

Lecture 24: Bioinformatics

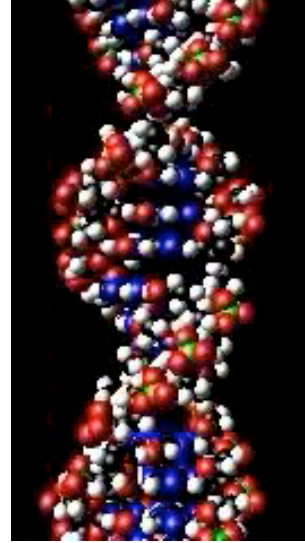
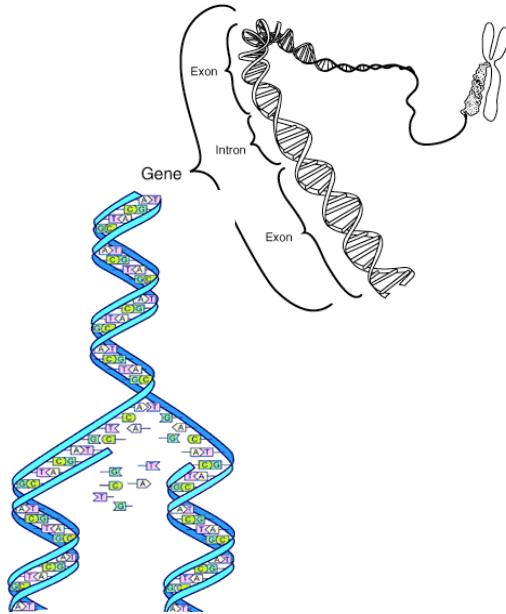
CS442: Great Insights in Computer Science
Michael L. Littman, Spring 2006

What is Bioinformatics?

Bioinformatics or **computational biology** involves the use of techniques from applied mathematics, informatics, statistics and computer science to solve problems in the biological sciences. Research in computational biology often overlaps with systems biology.

- **Bioinformatics:** techniques
- **Computational biology:** hypothesis testing
- Both use mathematical tools to extract information from data

The Biology Picture (video)

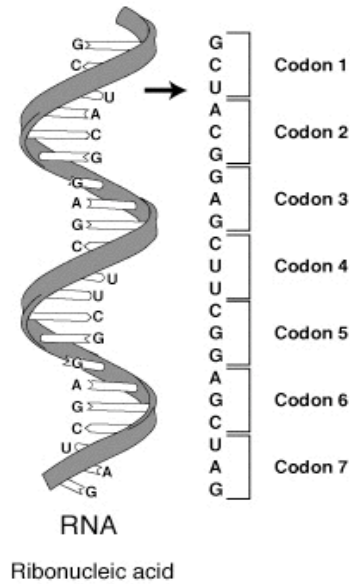


Gene Sequencing – WHY?

- DNA encodes the necessary information for living things to survive and reproduce.
- It is useful in research into why and how organisms live.
- Can help us understand, identify, diagnose and potentially treat genetic diseases.

Gene Sequencing Process

- We want genes. DNA is formed of nucleotides. Triples of nucleotides form a codon.
- Translations (interesting parts) starts with a special codon, named START and ends with another special codon named STOP (Terminator).

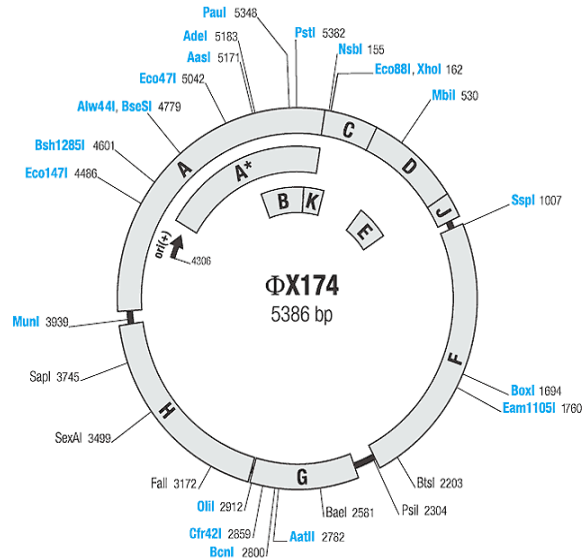


Gene Sequencing Process 2

- So, what we have to do is search for the START codon and read all codons until the Terminator.
- This process gives us the Translation (the interesting part).

Genes, Proteins, Computers

- Ok, biology hasn't changed much in the last 30 years
- First genome to be sequenced : phi-x174 phage in 1977
- Small amount of DNA
- 11 genomes
- 5386 base pairs



How Big is the Problem?

- Usually, the job is to find certain patterns in the genome.
- This is 0.57% of the information (amino acids) for *Y.pestis*.

```
>gi|22123923|ref|NP_667346.1| flavodoxin [Yersinia pestis KIM]
MADITLISGSTLGSAEYVAEHLADKLEEAGFSTEILHGPELDELTLNGLWLVITVSTHGAGDLPDMLQPLL
EQIEQQKPDLSQVRFVGAAGVLSSEYDTFCGAIKLDQQLIAQGAQRLGEILLEIDVIOHEIPEDPAEIUVK
DWINLL
>gi|22123924|ref|NP_667347.1| transcriptional regulator [Yersinia pestis KIM]
MSEIYQIDNLDRLKALMENARTPYAELAKNLAVSPGTIHVRVEKMRQAGIITAACVHVNPQKLGVDVC
CFIGIILKSAKDYPSALKKLESLEEVVEAYTTGHYSIFIKVMCKSIDALQQVLINKIQTIDEIQSTETL
ISLQNPIMRTIVP
>gi|22123925|ref|NP_667348.1| asparagine synthetase [Yersinia pestis KIM]
MKKQFIQKQQQISFVKSFFSRQLEQQGLIEVQAPILSRVGDGTQDNLGSEKAVQVKVSLPDSTFEVV
HSLAKWKRKTLGRFDGADQGVYTHMKALRPDEDRLSAIHSVYVDQWDWERVMGDGERNLAYLKSTVNKI
YAAIKETEAAISAEFGVKPFLPDHIQFIHSESLRARFPDLDAKGRERAIKELGAVFLIGIGKLDAGQS
HDVRAPDYDDWTSPSAEGFSGLNGLIIVWNPILEDAFEISSMGIKRVDAEALKRQLALTGDEDRLLELWHQ
SLLRGEMPQTIGGGIGQSRLVMLLLQKQHIGQVQCGVWGPEISEKVDGLL
```

People Need Help!

- But what kind...?

What Exactly Is Needed?

- Solve a formal problem.
- Do a lot of calculations.
- Maybe even automatically extract patterns in our data.
- Perhaps something that will do the work while we're tackling other big problems.

We need...



Actually, More Like...



Some Examples

- GBio
 - GPSA - “Grid Protein Sequence Analysis” : developing a grid portal devoted to Bioinformatics: <http://gpsa.ibcp.fr> .
 - GriPPS - “Grid Protein Pattern Scanning” : French ACI GRID project studying protein pattern/profile scanning application on the grid: <http://gripps.ibcp.fr> .
- Folding@home
 - <http://folding.stanford.edu/>.

Demo

- Other resources ...
- The BioMaPS Institute
 - <http://biomaps.rutgers.edu/>.
- Protein Data Bank @ Rutgers
 - <http://www.pdb.org> .

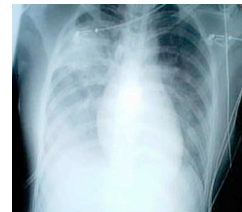
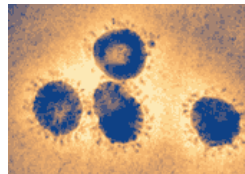
SARS Virus: Wikipedia

Severe acute respiratory syndrome (SARS) is an [atypical pneumonia](#) that first appeared in November 2002 in [Guangdong Province](#), in the city of Foshan, of the [People's Republic of China](#). The disease is now known to be caused by the [SARS coronavirus](#) (SARS CoV), a novel [coronavirus](#).

SARS was first reported in Asia in February 2003. Over the next few months, the illness spread to more than two dozen countries in North America, South America, Europe, and Asia before the SARS global outbreak of 2003 was contained. According to the World Health Organization (WHO), a total of 8098 people worldwide became sick with SARS during the 2003 outbreak. 774 of these died. SARS did not spread more widely in the community in the United States.

After the [Chinese government](#) suppressed news of the SARS outbreak, the disease spread rapidly, reaching [Hong Kong](#) and [Vietnam](#) in late February 2003, and then to other countries via international travelers. The last case in this outbreak occurred in June 2003. There were a total of 8437 known cases of the disease, with 813 deaths (a [mortality rate](#) of around 10 percent).

In May 2005, the [New York Times](#) reported that "not a single case of severe acute respiratory syndrome has been reported this year or in late 2004. It is the first winter without a case since the initial outbreak in late 2002. In addition, the epidemic strain of SARS that caused at least 813 deaths worldwide by June of 2003 has not been seen outside a laboratory since then." [\[1\]](#)



Computer Virus

- Can be downloaded at http://ybweb.bcgsc.ca/sars/TOR2_finished_genome_assembly_290403.fasta

```
>TOR2_finished_genome_assembly_290403 Release 3
ATATTAGGTTTTTACCTACCCAGGAAAAGCCAACCAACCTCGATCTCTTG
TAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTAGCTGTCGCTCGGC
TGCATGCCTAGTGCACCTACGCAGTATAAACAATAATAAATTTTACTGTC
GTTGACAAGAAACGAGTAACTCGTCCCTCTTCTGCAGACTGCTTACGGTT
TCGTCCGTGTTGCAGTCGATCATCAGCATACTAGGTTTCGTCCGGGTGT
GACCGAAAGGTAAGATGGAGAGCCTTGTTCTTGGTGTCAACGAGAAAACA
CACGTCCAACCTCAGTTTGCCTGTCCTTCAGGTTAGAGACGTGCTAGTGCG
TGGCTTCGGGGACTCTGTGGAAGAGGCCCTATCGGAGGCACGTGAACACC
TCAAAAATGGCACTTGTGGTCTAGTAGAGCTGGAAAAAGGCGTACTGCCC
CAGCTTGAACAGCCCTATGTGTTTCATTAACGTTTCTGATGCCTTAAGCAC
CAATCACGGCCACAAGGTCGTTGAGCTGGTTGCAGAAATGGACGGCATTG
AGTACGGTCGTAGCGGTATAACACTGGGAGTACTCGTGCCACATGTGG ...
```

Wrap Python Around It

```
sars = ['A', 'T', 'A', 'T', 'T', 'A', 'G', 'G', 'T', 'T', 'T', 'T', 'T',
'A', 'C', 'C', 'T', 'A', 'C', 'C', 'C', 'A', 'G', 'G', 'A', 'A', 'A',
'A', 'G', 'C', 'C', 'A', 'A', 'C', 'C', 'A', 'A', 'C', 'C', 'T', 'C',
'G', 'A', 'T', 'C', 'T', 'C', 'T', 'T', 'G', 'T', 'A', 'G', 'A', 'T',
'C', 'T', 'G', 'T', 'T', 'C', 'T', 'C', 'T', 'A', 'A', 'A', 'A', 'C', 'G',
'A', 'A', 'C', 'T', 'T', 'G', 'T', 'G', 'G', 'T', 'C', 'T', 'A', 'G',
'T', 'G', 'T', 'A', 'G', 'C', 'T', 'G', 'G', 'T', 'T', 'C', 'A', 'G',
'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A',
'A', 'A']
```

```
len(sars)
# Base pairs? 29751
sum([i == 'G' for i in sars])
# How many "G" in SARS? 6187
```

Codons

- 4 symbols (How many bits would you need?)
- 3 symbols (codon) work together to encode an amino acid. (How many codons?)
- Codons code for the amino acid alphabet (20 symbols).
- There are also two “syntactic” symbols (START, STOP)

Amino Acid Code (DNA)

TTT Phenylalanine (Phe) TTC Phe TTA Leucine (Leu) TTG Leu	TCT Serine (Ser) TCC Ser TCA Ser TCG Ser	TAT Tyrosine (Tyr) TAC Tyr TAA STOP TAG STOP	TGT Cysteine (Cys) TGC Cys TGA STOP TGG Tryptophan (Trp)
CTT Leucine (Leu) CTC Leu CTA Leu CTG Leu	CCT Proline (Pro) CCC Pro CCA Pro CCG Pro	CAT Histidine (His) CAC His CAA Glutamine (Gln) CAG Gln	CGT Arginine (Arg) CGC Arg CGA Arg CGG Arg
ATT Isoleucine (Ile) ATC Ile ATA Ile ATG Methionine/START	ACT Threonine (Thr) ACC Thr ACA Thr ACG Thr	AAT Asparagine (Asn) AAC Asn AAA Lysine (Lys) AAG Lys	AGT Serine (Ser) AGC Ser AGA Arginine (Arg) AGG Arg
GTT Valine (Val) GTC Val GTA Val GTG Val	GCT Alanine (Ala) GCC Ala GCA Ala GCG Ala	GAT Aspartic acid (Asp) GAC Asp GAA Glutamic acid (Glu) GAG Glu	GGT Glycine (Gly) GGC Gly GGA Gly GGG Gly

Translation

GCATGCCTAGTGCACTACGCAGTATAAA CAATAAT

- ATG: START
- CCT: Pro
- AGT: Ser
- GCA: Ala
- ACG: Thr
- CAG, CAA: Gln
- TAT: Tyr
- AAA: Lys
- TAA: STOP

"Gene" Finder

```
# returns the first position (starting at i)
# containing a start codon or -1 if none
def findStart(dna,i):
    while i <= len(dna)-3:
        triple = dna[i] + dna[i+1] + dna[i+2]
        if triple == 'ATG': return i
        i = i + 1
    return -1
```

```
def findStop(dna,i):
    while i <= len(dna)-3:
        triple = dna[i] + dna[i+1] + dna[i+2]
        if triple == 'TAA' or triple == 'TAG' or triple == 'TGA':
            return i
        i = i + 3
    return -1
```

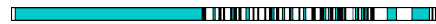
```
def translate(dna):
    i = 0
    while i <= len(dna)-3:
        i = findStart(dna, i)
        if i == -1:
            print "no start found"
            return
        print "found start at ", i
        i = findStop(dna, i+3)
        if i == -1:
            print "no stop found"
            return
        print "    stop at ", i
        i = i + 3
```

Output

found start at 103
stop at 133
found start at 264
stop at 13410
found start at 13455
stop at 13515
found start at 13548
stop at 13566
found start at 13598
stop at 21482
found start at 21491
stop at 25256
found start at 25267
stop at 26089
found start at 26116
stop at 26344
found start at 26397
stop at 27060

found start at 27073
stop at 27262
found start at 27272
stop at 27638
found start at 27641
stop at 27653
found start at 27706
stop at 27769
found start at 27778
stop at 27895
found start at 27958
stop at 27988
found start at 28053
stop at 28098
found start at 28119
stop at 29385
found start at 29394
stop at 29424

found start at 29474
stop at 29489
found start at 29544
stop at 29625
found start at 29648
stop at 29660
found start at 29695
stop at 29698
found start at 29722
no stop found



Next Time

- Statistical Natural Language (crossword puzzles)