Introduction to Bioinformatics
BINF 630

D. Andrew Carr

Course information

- http://binf.gmu.edu/sjafri/binf630
- Professor:
  - Saleet Jafri, Professor and Chair.
    Ph.D. Biomedical Sciences, City University of New York/Mount Sinai School of Medicine, 1993.
    Cellular Signaling, Cardiac Physiology, High-Performance Computing and Modeling.
    - Office: 703-993-8420
    - Email: sjafri@gmu.edu
- Computer Systems Administrator
  - Chris Ryan
    - Office: 703-993-8394
    - Email: cryan1@gmu.edu

Grading Scheme

- Mid-Term 30%
- Final 30%
- Homework 40%

Course Book

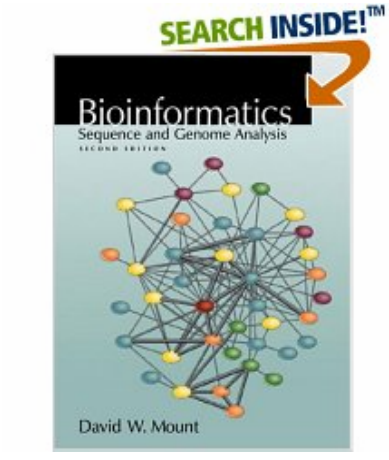


Image from www.amazon.com

Mount, David W.
Bioinformatics: Sequence and Genome Analysis *Second Edition*

Cold Springs Harbor Laboratory Press, 2004

Bioinformatics

Bioinformatics is a field that deals with biological information, data, and knowledge, and their storage, retrieval, management, and optimal use for problem solving and decision making.

NIH working definition of bioinformatics and computational biology (July 2000)

The NIH Biomedical Information Science and Technology Initiative Consortium agreed on the following definitions of bioinformatics and computational biology recognizing that no definition could completely eliminate overlap with other activities or preclude variations in interpretation by different individuals and organizations.

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

OVERLAPPING DISCIPLINES

COMPUTATIONAL STRUCTURAL BIOLOGY

COMPUTATIONAL MOLECULAR BIOLOGY

BIOINFORMATICS

GENOMICS

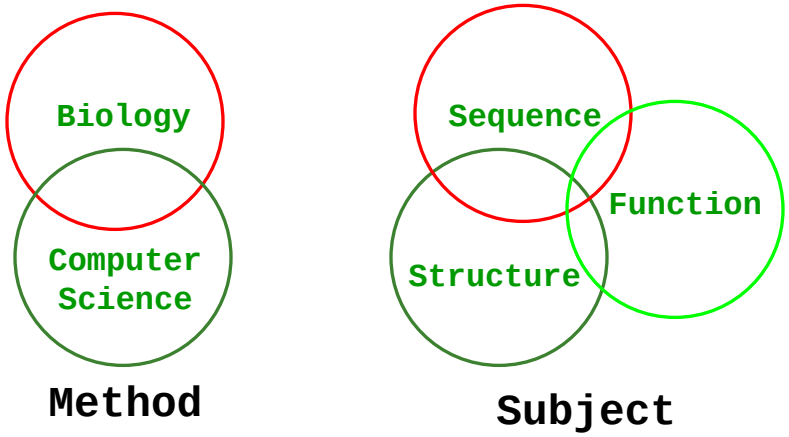
STRUCTURAL GENOMICS

PROTEOMICS

COMPUTATIONAL BIOLOGY

BIOENGINEERING

Bioinformatics



What is Informatics?

in•for•mat•ics (in′fər mat′iks) *n.* (*used with a sing. v.*)
the study of information processing; computer science.
[trans. of Russ informátika (1966); see INFORMATION, -ICS]

Random House Unabridged Dictionary

What is information?

- Definition: knowledge or intelligence communicated, received or gained
- Information is a decrease in uncertainty.

Information

General

knowledge or intelligence communicated, received or gained

Information theory

indication of the number of possible choices

```
Th_ qui_k br_wn _ox ju_ps ov__ th_ laz_ d_g
```

Information

```
Th_ qui_k br_wn _ox ju_ps ov__ th_ laz_ d_g
```

```
The quick brown fox jumps over the lazy dog
```

Shannon Entropy

- Claude E. Shannon defined entropy as a measure of the average information content associated with a random outcome.
- Shannon information entropy relates to the amount of *uncertainty* about an event associated with a given probability distribution.
- Shannon Entropy:
  - The entropy of the event *x* is the sum, over all possible outcomes *i* of *x*, of the product of the probability of outcome *i* times the log of the inverse of the probability of *i*

$$H(x) = -\sum_{i=1}^{M} P_i \log_2 P_i$$

- *http://en.wikipedia.org/wiki/Information_entropy*

Example 1 of uncertainty as applied to Shannon entropy.

- Alphabet #1:
  - {A,B,C,D,E,F,G,H,I,J,K,L}
- Alphabet #2
  - {A,A,A,A,A,A,C,B,D,E,F,G}
- In case #1 uncertainty at selecting any one character is maximal.
- In case #2 there are more {A}'s and less uncertainty
- Selecting and removing a character from each set provides information.
  - More information is gained from case #1 because the uncertainty is higher.

Information and uncertainty

Information is a decrease in uncertainty

$$\log_2 (M) = - \log_2 (M^{-1}) = - \log_2 (P)$$

Shannon's formula for uncertainty

$$H(x) = - \sum_{i=1}^{M} P_i \log_2 P_i$$

only infrmatn esentil to understandn mst b tranmitd

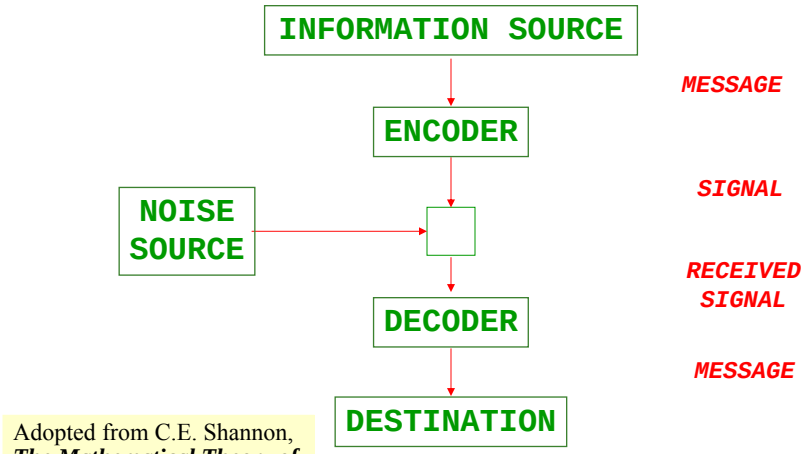
Communication

Fundamental problem of communication:

reproducing at one point either exactly or approximately a message selected at another point

The Mathematical Theory of Communication
Claude Shannon and Warren Weaver

Communication system



Adopted from C.E. Shannon, ***The Mathematical Theory of Communication***, 1949

Communication system duality

"This duality can be pursued further and is related to the duality between past and future and the notions of control and knowledge. Thus we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it."

C. E. Shannon (1959)

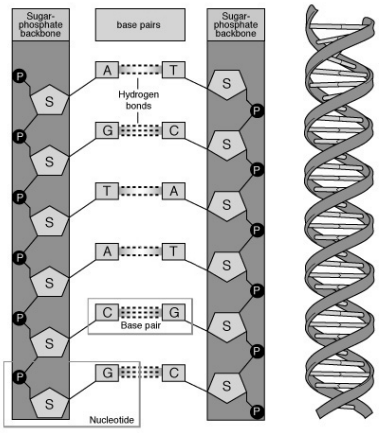
The signal of bioinformatics

- The molecular components common to all life
  - Nucleic Acids
    - DNA
    - RNA
  - Proteins

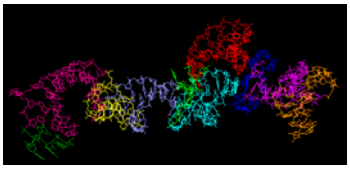
- Central dogma
  - DNA → RNA → Protein
    - Flow is unidirectional
      - Except reverse transcriptase (virus)

DNA

- DNA (deoxyribonucleic acid)
  - Helix formed by pairing of bases
  - Four bases
    - Two complement pairs
      - (A) adenine ~ purine
      - (T) thymine ~ pyrimidine
      - (G) guanine ~ purine
      - (C) cytosine ~ pyrimidine
  - Location:
    - Nucleus of Eukaryotes
    - Prokaryotes
    - Achaebacteria



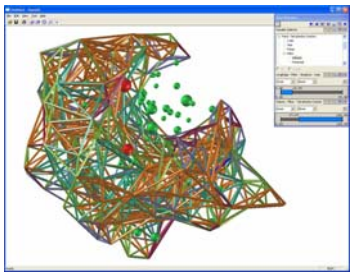
RNA



Haloarcula marismortui : http://rose.man.poznan.pl/5SData/

- Ribonucleic acid
  - Formed with ribose not 2'-deoxyribose as sugar
  - Does not form double helix
  - Can have a complicated 3D structure
  - Takes on different forms and functions
    - mRNA ~ messenger RNA
      - Is the transcribed signal that travels to ribosome for translation
    - tRNA ~ transfer RNA
      - Carries amino acid to ribosome
    - rRNA ~ ribosomal RNA
      - Combines with protein to form the ribosome
    - sRNA ~ small RNA
      - Facilitate other functions within the cell
    - Others …

Proteins

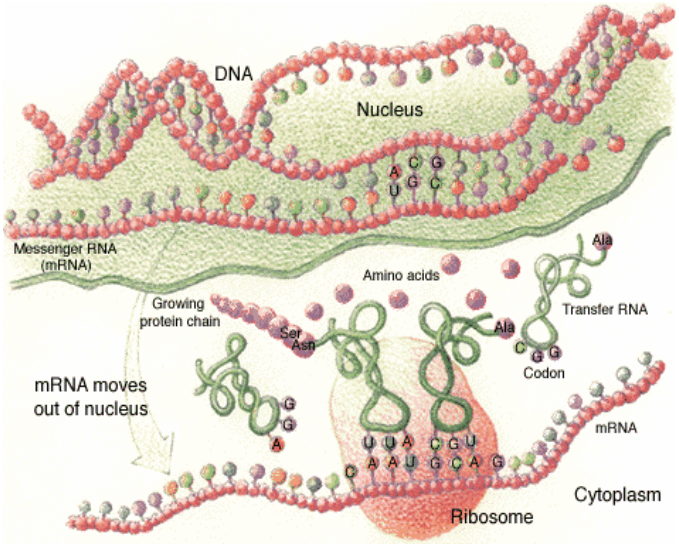


Top Chime image of 1GBG

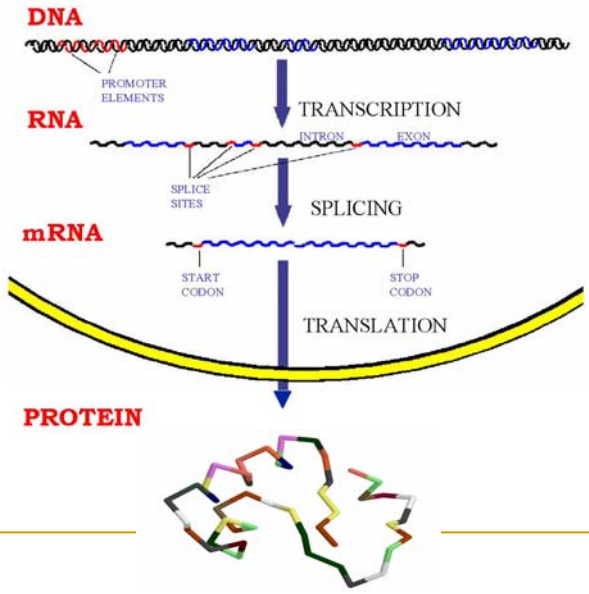
GLISTEN image of 1GBG

- Chains of amino acids
- Functions:
  - Structural proteins
  - Enzymes
  - Facilitate transport
  - Participate in cell signaling
- Structure = Function…
  - Sequence → Structure?
- Typical size ~300 residues

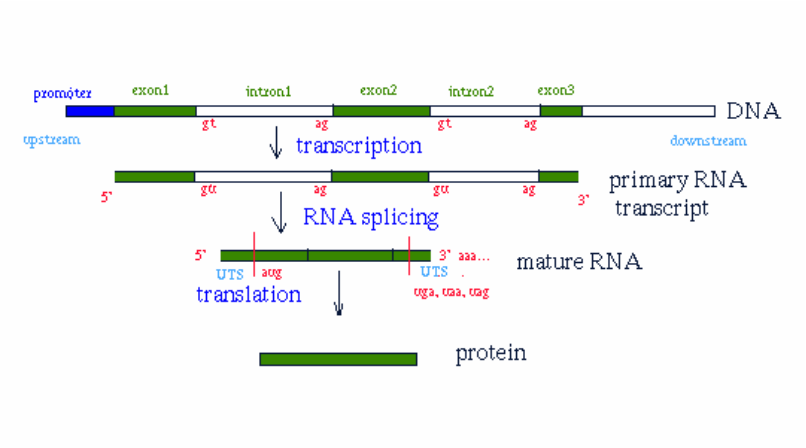
Cell Informatics



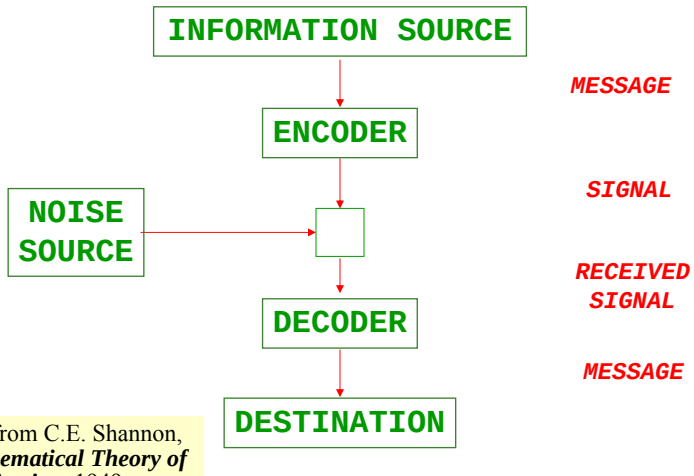
Cell Informatics



Cell Informatics



Communication system

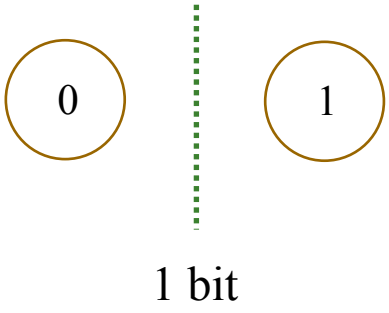


Adopted from C.E. Shannon, ***The Mathematical Theory of Communication***, 1949

Back to informatics…

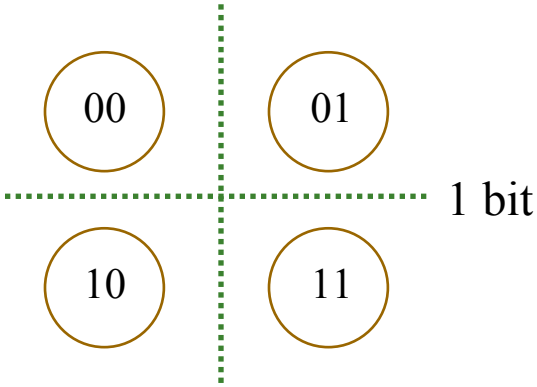
- What is the information content of DNA and RNA?
- What is the information content of a protein sequence?

Information Theory

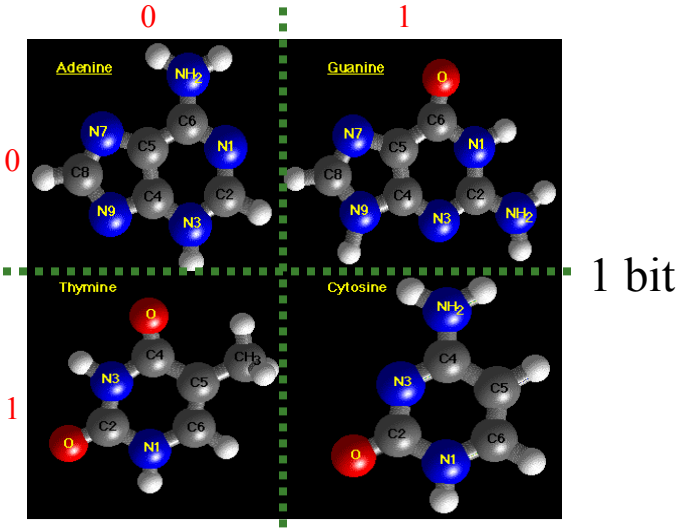


1 bit

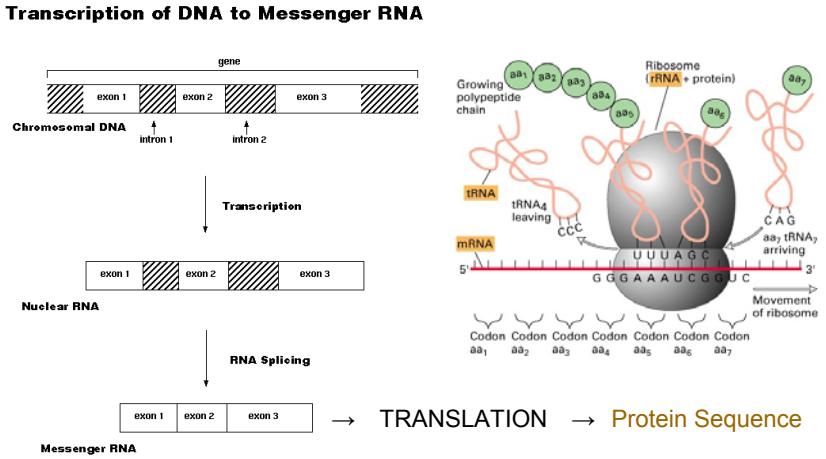
Information Theory



Nucleotide permutation space



Translation from mRNA to Protein Sequence



→ TRANSLATION → Protein Sequence

Translation: mRNA to Protein Sequence

- {A,T,C,G} can be encoded by 2 bits
  - 1 base position
    - $4^1 = 4$
  - 2 bases
    - $4^2 = 16$
    - Not enough
  - 3 bases
    - $4^3 = 64$
    - Too many
- Redundancy ~ many codons to one amino acid
- Error correcting code
- Third position wobble

Standard genetic code

TTT	F Phe	TCT	S Ser	TAT	Y Tyr	TGT	C Cys
TTC	F Phe	TCC	S Ser	TAC	Y Tyr	TGC	C Cys
TTA	L Leu	TCA	S Ser	TAA	* Stop	TGA	* Stop
TTG	L Leu	TCG	S Ser	TAG	* Stop	TGG	W Trp
CTT	L Leu	CCT	P Pro	CAT	H His	CGT	R Arg
CTC	L Leu	CCC	P Pro	CAC	H His	CGC	R Arg
CTA	L Leu	CCA	P Pro	CAA	Q Gln	CGA	R Arg
CTG	L Leu	CCG	P Pro	CAG	Q Gln	CGG	R Arg
ATT	I Ile	ACT	T Thr	AAT	N Asn	AGT	S Ser
ATC	I Ile	ACC	T Thr	AAC	N Asn	AGC	S Ser
ATA	I Ile	ACA	T Thr	AAA	K Lys	AGA	R Arg
ATG	M Met	ACG	T Thr	AAG	K Lys	AGG	R Arg
GTT	V Val	GCT	A Ala	GAT	D Asp	GGT	G Gly
GTC	V Val	GCC	A Ala	GAC	D Asp	GGC	G Gly
GTA	V Val	GCA	A Ala	GAA	E Glu	GGA	G Gly
GTG	V Val	GCG	A Ala	GAG	E Glu	GGG	G Gly

Noise Sources

- Vector sequences
- Heterologous sequences
- Rearranged & deleted sequences
- Repetitive element contamination
- Sequencing errors / Natural polymorphisms
- Frameshift errors

Standard genetic code

```
  AAs = FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = ---M---------------M---------------M----------------------------
 Base1 = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
 Base2 = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
 Base3 = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

Frameshift Errors

```
ATGAAATTTGGAAACTTCCTTCTCACTTATCAGCCACCTGAGCTATCTCAGACCGAAGTGATGAAGCGATTGGTTAATCT

          5'3'Frame1 MKFGNFLLTYQPPELSQTEVMKRLVN
          5'3'Frame2 -NLETSFSLISHLSYLRPK--SDWLI
          5'3'Frame3 EIWKLPSHLSAT-AISDRSDEAIG-S
          3'5'Frame1 RLTNRFITSV-DSSGG--VRRKFPNF
          3'5'Frame2 D-PIASSLRSEIAQVADK-EGSFQIS
          3'5'Frame3 INQSLHHFGLR-LRWLISEKEVSKFH
```

Comparative Sequence Sizes

- Watson and Crick measure sequence size as base pairs (bp)

Yeast chromosome 3	350,000
Escherichia coli (bacterium) genome	4,600,000
Largest yeast chromosome now mapped	5,800,000
Entire yeast genome	15,000,000
Smallest human chromosome (Y)	50,000,000
Largest human chromosome (1)	250,000,000
Entire human genome	3,000,000,000

Computation

- Login information:
  - Passwd to change password
- Course website
  - http://binf.gmu.edu/jafir/binf630/
- Second Vaisman's website.
  - http://binf.gmu.edu/vaisman/binf731/work/
- BiologyWorkbench
  - Research tool for bioinformaticians.
  - http://workbench.sdsc.edu/

Exercise 1

- Look at the information content of the two DNA sequences present on the web page.
  - BLAST them and look at the results
    - Briefly discuss BLAST and BiologyWorkbench.
- Translate them and look at the results….
  - Discuss the relationship to frame shift.
- What is the information content of the two sequences and how does it differ?