BINF 731

Protein Structure Analysis

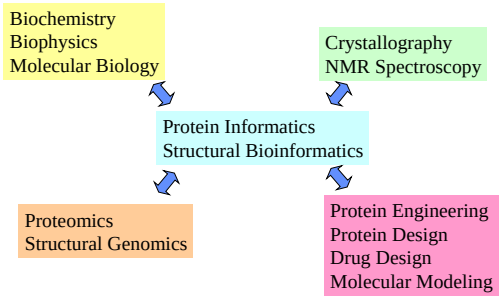
http://binf.gmu.edu/vaisman/binf731/

Iosif Vaisman

2004

COMPUTATIONAL BIOLOGY
COMPUTATIONAL STRUCTURAL BIOLOGY
COMPUTATIONAL MOLECULAR BIOLOGY
BIOINFORMATICS
STRUCTURAL BIOINFORMATICS
GENOMICS
STRUCTURAL GENOMICS
PROTEOMICS
…
…

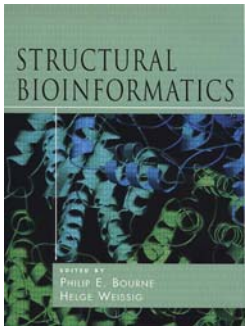
Protein Science (Proteinology?)



Protein Structure and…

Business
Law
Ethics
Medicine
...

Recommended book



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

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.

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	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

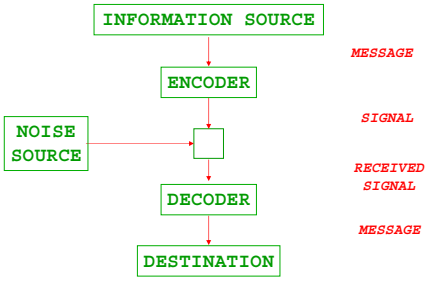
Communication

Fundamental problem of communication:

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

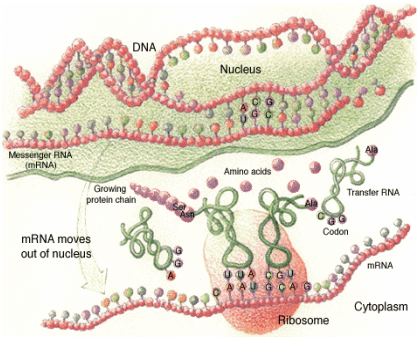
The Mathematical Theory of Communication
Claude Shannon and Warren Weaver

Communication system

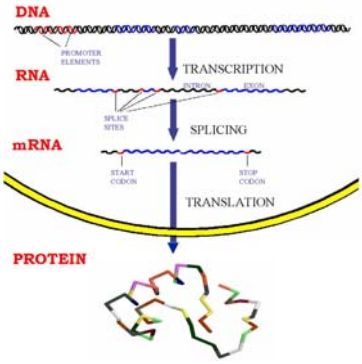


Adopted from C.E. Shannon, ***The Mathematical Theory of Communication***, 1949

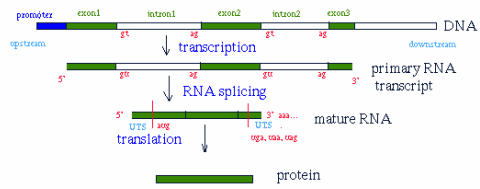
Cell Informatics



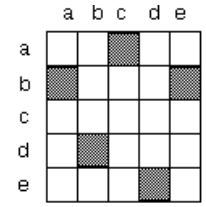
Cell Informatics



Cell Informatics



Error correcting codes



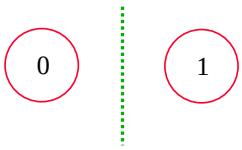
Code words ac, ba, be, db, ed in the permutation space of [a..e]x[a..e]

Hamming metric

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

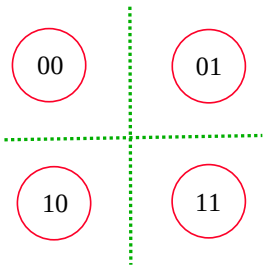
0000 and 0111 are separated by Hamming distance of 3:
0000 - 0001 - 0011 - 0111

Information Theory



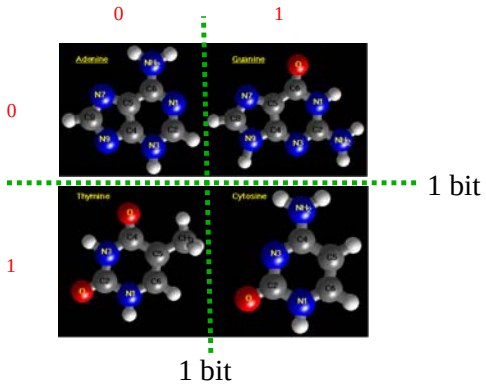
1 bit

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
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

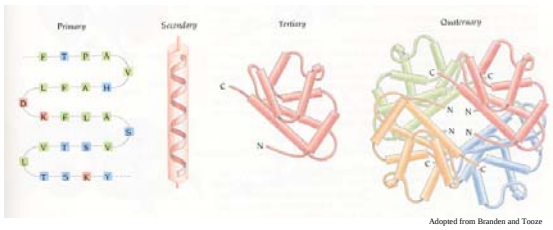
```
  AAs = FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = ---M---------------M---------------M----------------------------
 Base1 = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
 Base2 = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
 Base3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

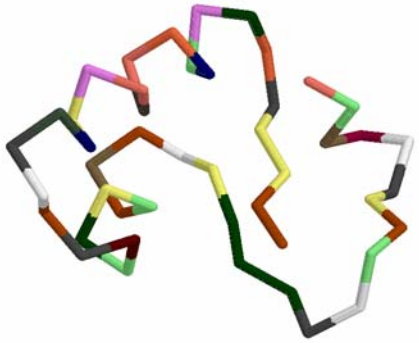
Protein Structure Hierarchy



Adopted from Branden and Tooze

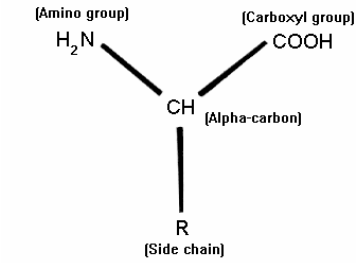
•Primary - the sequence of amino acid residues
•Secondary - ordered regions of primary sequence
(helices, beta-sheets, turns)
•Tertiary - the three-dimensional fold of a protein subunit
•Quaternary - the arrangement of subunits in oligomers.

Protein backbone

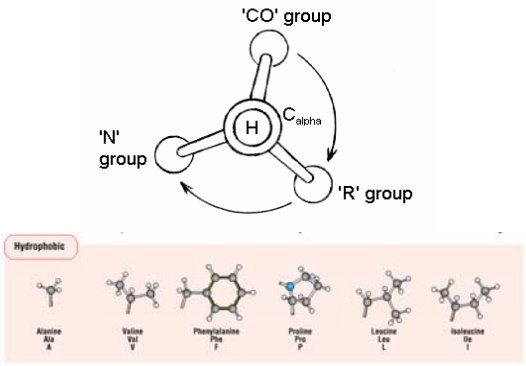


NWVLSTAADMQGVVTDGMASGLDKD

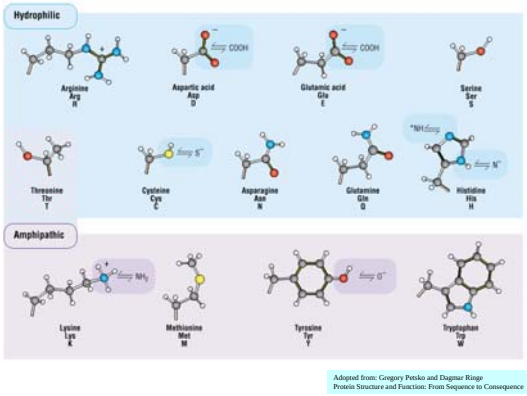
Amino Acid Residue



Amino Acid Residue

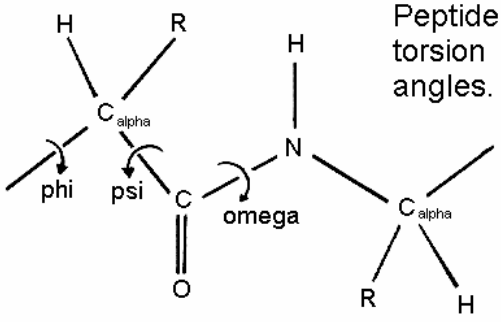


Amino Acids

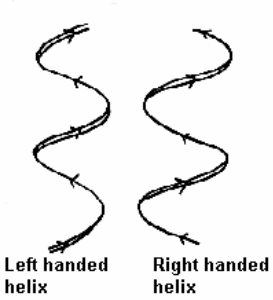


Adopted from: Gregory Petsko and Dagmar Ringe
Protein Structure and Function: From Sequence to Consequence

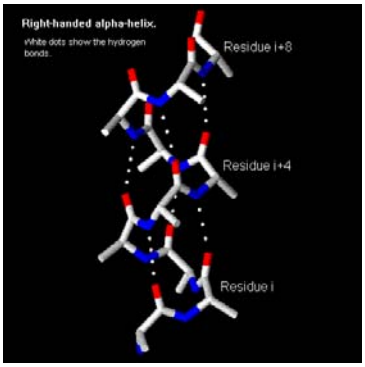
Peptide torsion angles.



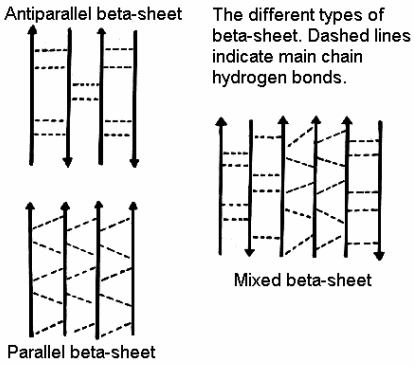
Secondary Structure (Helices)

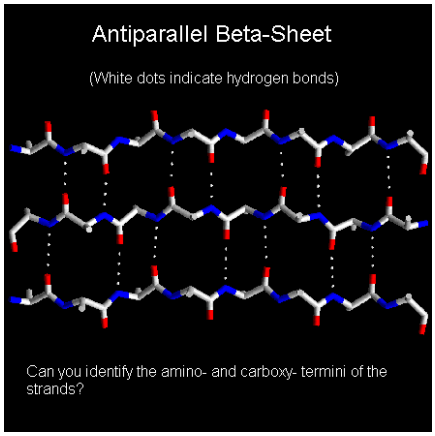


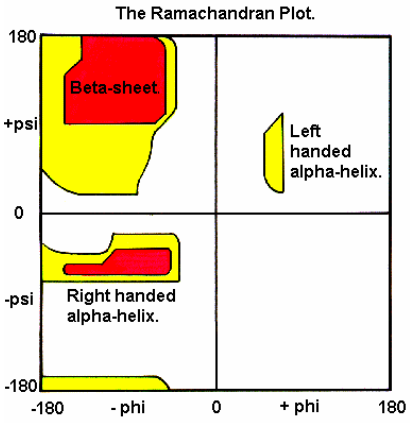
Helix

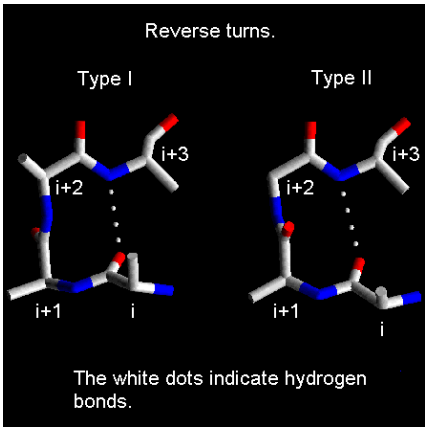


Secondary Structure (Beta-sheets)

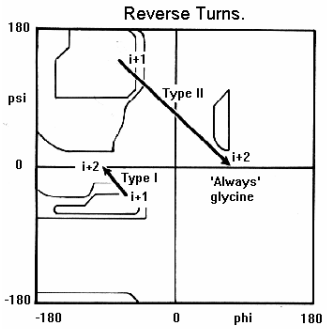


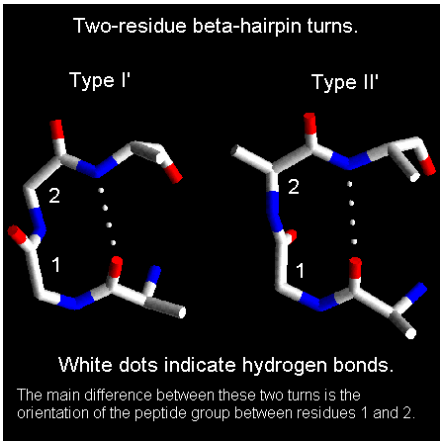




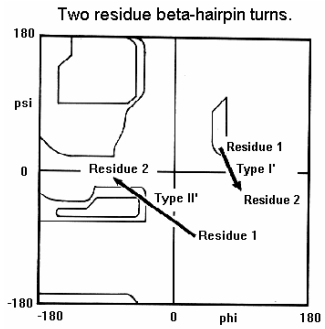


Reverse Turns on a Ramachandran Plot

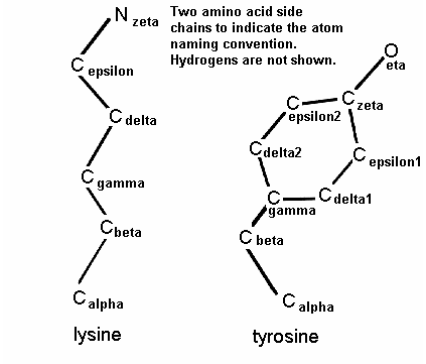




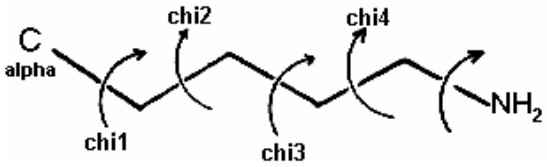
Beta-hairpin Turns on a Ramachandran Plot



Side-Chain Atom Nomenclature



Side-Chain Torsional Angles



Secondary Structure: Computational Problems

Secondary structure characterization
Secondary structure assignment
Secondary structure prediction
Protein structure classification