BINF 731

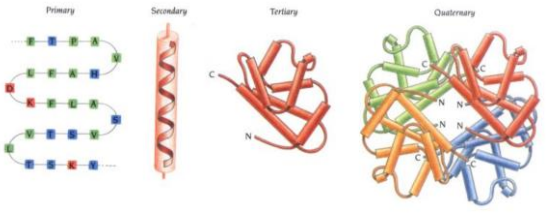
Protein Structure Analysis

http://binf.gmu.edu/vaisman/binf731/

Iosif Vaisman

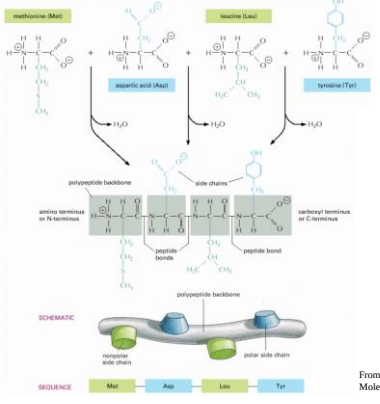
2020

Proteins



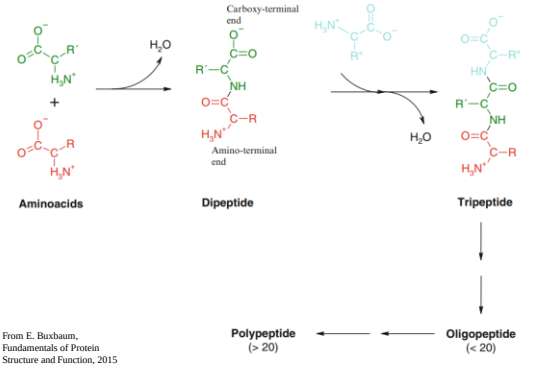
Adopted from Branden and Tooze

Proteins



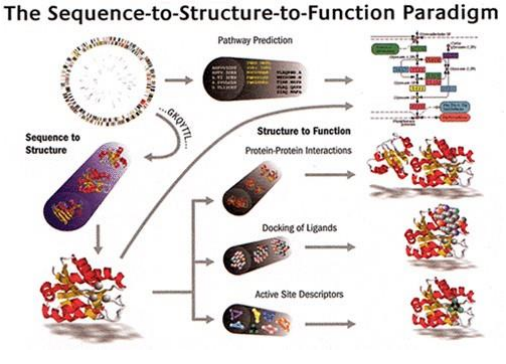
From Alberts B, et al.
Molecular Biology of the Cell.

Proteins



From E. Buxbaum,
Fundamentals of Protein
Structure and Function, 2015

Proteins: Sequence, Structure, Function



J. Skolnick, 2003

Proteins: Sequence, Structure, Function

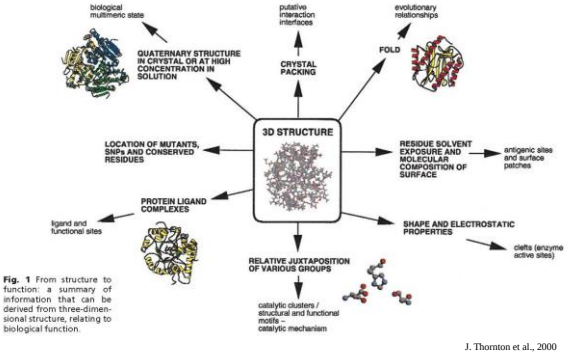
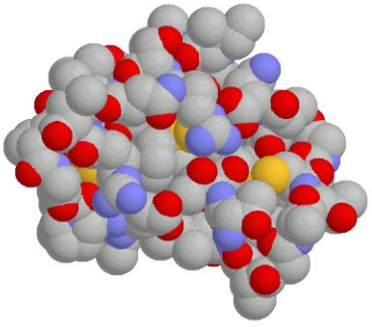


Fig. 1 From structure to function: a summary of information that can be derived from three-dimensional structure, relating to biological function.

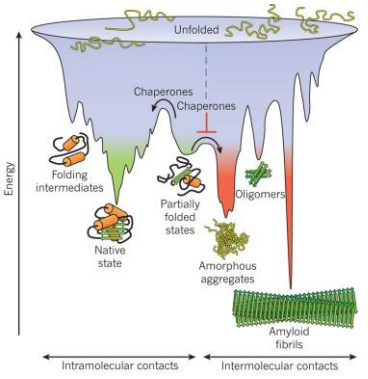
J. Thornton et al., 2000

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

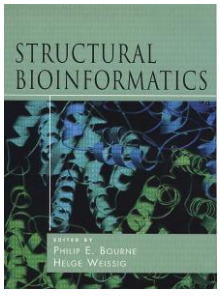
Proteomics
Structural Genomics

Protein Engineering
Protein Design
Drug Design
Molecular Modeling

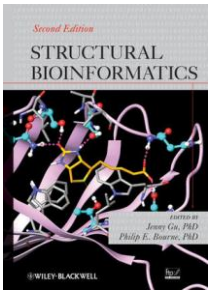
Protein Structure and…

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
↓
DYNAMICS
↓
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

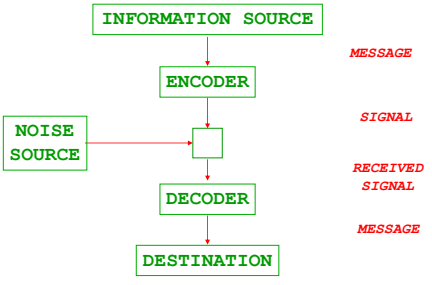
Communication

Fundamental problem of communication:

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

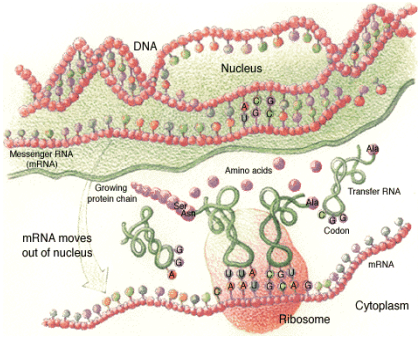
The Mathematical Theory of Communication
Claude Shannon and Warren Weaver

Communication system

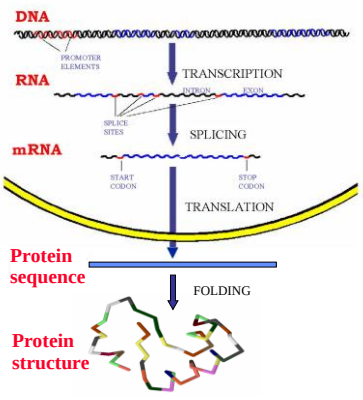


Adopted from C.E. Shannon, ***The Mathematical Theory of Communication***, 1949

Cell Informatics



Cell Informatics



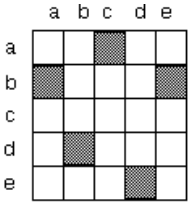
DNA Sequence – Protein Sequence – Protein Structure

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



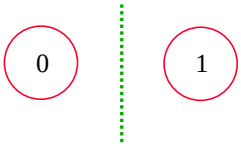
Code words ac, ba, be, db, ed in the permutation space of [a..e]x[a..e]

Hamming metric

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

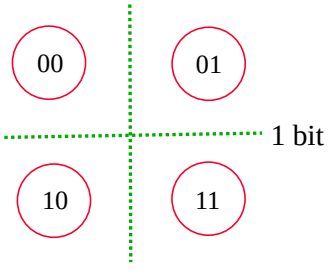
0000 and 0111 are separated by Hamming distance of 3:
0000 - 0001 - 0011 - 0111

Information Theory



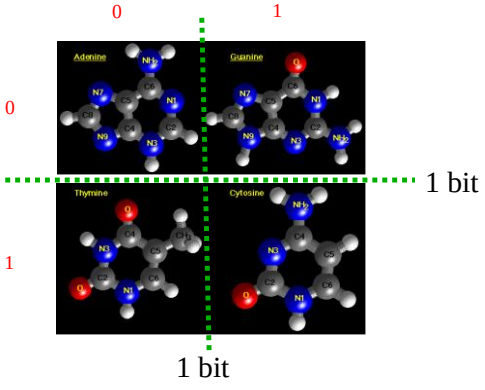
1 bit

Information Theory



1 bit

Nucleotide permutation space



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** (Sec)
TTG	L **Leu**	TCG	S Ser	TAG	* **Ter** (Pyl)	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

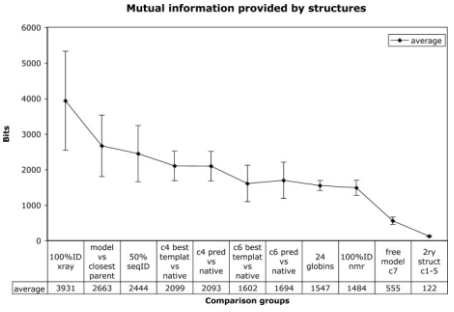
Standard genetic code

```
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
```

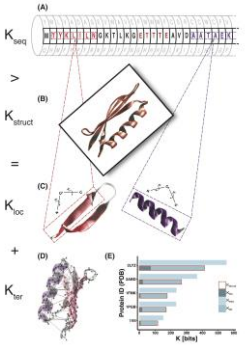

Information theory in protein structure evaluation



	100%ID xray	model vs closest parent	50% seqID	c4 best templat vs native	c4 pred vs native	c6 best templat vs native	c6 pred vs native	24 globins	100%ID nmr	free model c7	2ry struct c1-5
average	3931	2663	2444	2099	2093	1602	1694	1547	1484	555	122

R.Swanson et al., 2008

Information content in protein sequences and structures



A. Possenti et al., 2018

Amino acid naming conventions

- Amino acids with a unique first letter: **C**ys, **H**is, **I**le, **M**et, **S**er, **V**al
- Where several amino acids start with the same letter, common amino acids are given preference: **A**la, **G**ly, **L**eu, **P**ro, **T**hr
- Letters other than the firstst letter are used for Asn (asparagi**N**), Arg (a**R**ginine), Tyr (t**Y**rosine)
- Similar sounding names: Asp (aspar**D**ic acid), Glu (glut**E**mate), Gln (**Q**tamine), Phe (**F**enylalanine)
- The remaining amino acids have letters that do not occur in their name: Lys (**K** close to L), Trp (**W** reminds of double ring), Sec (**U**), Pyl (**O**)
- **X** is used as placeholder, meaning "any amino acid". **B** is used for "Asp or Asn", **Z** for "Gln or Glu", **J** for "Ile or Leu". The **-** is used to denote gaps in a protein sequence, e.g., in sequence alignments. **h** is used to denote hydrophobic amino acids (do not confuse with H for His!)

E.Buxbaum, 2015