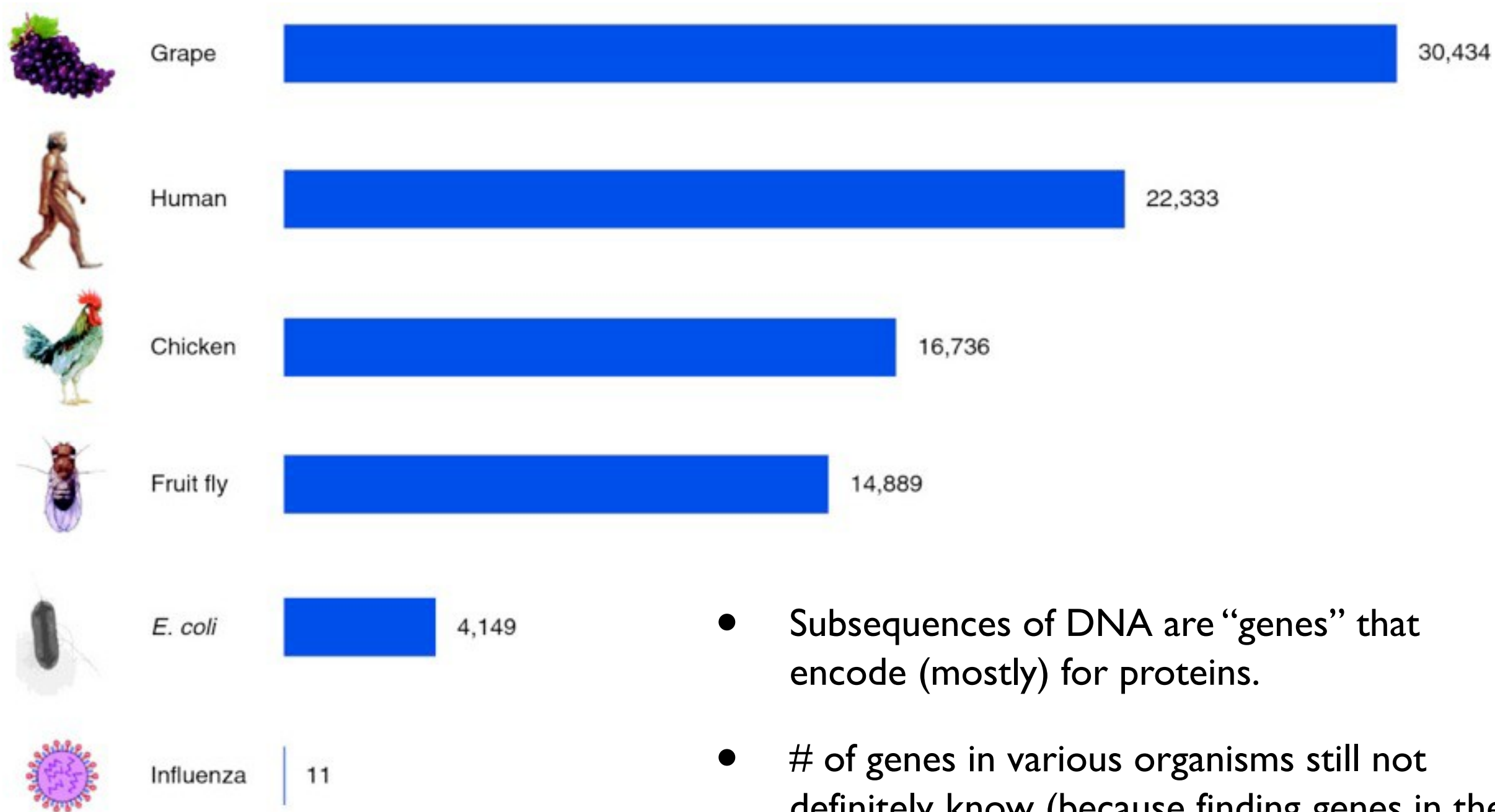
Estimates for the # of Human Genes

Before human genome sequence was available, many (but not all) estimates for # of genes were high (> 80,000).



Now estimates are < 23,000.

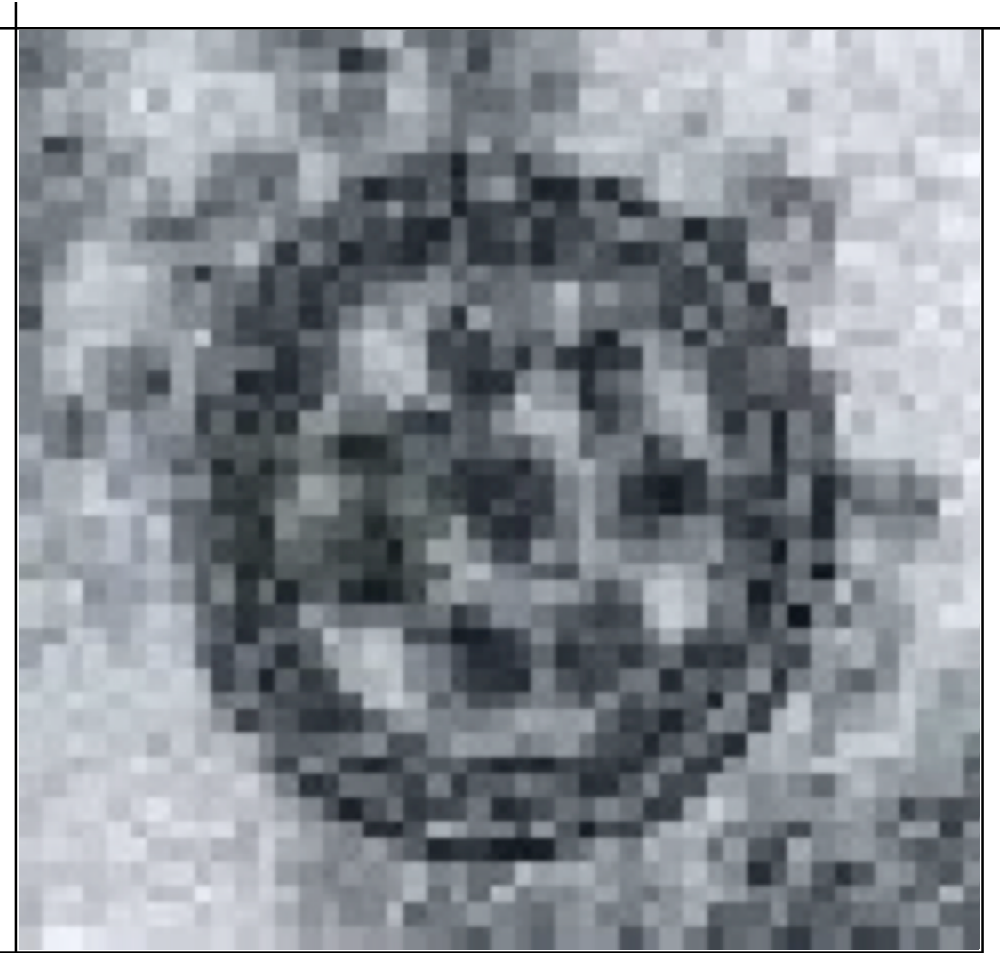
of Genes in Various Organisms



- Subsequences of DNA are “genes” that encode (mostly) for proteins.
- # of genes in various organisms still not definitely known (because finding genes in the sequence is a hard problem that we will talk about).
- But there are reasonably good estimates.

Tracking the evolution of influenza

Influenza Virus



(Toda et al., 2006)

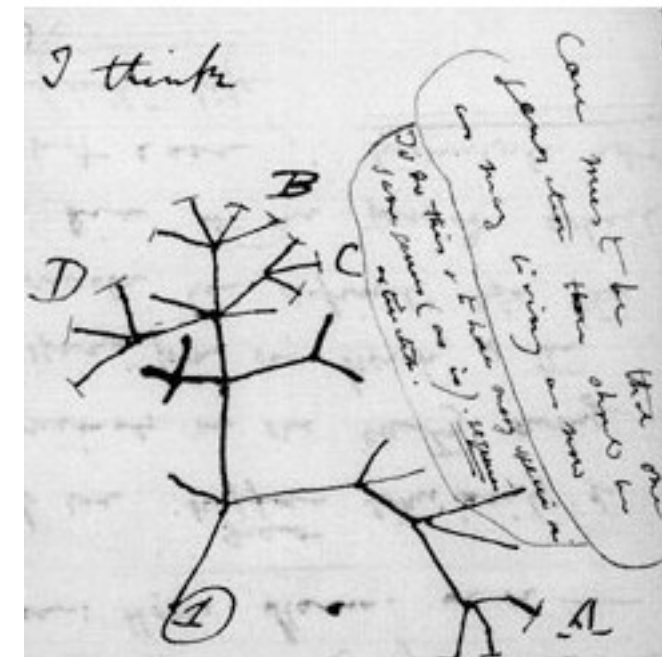
Rapidly evolving (it's genome is mutating):

that's why you have to get a different flu shot every year

3 strains must be selected each year to include in the vaccine.

So, the evolution of the virus is must be predicted.

HA



$?L_{-} \quad \text{---} \quad !\{ \quad -$
 $1-)... \quad f \quad <: -1- \quad lJ .a:::$
 $/vi-1- \quad 7\$; . Z>$
 $<2....::;1$
 $?L_{-} \quad \text{---} \quad r,...,....a \quad L$
 $- \quad rtL...A;$

<http://wwwamnh.org/exhibitions/darwin/idea/treelgp>
PoliciesFall2012.pdf

Tentative Course Topics & Outline

- Sequence search & comparison
 - Dynamic programming
 - Local / global alignments
 - Aligning multiple sequences
 - Suffix trees / suffix arrays
 - Burrows-Wheeler Transform
- Hidden Markov Models
 - For gene finding
 - For sequence pattern finding
- Phylogenetics
 - Algorithms for building trees
 - Genome Rearrangements
- Current Research Topics
 - Genome sequencing & assembly
 - RNA-seq expression
 - Pathway comparison / analysis

Algorithmic Techniques & Data Structures

- Dynamic programming
- Divide and conquer
- Branch and bound
- Linear programming
- Gibbs sampling / expectation maximization

Very general algorithmic techniques

- Hidden Markov Models
- Burroughs-Wheeler transform
- Suffix arrays / suffix trees

Techniques widely used in string / sequence algorithms.