Great Ideas in Computational Biology

02-251

Phillip Compeau & Carl Kingsford

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.

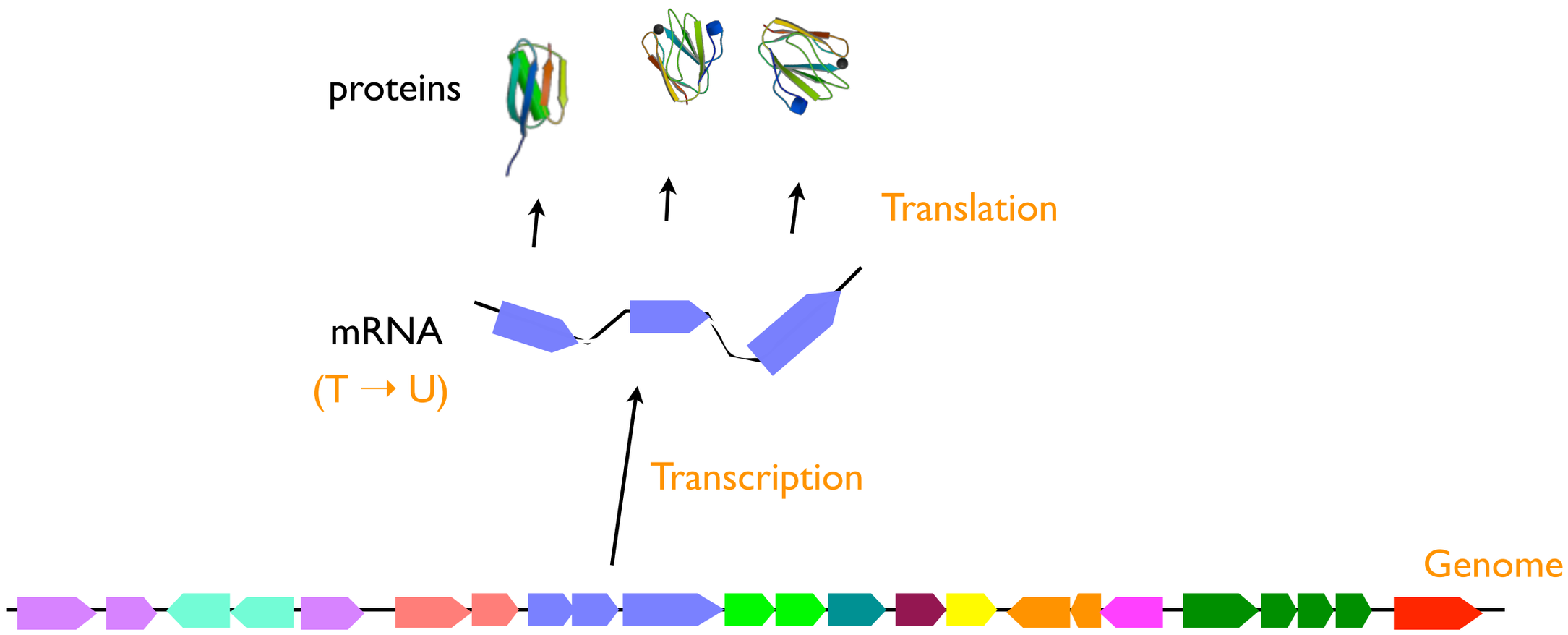
Example Computational Ideas

- Dynamic programming
- String data structures
- Hidden Markov Models
- Machine learning

Example 1: Genomics

- Algorithms for understanding and processing DNA and RNA data
- String algorithms are central

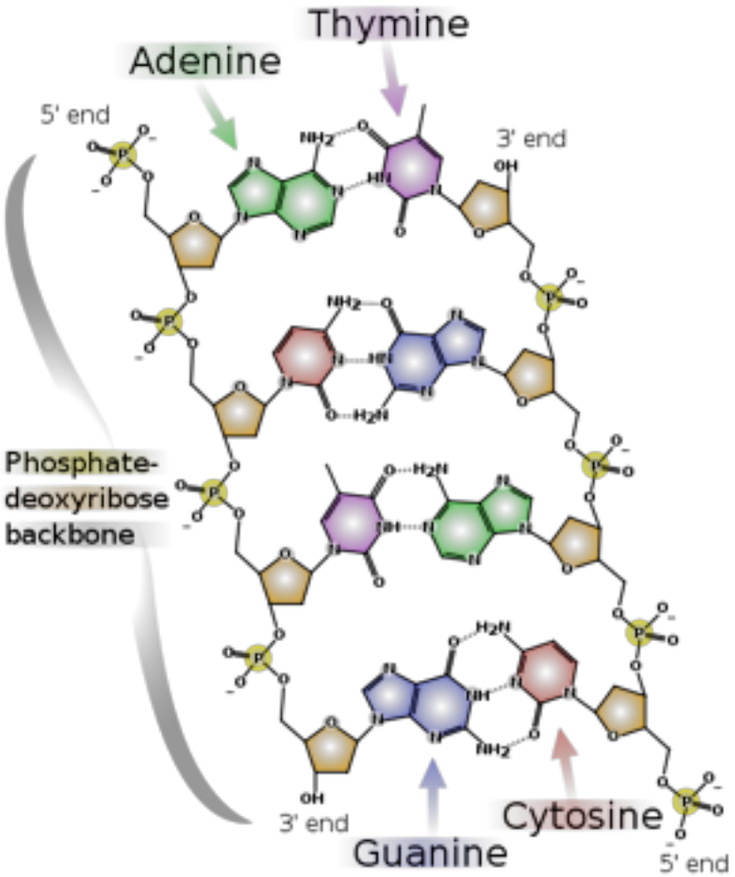
Central Dogma of Biology

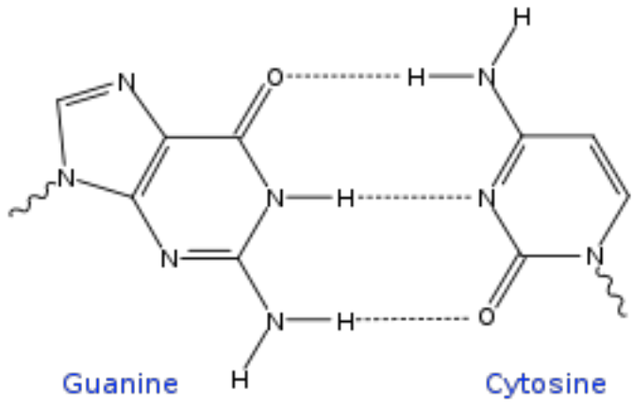


DNA =

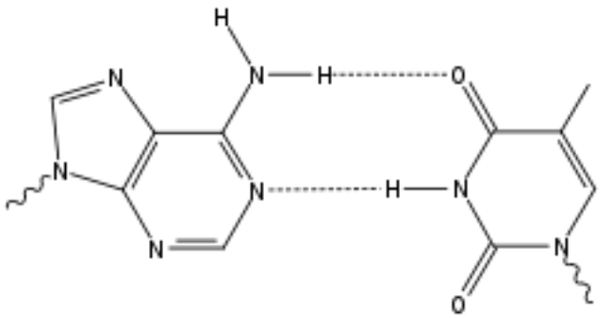
- double-stranded, linear molecule
- each strand is string over {A,C,G,T}
- strands are complements of each other (A ↔ T; C ↔ G)
- substrings encode for genes most of which encode for proteins

DNA


Thymine
Adenine
5' end
3' end
Phosphate-
deoxyribose
backbone
3' end
Guanine
Cytosine
5' end


Guanine
Cytosine

G C


Adenine
Thymine

A T

Genome of the Cow

a sequence of 2.86 billion letters

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

```
TATGGAGCCAGGTGCCTGGGGCAACAAGACTGTGGTCACTGAATTCATCCTTCTTGGTCTAACAGAGAACATAG
AACTGCAATCCATCCTTTTTGCCATCTTCCTCTTTGCCTATGTGATCACAGTCGGGGGCAACTTGAGTATCCTG
GCCGCCATCTTTGTGGAGCCCAAACTCCACACCCCCATGTACTACTTCCTGGGGAACCTTTCTCTGCTGGACAT
TGGGTGCATCACTGTCACCATTCCTCCCATGCTGGCCTGTCTCCTGACCCACCAATGCCGGGTTCCCTATGCAG
CCTGCATCTCACAGCTCTTCTTTTTCCACCTCCTGGCTGGAGTGGACTGTCACCTCCTGACAGCCATGGCCTAC
GACCGCTACCTGGCCATTTGCCAGCCCCTCACCTATAGCATCCGCATGAGCCGTGACGTCCAGGGAGCCCTGGT
GGCCGTCTGCTGCTCCATCTCCTTCATCAATGCTCTGACCCACACAGTGGCTGTGTCTGTGCTGGACTTCTGCG
GCCCTAACGTGGTCAACCACTTCTACTGTGACCTCCCGCCCCTTTTCCAGCTCTCCTGCTCCAGCATCCACCTC
AACGGGCAGCTACTTTTCGTGGGGGCCACCTTCATGGGGGTGGTCCCCATGGTCTTCATCTCGGTATCCTATGC
CCACGTGGCAGCCGCAGTCCTGCGGATCCGCTCGGCAGAGGGCAGGAAGAAAGCCTTCTCCACGTGTGGCTCCC
ACCTCACCGTGGTCTGCATCTTTTATGGAACCGGCTTCTTCAGCTACATGCGCCTGGGCTCCGTGTCCGCCTCA
GACAAGGACAAGGGCATTGGCATCCTCAACACTGTCATCAGCCCCATGCTGAACCCACTCATCTACAGCCTCCG
GAACCCTGATGTGCAGGGCGCCCTGAAGAGGTTGCTGACAGGGAAGCGGCCCCCGGAGTG ...
```

```
   1 atactataaa tccacctctc attttattca cttcatacat gctattacac actctgtgcc
  61 atcatagtat gttttcatac atcctccctt ctttcacacc ctatgtatat cgtacattaa
 121 tggtgtaccc cccctccccc tatgtatatc gtgcattaat ggcgtgcccc atgcatataa
 181 gcatgtacat actgtgcttg gctttacatg aggatactca ttacaagaac ttatttcaag
 241 cgatagtcta tgagcatgta tttcacttag tccaagagct tgatcaccaa gcctcgagaa
 301 accagcaatc cttgcgagta cgtgtacctc ttctcgctcc gggcccataa tttgtggggg
 361 tttctatact gaaactatac ctggcatctg gttcttacct cagggccatg ttagcgtcaa
 421 ctcaatccta ctaacccttc aaatgggaca tctcgatgga ctaatgacta atcagcccat
 481 gatcacacat aactgtggtg tcatgcattt ggtatttttt aattttaggg ggggaacttg
 541 ctatgactca gctatgaccg taaaggtctc gtcgcagtca aatcagctgt agctgggctt
 601 attcatcttt cgaggctcct catggacacc cataaggtgc aattcagtca atggtcacag
 661 gacataacac tatagatcac ccggactggc 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 tagacaaacc ccccttaccc cccgtaactt
1081 caagaagctt acatatactt atggatgtcc tgccaaaccc caaaaacaag actaaatata
1141 tgcgcaaaca tgaagtcact tacacctaaa cccatataat taagctaacc ccccagccaa
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 taccccatag acataaaggt ttggtcctag
1441 ccttcctatt agccattaac aagattacac atgtaagtct ccacgctcca gtgaaaatgc
1501 cccttaagtc ctcttagacg acctaaagga gcgggtatca agcacacctt atggtagctc
1561 acaacgcctt gcttagccac acccccacgg gaaacagcag tgataaaaat taagctatga
1621 acgaaagttc gactaagcta tgttaatact agggttggta aatctcgtgc cagccaccgc
1681 ggtcatacga ttaactcgag ttaataggcc tacggcgtaa agcgtgtaaa agaaaaaatc
1741 tcctctacta aagttaaagt atgattaagc tgtaaaaagc taccattaat actaaaataa
1801 actacgaaag tgactttaaa atttctgatt acacgatagc tagggcccaa actgggatta
1861 gataccccac tatgcctagc tctaaacata gatattttac taaacaaaac tattcgccag
1921 agaactacta gcaacagctt aaaactcaaa ggacttggcg gtgctttata tccccctaga
1981 ggagcctgtt ctgtaatcga taaaccccga tagacctcac catcccttgc taattcagtt
2041 tatataccgc catcttcagc aaacccttaa aaggaaaaaa agtaagcata actaccctac
2101 ataaaaaagt taggtcaagg tgtaacctat gggctgggaa gaaatgggct acattttcta
2161 ttcaagaaca acttctacga aaacttttat gaaactaaaa gctaaaggcg gatttagtag
2221 taaattaaga atagagagct taattgaaca gggcaatgaa gcacgcacac accgcccgtc
2281 accctcctcg agtgatataa tttaattata acctatttaa actaagcaaa gcataagagg
2341 agacaagtcg taacaaggta agcatactgg aaagtgtgct tggatgagcc aaagtgtagc
2401 ttaaacaaag cgtctggctt acatccagaa gatttcatta atatatgact actttgaacc
2461 caaagctagc ccaagcaaca atgactagta aaaccattat gaaacattca aacaaaacat
2521 ttagtagcat gactagagta taggagatag aaatttttaa ctggagctat agagagagta
2581 ccgcaaggga atgatgaaag attacctaaa gtgataaaca gcaaagattg ccccttctac
2641 cttttgtata atgagttagc tagaaataac ttaacaaaga gaacttaagc taagtccccc
2701 gaaaccagac gagctaccta tgaacaatcc actgggatga actcatctat gttgcaaaat
2761 agtgagaaga tccataggta gaggtgaaag gcctaacgag cctggtgata gctggttgcc
2821 cagaatagaa ttttagttcg actttaaacc tgcctacaaa actaataatt ctaatgcaga
2881 tttaaaatat attctaaaaa ggtacagctt tttagagtta aggatacagc cttacttaga
2941 gagtaaatat ttatataagc catagtaggc ctagaggcag ccatcaatta agaaagcgtt
3001 aaagctcaac atctctatta acttaatacc aagaatattt aatcaactcc taatgtatta
3061 ctgggtcaat ctatttaaat atagaagtga taatgctaat atgagtaaca agaaatattt
3121 ctcccatgca taagcttata acagcaacgg ataaccactg atagttaaca acaacataga
3181 aataacctaa tgataaaaca cctattaaat caattgttag tccaacacag gcatgcaatc
3241 agggaaagat taaaagaagt gaaaggaact cggcaaatac aaaccccgcc tgtttaccaa
3301 aaacatcacc tccagcattt ccagtattgg aggcactgcc tgcccggtga catcagttaa
3361 acggccgcgg tattctgacc gtgcaaaggt 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 tagggataac agcgcaatcc tatccgagag tccatatcga caatagggtt
3781 tacgacctcg atgttggatc aggacatcct aatggtgcag cagctattaa aggttcgttt
3841 gttcaacgat taaagtccta cgtgatctga gttcagaccg gagcaatcca ggtcggtttc
3901 tatctattca aataatttct cccagtacga aaggacaaga gaaataaggc ctacttctct
3961 gaagcgcctt aagaccaata gatgaattta tctaaatcta gtaaatctaa ctccaatatt
4021 gcccaagaga cagggctttg ttagggtggc agagcccggt aattgtgcaa aacttaaact
4081 cttgtgtcca gaggttcaat tcctctccct agcatatgtt 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?

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

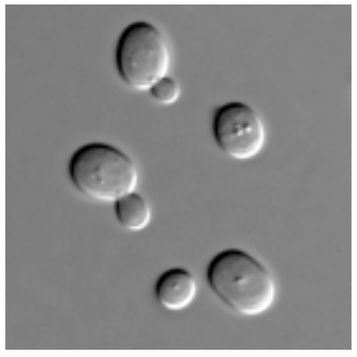
Help understand how to make animals and plants more hardy, resistant to disease, and understand their biology.

New technologies and larger genomes require new algorithms.

Other Sequenced Genomes

Arabidopsis thailiana

Callithrix jacchus
(marmoset)

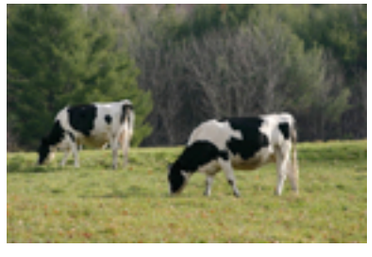


Saccharomyces cerevisiae
(baker's yeast)

Canis lupus familiaris
(dog)

Apis mellifera
(honey bee)

Drosophila melanogaster
(fruit fly)



Bos torus
(cow)

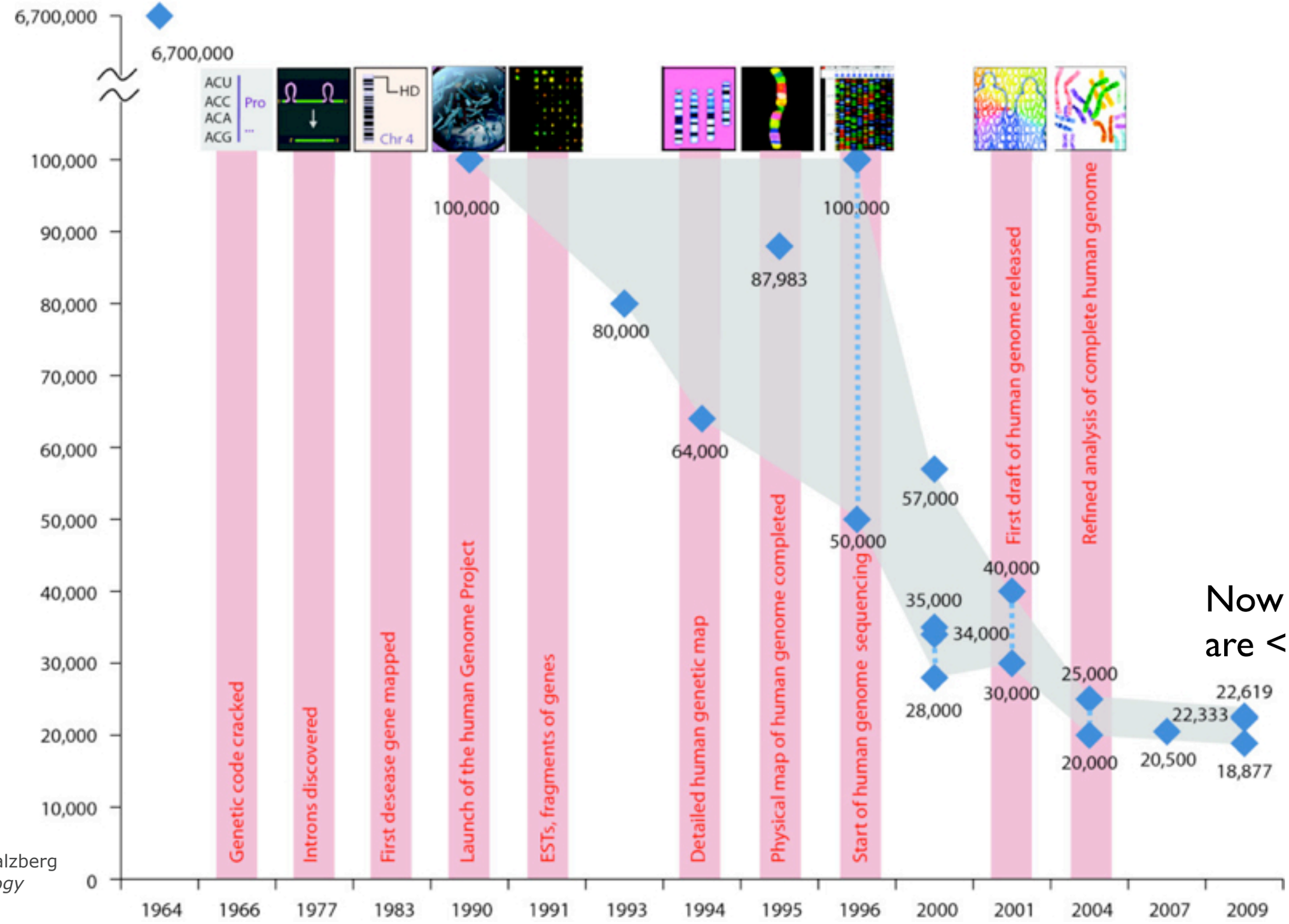
Equus caballus
(horse)

and many more...

Sequenced Eukaryotic Chromosomes

Estimates for the # of Human Genes

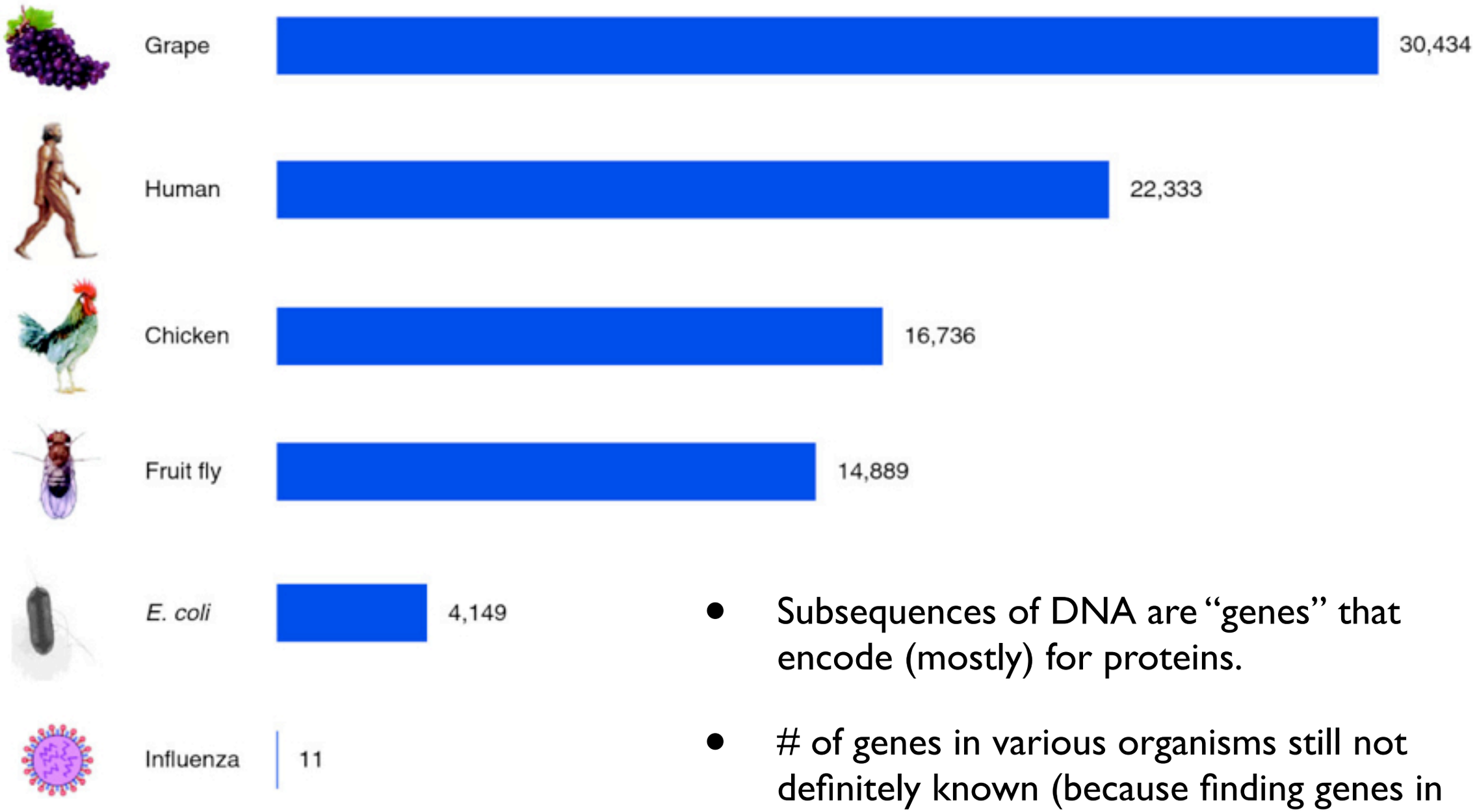
Before human genome sequence was available, many
(but not all) estimates for # of genes were high (> 80,000).



Now estimates are < 23,000.

Pertea and Salzberg
Genome Biology
2010 **11**:206

of Genes in Various Organisms



Pertea and Salzberg
Genome Biology 2010
11:206

- 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.

Example 2: Phylogenetics

- Algorithms for modeling how organisms evolved
- Algorithms on trees are central

Influenza Virus

(Toda et al., 2006)

Rapidly evolving (it's genome is mutating):

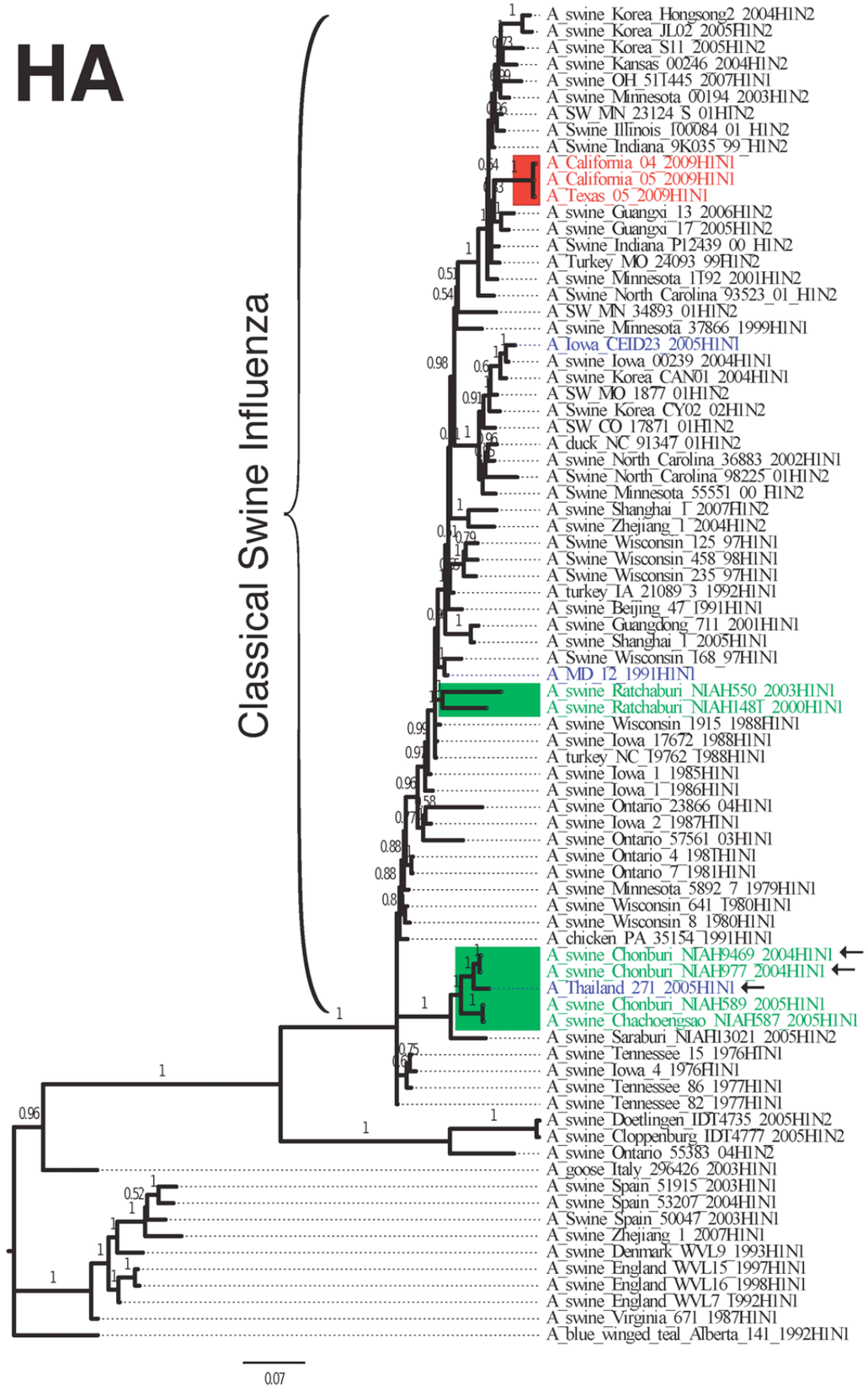
that's why you have to get a different flu shot every year

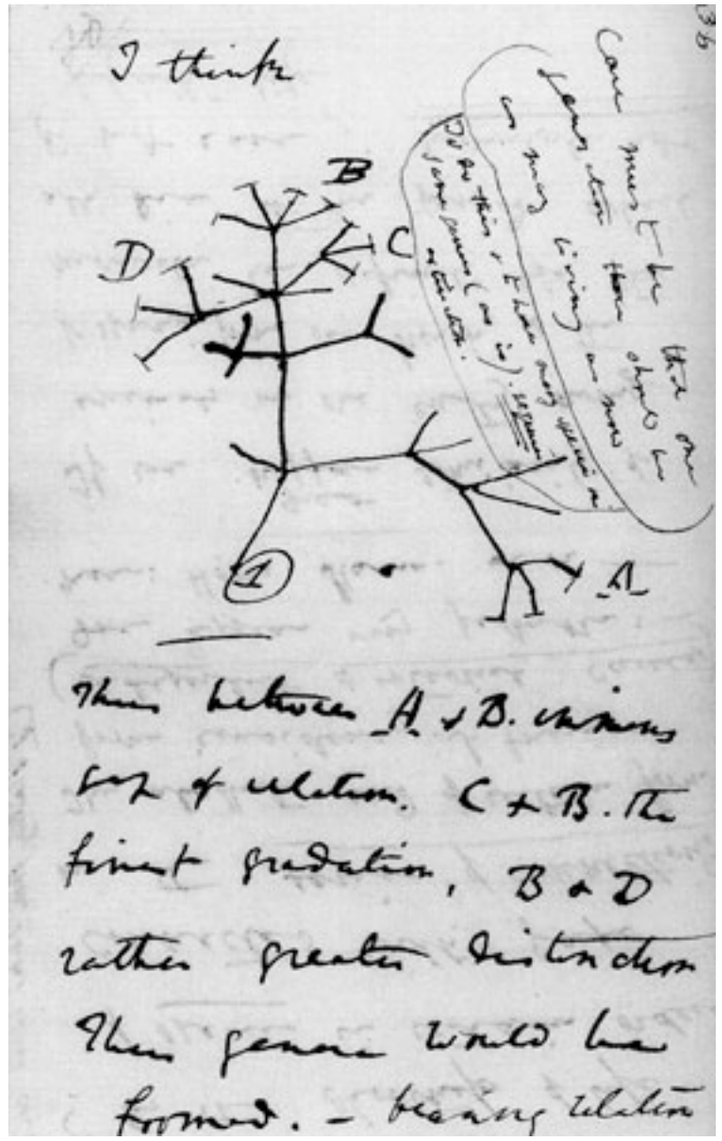
3 strains must be selected each year to include in the vaccine.

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

Evolutionary Trees

HA





http://www.amnh.org/exhibitions/darwin/idea/treelg.php

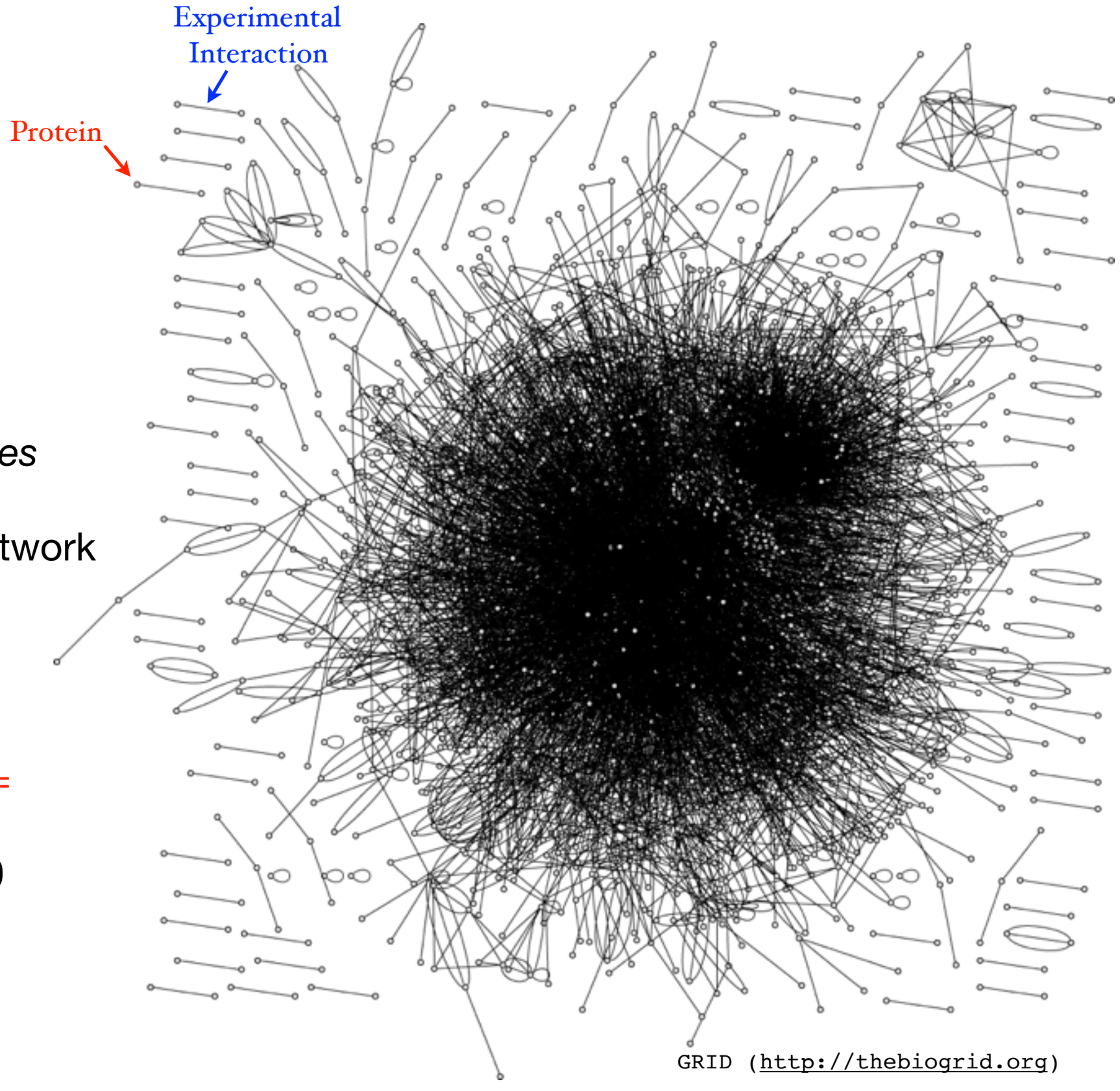
Example 3: Network Biology

- Algorithms for modeling how cellular components work together
- Graph algorithms are central

Yeast (aka *Saccharomyces cerevisiae)* interaction network

8,742 edges

3113 nodes (= proteins) (out of ~6,000 genes)



Network Questions

- Can we find groups of molecules that work together to perform a function?

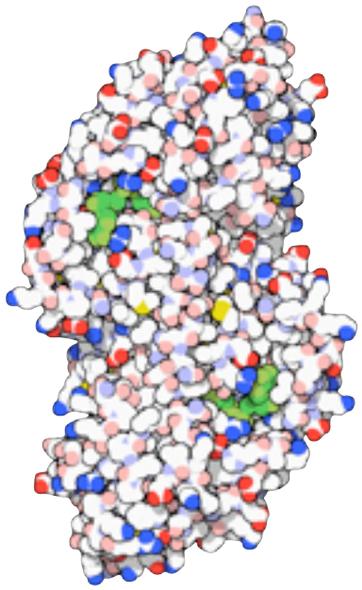
- Can we use interactions between proteins to predict what a protein does?

- Can we determine the how such a complex system evolved?

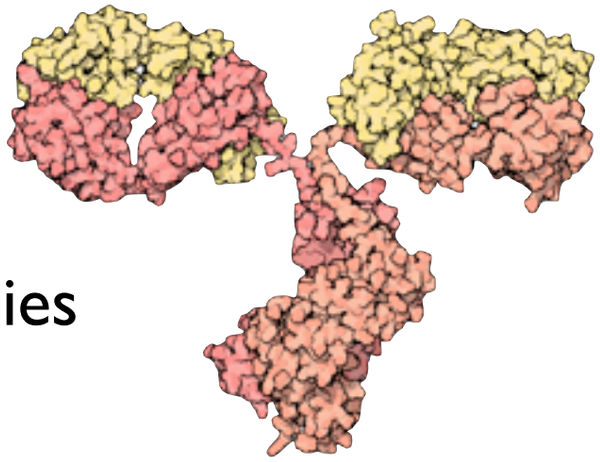
Example 4: Protein Structure

- Predicting the 3D structure of proteins
- Optimization algorithms are central

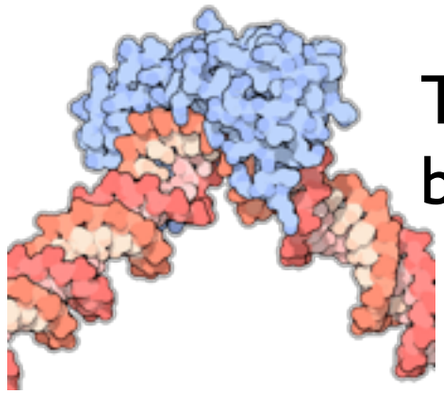
Examples of Proteins



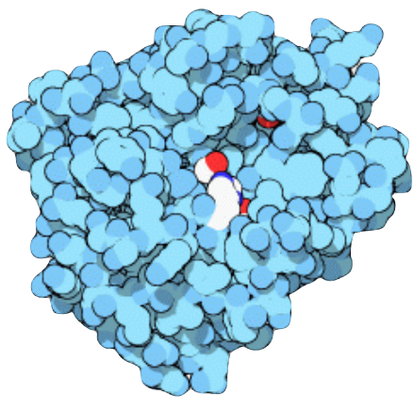
Alcohol dehydrogenase



Antibodies



TATA DNA binding protein



Trypsin: breaks down other proteins

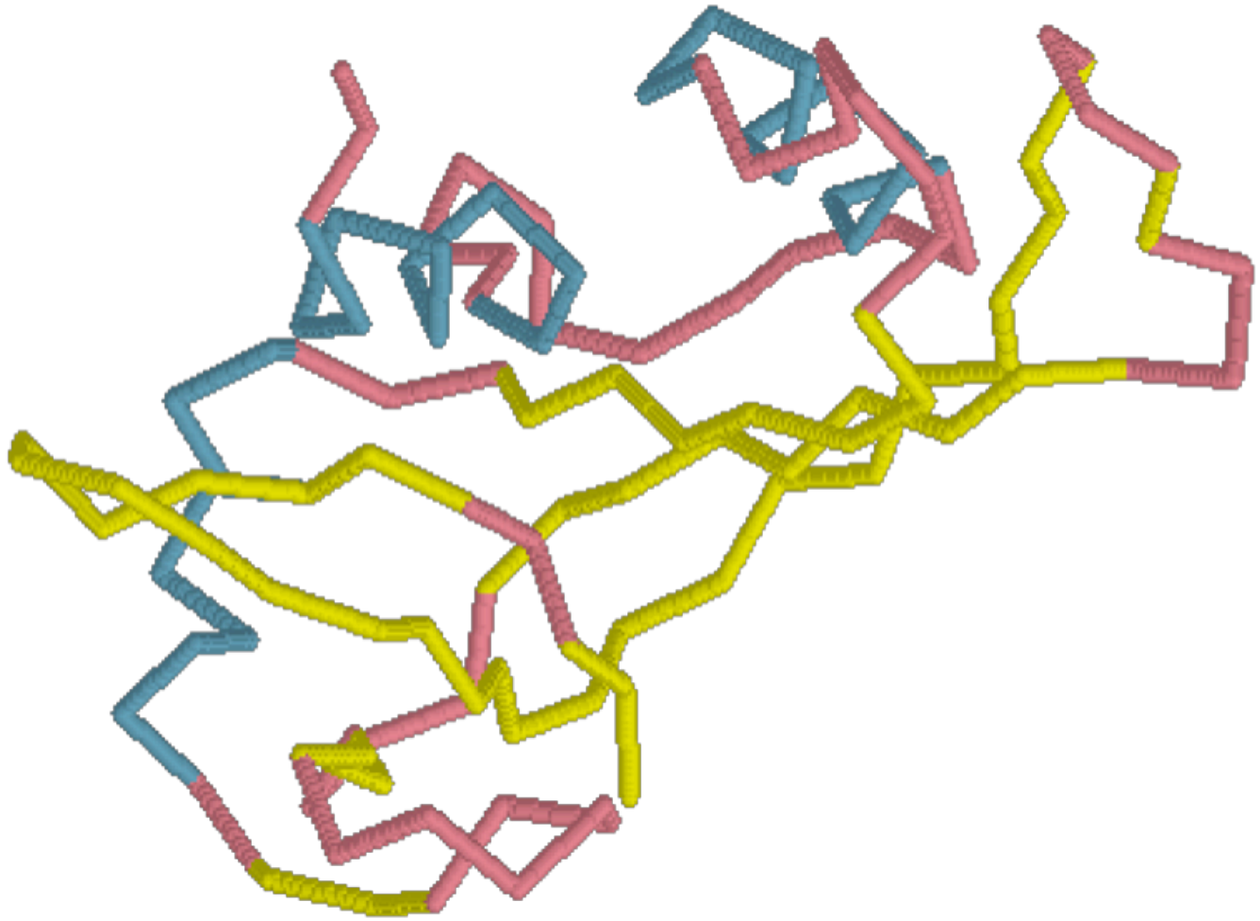
Collagen: forms tendons, bones, etc.

Examples of "Molecules of the Month" from the Protein Data Bank
http://www.rcsb.org/pdb/

Predicting Structure

Given: ATGAAGATGATAGATGGGGCCCGACAG…

Determine:



Course Mechanics

- **Homework** (30%): mix of “theory” and programming assignments
  - Theory assignments submitted via Gradescope
  - Programming assignments submitted via Stepik and autograded.
  - You can use any programming language.
  - NO LATE ASSIGNMENTS ACCEPTED

- **Exams** (40%): 2 in-class midterms (10% each) + final (20%)
  - Midterm 1: Thursday, February 14
  - Midterm 2: Tuesday, April 2
  - Final: TBD by the university (do not schedule flights until this is posted)

Course Mechanics, cont.

- **Project** (20%): more details later in the semester
  - Will involve applying ideas we've learned to some real data
  - Write-up + in-class presentation

- **Participation** (10%):
  - Attendance of lectures is required
  - You are allowed 3 missed classes
  - New material will be introduced in recitations.
  - **NO UNAPPROVED USE OF ELECTRONICS IN CLASS**

- **Mini-Lectures:**
  - Two lectures will be on-your-own watching of a computational biology lecture from a menu we will provide.
  - You will write a summary of these lectures in lieu of attending the two lectures before the midterms.

Academic Honesty

- You may discuss programming assignments with classmates

- **However, you must not share or show or see the code of your classmates. You must write your own code entirely.**

- You may never use, look at, study, or copy any answers from previous semesters of this course or from the internet.

- All class work should be done independently unless explicitly indicated on the assignment handout.

Prerequisites

- You should be comfortable programming
- We generally don't assume any biological background – we'll introduce the relevant biology as it is needed
- Mathematics: we assume elementary math & logical thinking & some familiarity with probability
  - We'll review probability facts as needed
  - Assume you are familiar with big-O notation

Questions?