BINF 731

Protein Structure Analysis

http://binf.gmu.edu/vaisman/binf731/
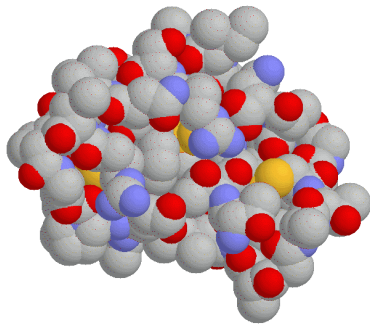
Iosif Vaisman
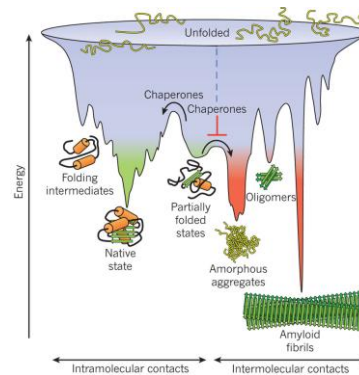
2013

Proteins
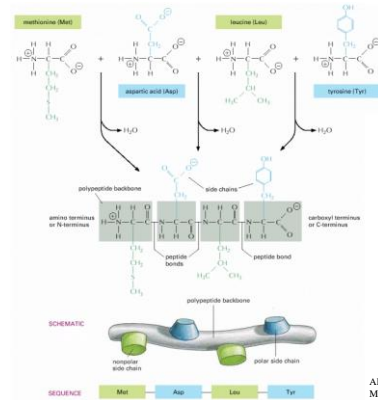
Alberts B, et al.
Molecular Biology of the Cell.

Proteins

TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIIIPGATCPGDYAN

Proteins

Hartl F.U. et al., Nature, 2011

Protein Science

Biochemistry
Biophysics
Molecular Biology

Crystallography
NMR Spectroscopy

Protein Informatics
Structural Bioinformatics
Computational Structural Biology

Proteomics
Structural Genomics

Protein Engineering
Protein Design
Drug Design
Molecular Modeling
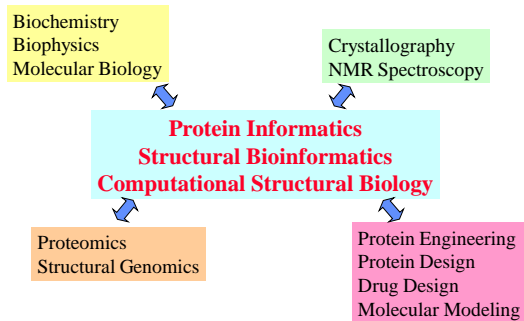
Protein Structure and…

Business
Law
Ethics
Medicine
...

Recommended book

Philip Bourne, Helge Weissig (Eds)
Structural bioinformatics
Hoboken, N.J. : Wiley-Liss, 2003.

Jenny Gu, Philip Bourne (Eds)
Structural bioinformatics
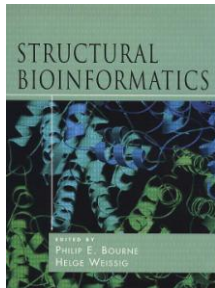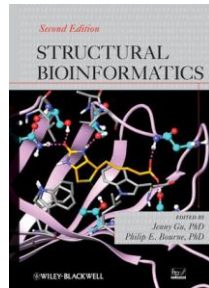Hoboken, N.J. : Wiley-Liss, 2009.

Protein Informatics

SEQUENCE

⬇

STRUCTURE

⬇

FUNCTION

Information

General	Information theory
knowledge or intelligence communicated, received or gained	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

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

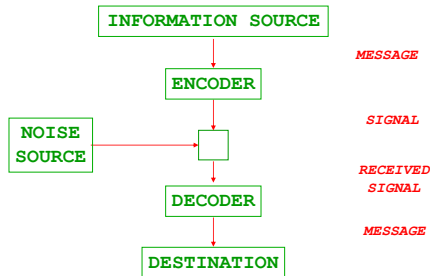only infrmatn esentil to understandn mst b tranmitd

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

ENCODER

MESSAGE

NOISE
SOURCE

SIGNAL

RECEIVED
SIGNAL

DECODER

MESSAGE

DESTINATION

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

Cell Informatics
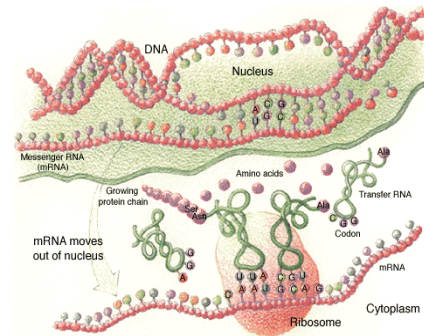
Cell Informatics
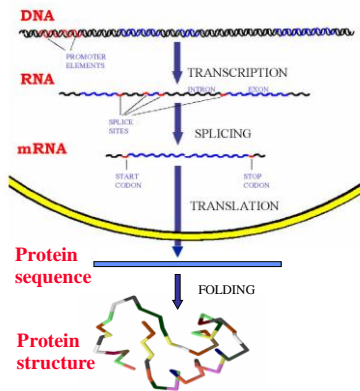
DNA Sequence – Protein Sequence – Protein Structure

ATGAAATTTGGAAACTTCCTTCTCACTTATCAGCCACCT...

MKFGNFLLTYQPPELSQTEVMKRLVN.....
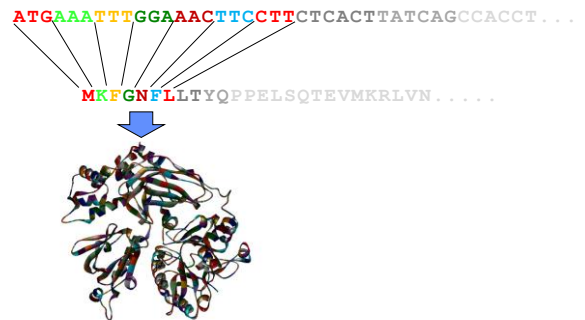
Communication system duality

"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)

Error correcting codes

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

Information Theory

0	1

1 bit

Information Theory

00	01
10	11

1 bit

1 bit

Nucleotide permutation space

0   1

0

1 bit

1

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
```

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