BINF630/BIOL580/BINF401

Bioinformatics Methods

Iosif Vaisman

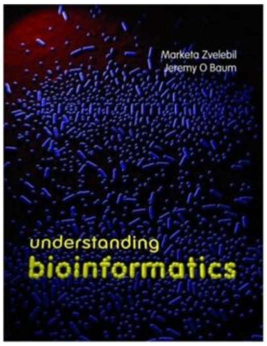
`Email: ivaisman@gmu.edu`

Spring 2016

Major focus areas

- Informatics infrastructure
- DNA and protein sequence analysis and genomics
- Protein structure and function analysis

Recommended book



Marketa J Zvelebil,
Jeremy O Baum

UNDERSTANDING BIOINFORMATICS

New York: Garland Science, 2008.

Class webpage

http://binf.gmu.edu/vaisman/binf630/

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.

Bioinformatics and Computational Biology

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.

COMPUTATIONAL BIOLOGY

COMPUTATIONAL STRUCTURAL BIOLOGY

COMPUTATIONAL MOLECULAR BIOLOGY

BIOINFORMATICS

GENOMICS

STRUCTURAL GENOMICS

PROTEOMICS

…

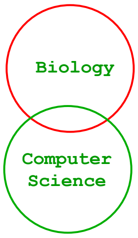
…

Omics sciences

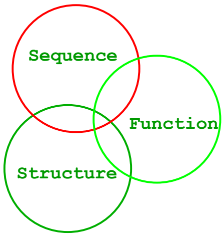
Connectomics
Cytomics
Epigenomics
Exposomics
Exomics
Genomics
Glycomics
Interferomics
Interactomics
Ionomics
Kinomics
Lipidomics
Mechanomics

Metabolomics
Metagenomics
Metallomics
ORFeomics
Organomics
Pharmacogenomics
Phenomics
Physiomics
Proteomics
Regulomics
Secretomics
Speecheomics
Transcriptomics

Bioinformatics

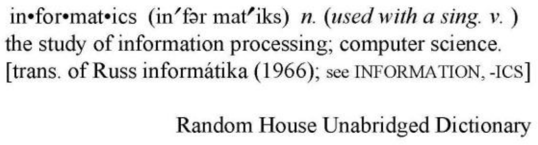


Method



Subject

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

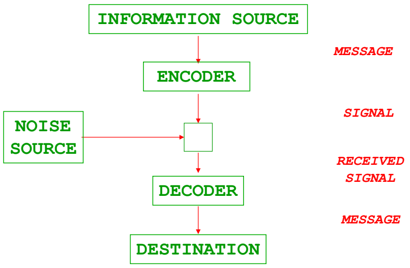
Communication

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



INFORMATION SOURCE

MESSAGE

ENCODER

SIGNAL

NOISE SOURCE

RECEIVED SIGNAL

DECODER

MESSAGE

DESTINATION

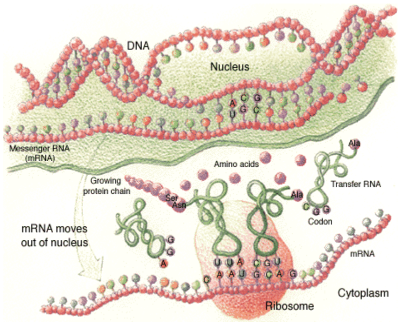
Adopted from C.E. Shannon, ***The Mathematical Theory of Communication***, 1949

Communication system duality

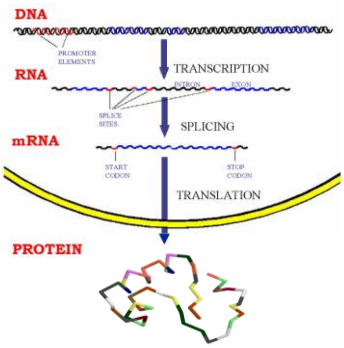
"This duality can be pursued further and is related to the duality between past and future and the notions of control and knowledge. Thus we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it."

C. E. Shannon (1959)

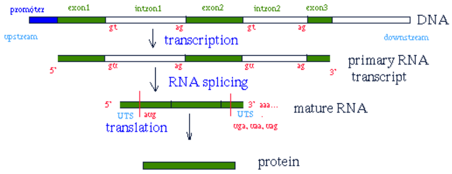
Cell Informatics



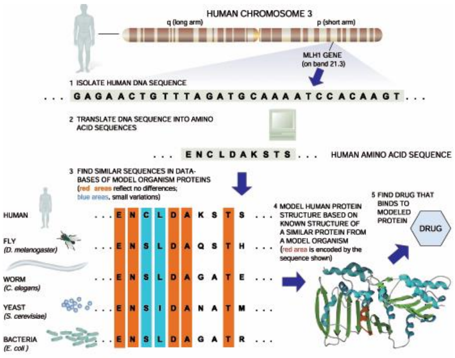
Cell Informatics



Cell Informatics

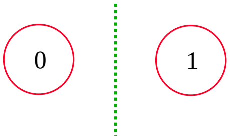


Sequence – structure – function


HUMAN CHROMOSOME 3
q (long arm)
p (short arm)
MLH1 GENE
(on band 21.3)
1 ISOLATE HUMAN DNA SEQUENCE
. . . G A G A A C T G T T T A G A T G C A A A A T C C A C A A G T . . .
2 TRANSLATE DNA SEQUENCE INTO AMINO ACID SEQUENCES
. . . E N C L D A K S T S . . .
HUMAN AMINO ACID SEQUENCE
3 FIND SIMILAR SEQUENCES IN DATA-BASES OF MODEL ORGANISM PROTEINS (red areas reflect no differences; blue areas, small variations)
HUMAN
. . . E N C L D A K S T S . . .
FLY (D. melanogaster)
. . . E N S L D A Q S T H . . .
WORM (C. elegans)
. . . E N S L D A G A T E . . .
YEAST (S. cerevisiae)
. . . E N S I D A N A T M . . .
BACTERIA (E. coli)
. . . E N S L D A G A T R . . .
4 MODEL HUMAN PROTEIN STRUCTURE BASED ON KNOWN STRUCTURE OF A SIMILAR PROTEIN FROM A MODEL ORGANISM (red area is encoded by the sequence shown)
5 FIND DRUG THAT BINDS TO MODELED PROTEIN
DRUG

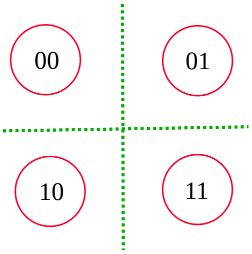
Luscombe *et al.*, 2001

Information Theory



1 bit

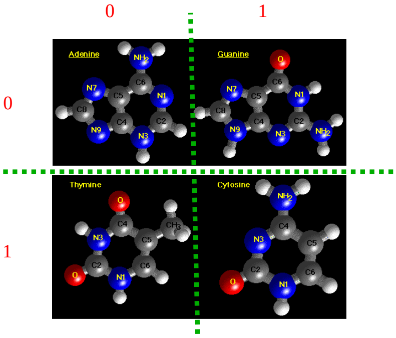
Information Theory



1 bit

1 bit

Nucleotide permutation space



1 bit

1 bit

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

Error correcting codes

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

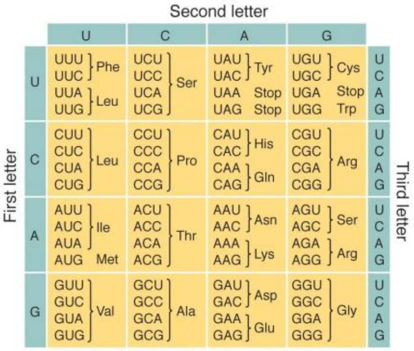
Code words ac, ba, be, db, ed in the permutation space of [a..e]x[a..e]

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

Standard genetic code



First letter	Second letter: U		C		A		G		Third letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Griffiths *et al.*, 2004

Differences from the Standard Code

Vertebrate Mitochondrial Code

AGA	Ter *	Arg R
AGG	Ter *	Arg R
AUA	Met M	Ile I
UGA	Trp W	Ter *

Yeast Mitochondrial Code

AUA	Met M	Ile I
CUU	Thr T	Leu L
CUC	Thr T	Leu L
CUA	Thr T	Leu L
CUG	Thr T	Leu L
UGA	Trp W	Ter *
CGA	absent	Arg R
CGC	absent	Arg R

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
```

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

Number of sequenced genomes:

different organisms (as of 2016)	68,000
human (as of 2014)	228,000