Biomolecular Informatics: Sequence to Structure to Function

Iosif Vaisman

`Email: Iosif_Vaisman@unc.edu`

`Class page`
`http://www.unc.edu/courses/bioc156`

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.

COMPUTATIONAL BIOLOGY

COMPUTATIONAL STRUCTURAL BIOLOGY

COMPUTATIONAL MOLECULAR BIOLOGY

BIOINFORMATICS

GENOMICS

STRUCTURAL GENOMICS

PROTEOMICS

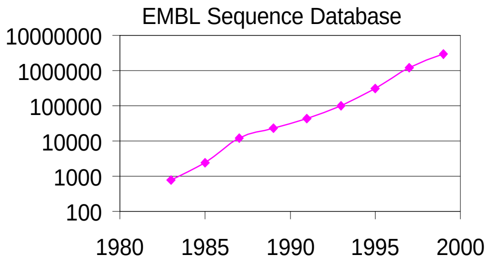
…

…

Comparative Sequence Sizes

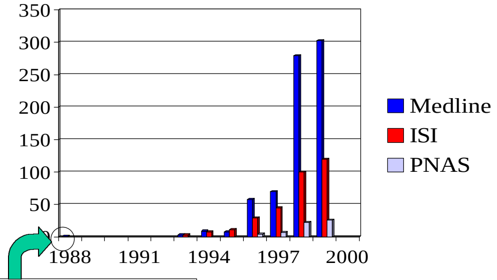
• Yeast chromosome 3	350,000
• Escherichia coli (bacterium) genome	4,600,000
• Largest yeast chromosome now mapped	5,800,000
• Entire yeast genome	15,000,000
• Smallest human chromosome (Y)	50,000,000
• Largest human chromosome (1)	250,000,000
• Entire human genome	3,000,000,000

Dynamics of Database Growth



Bioinformatics bibliography

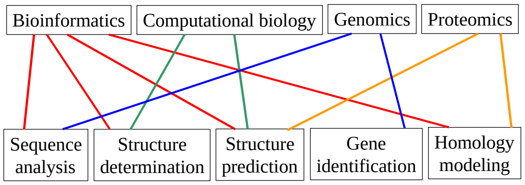
(papers with the word “bioinformatics” in title or abstract)



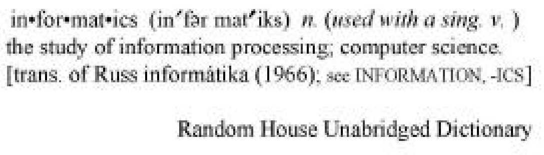
Liebman MN, Molecular modeling of protein structure and function: a bioinformatic approach.

J Comput Aided Mol Des 1988, **1**(4):323-41

Bioinformatics and related fields



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

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

Ae_h uz_ ko_ wm so_g oqr_it ypu_vn tr_e oj_

Information

Th_ qui_k br_wn _ox ju_ps ov__ th_ laz_ d_g

Ae_h uz_ ko_ wm so_g oqr_it ypu_vn tr_e oj_

The quick brown fox jumps over the lazy dog

Aedh uzh kox wm sobg oqrfit ypulvn tree ojc

Information and uncertainty

Information is a decrease in uncertainty

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

Shannon's formula for uncertainty

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

only infrmatn esentil to understandn mst b tranmitd

Information and uncertainty

Information is a decrease in uncertainty

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

Shannon's formula for uncertainty

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

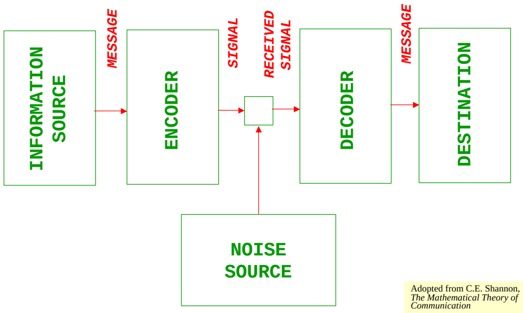
Communication

Fundamental problem of communication:

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

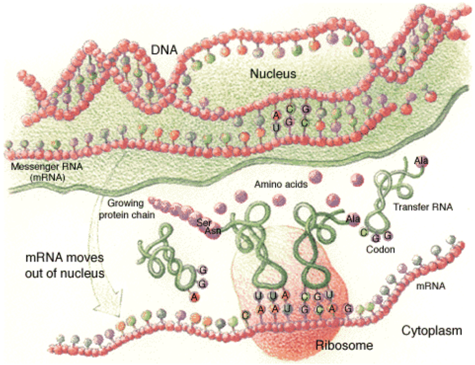
The Mathematical Theory of Communication
Claude Shannon and Warren Weaver

Communication system

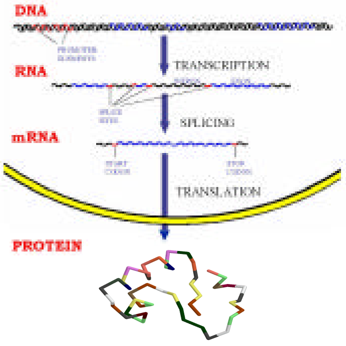


Adopted from C.E. Shannon, *The Mathematical Theory of Communication*

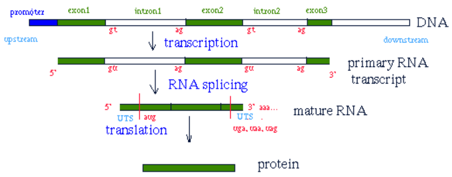
Cell Informatics



Cell Informatics



Cell Informatics



Hamming metric

The sum of bit changes necessary to move from one point in the permutation space to another point in the permutation space

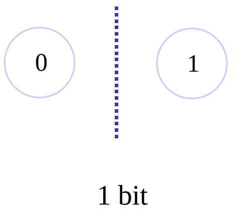
0000 and 0111 are separated by Hamming distance of 3:
0000 - 0001 - 0011 - 0111

Error correcting codes

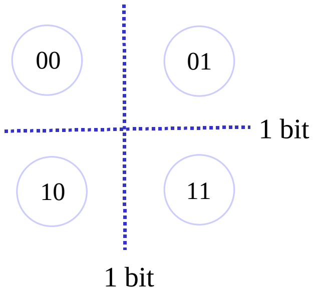
	a	b	c	d	e
a			▨		
b	▨				▨
c					
d		▨			
e				▨	

Code words ac, ba, be, db, ed
in the permutation space of
[a..e]x[a..e]

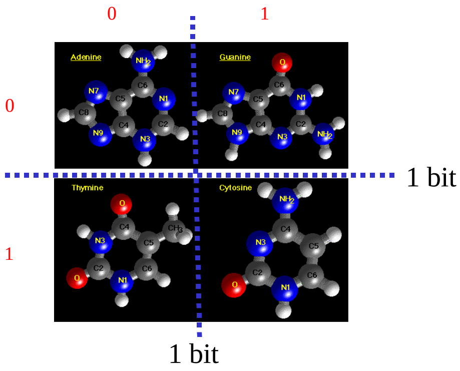
Information Theory



Information Theory



Nucleotide permutation space



Error Correcting Code

A - 00	A - 00000
G - 01	G - 01101
T - 10	T - 10110
C - 11	C - 11011
DNA base code	Error Correcting Code

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 * Ter	TGA * Ter
TTG L Leu	TCG S Ser	TAG * Ter	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 = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

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
```