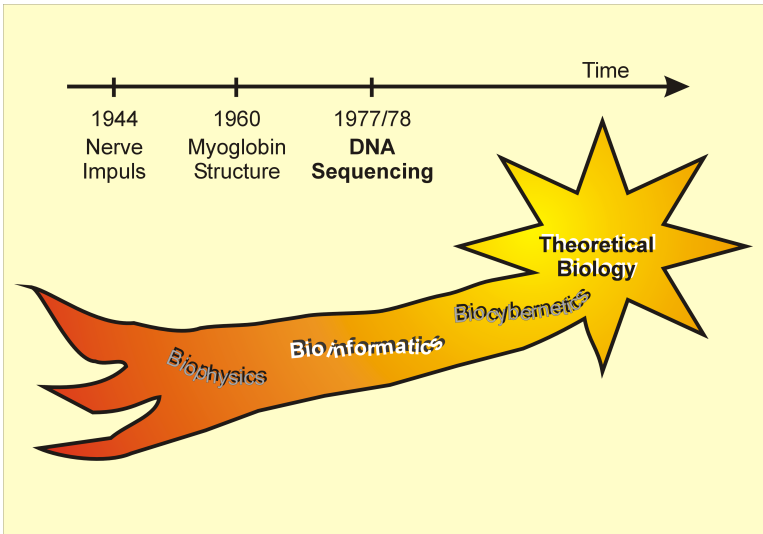
Bioinformatics



1

Bioinformatics

- **Margaret Dayhoff**

- **-Atlas of Protein Sequence and Structure, 1965-1969**
- **-PAM Matrices**
- **(http://pir.georgetown.edu/pirwww/)**

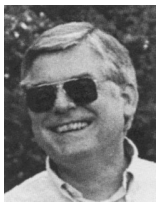
2

Bioinformatics

- **Temple Smith, Mike Waterman**
- **-Global alignment algorithm (1981)**

3

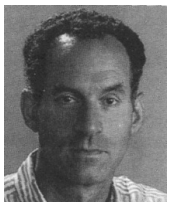
Bioinformatics



- **Bill Pearson**
- **-FASTA algorithm (1988)**

4

Bioinformatics



- **David Lipman**
- **-ENTREZ (http://www.ncbi.nlm.nih.gov/ENTREZ)**

5

Bioinformatics

- **Altschul et al.**
- **-BLAST (basic local alignment search tool) (1990)**

6

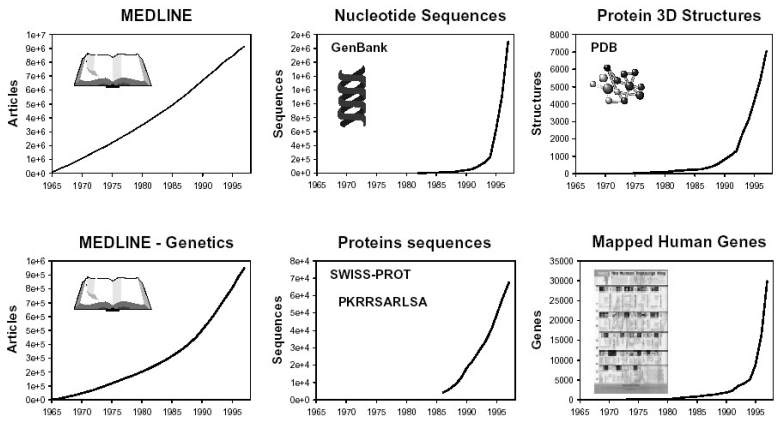
Bioinformatics

Changes in Experimental Life Sciences:

- Increased quantity of biological data
  - Increased number of researchers
  - High-throughput experimental technologies
- Changes in data quality
  - Uniform data
  - Abstract data with large gap to (often unknown) biological function

7

Bioinformatics



Ouellette F, Bethesda, 2000

8

Bioinformatics

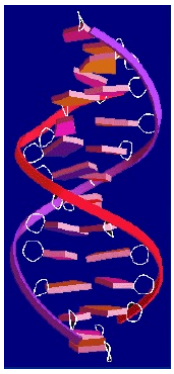
- **G. von Heijne, S. Brunak, G. Cameron, A. Tramontano, G. Vriend (ESF):**
- **„The use of computational techniques to handle, analyze, and add value to the flood of data coming from modern genomics and proteomics“**
- **Theoretical analysis of macromolecular sequences and structures**
- **Analyses and comparisons of genomic data**
- **Modeling protein structures**
- **Organizing biological knowledge in databases and data mining**
- **Target identification for drug development**
- **DNA chip analyses**
- **Analyses of metabolic and functional networks**

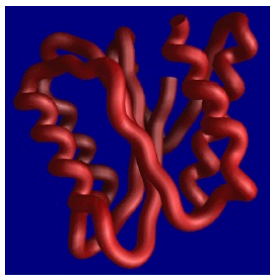
9

Bioinformatics

•Gene

Function



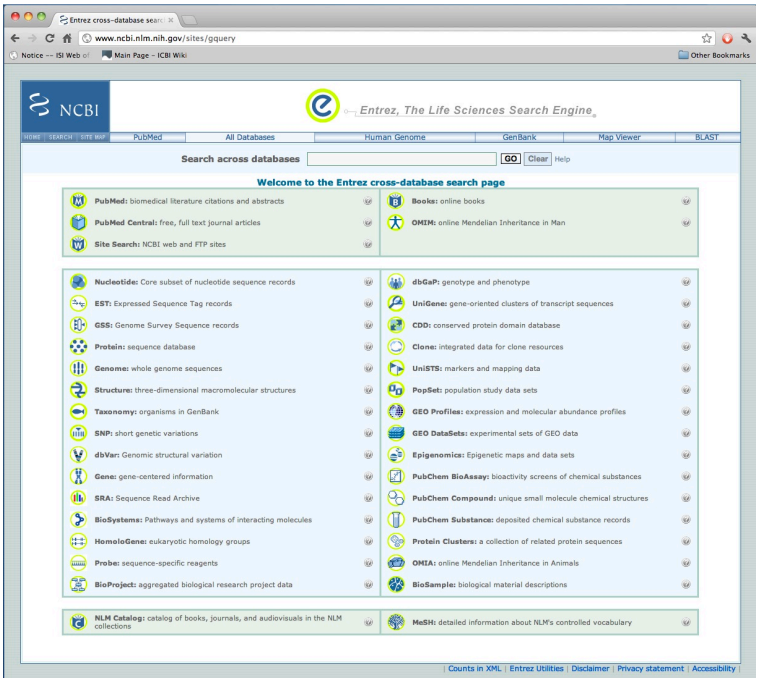


10

Bioinformatics

- Biological Databases: Information Retrieval
- Biological Databases: Defining and Building
- Predictive Methods using DNA and Protein Sequences (Part I & II)

11



12

Data Visualization

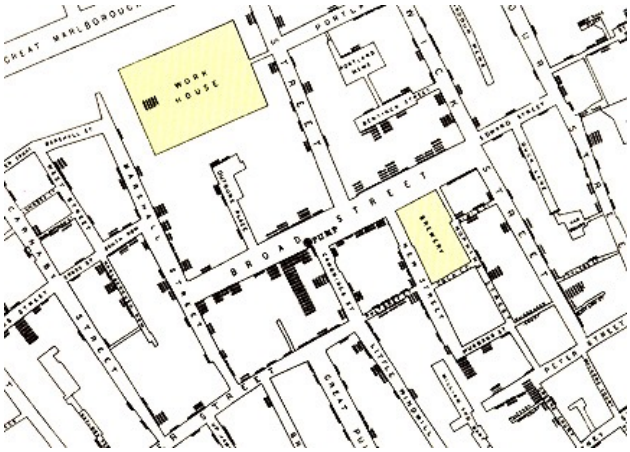
- John Snow (1813-1858)

a legendary figure in the history of public health, epidemiology and anesthesiology.

http://www.ph.ucla.edu/epi/snow.html

The year was 1854; the scene was the Soho District of west London. During the stifling beat of August, there had been a handful of deaths from the dreaded disease cholera. Not unusual, in itself. But on August 31, the situation exploded: In a single evening, within a radius of only blocks, doctors reported 56 new cholera cases. By the next evening there were 143, and the death toll had reached 70 and was climbing. Residents started fleeing the district in panic. Medical authorities debated around the clock but couldn't settle on a plan of action.

13



14

Biological Databases: Defining and Building

15

Databases

- Organized array of information
- Put things in, and being able to get them out again.
- Make discoveries.
- Simplify the information space by specialization.
- Resource for other databases and tools.

16

Database Components

- Definition and description
- Unique key
- Update version
- Links to other databases
- Documentation
- Submission/update/correction process

17

A Bioseq defines an integer coordinate system.

- ASN.1 definition

```
Bioseq ::= SEQUENCE {
id                  SET OF Seq-id ,
descr  Seq-descr                        OPTIONAL,
inst     Seq-inst ,
annot  SET OF Seq-annot    OPTIONAL}
```

- The minimum required elements are an ID and the instance (e.g. length, topology, residues).

18

Primary Data

- DNA/RNA and protein sequences are the primary data for computational biology.
- In most cases protein sequences are interpreted sequences.
- Understanding the various types sequences present in GenBank is key to any interpretation in computational biology.
- Also understand that, as careful as NCBI and others are, errors do creap in, and one needs to always keep that critical eye open.

19

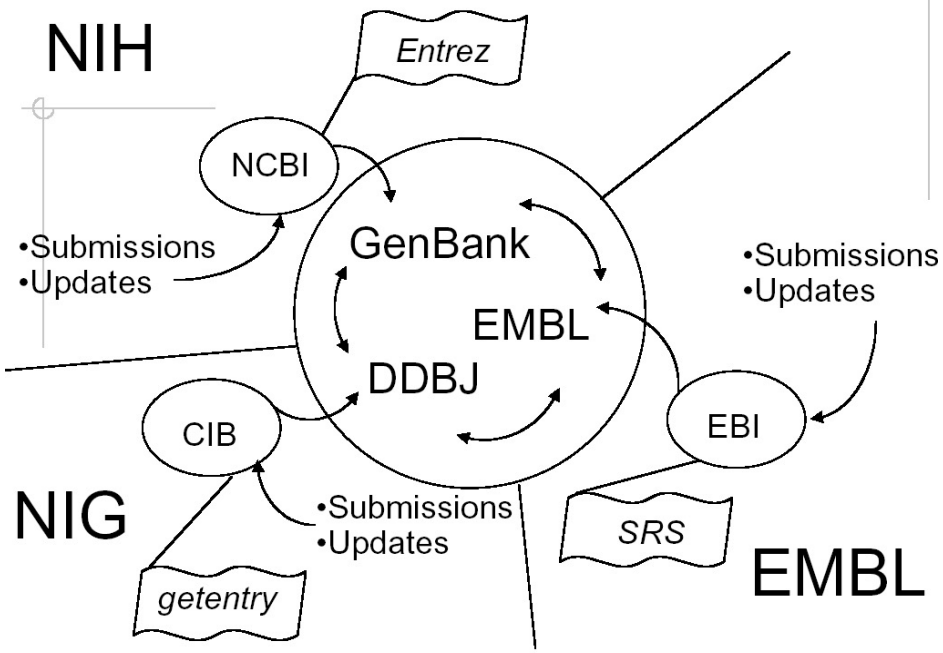
What is GenBank?

- GenBank is the NIH genetic sequence database of all publicly available DNA and derived protein sequences, with annotations describing the biological information these records contain.

http://www.ncbi.nlm.nih.gov/genbank

Nucleic Acids Res. 2011, **39(database issue)**:D32-7

20



21

GenBank - Release 188 - Feb 2012

149,819,246 sequences
137,384,889,783 nucleotides

- Full release of GenBank every 2 months.
- Incremental and cumulative releases: daily.
- GenBank is only available from the Internet.

22

Some insights into using GenBank

- GenBank is a nucleotide-centric view of the information space.
- GenBank is a repository of all publicly available sequences. If it's not in GenBank, it might as well not be considered part of the "public domain".
- In GenBank, records are grouped for various reasons: understand this is key.
- Data in GenBank is only as good as what you put in: applying this is quite important.

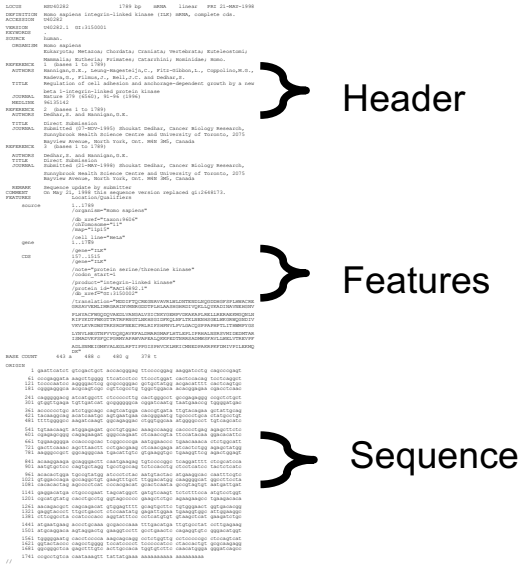
23

Sample GenBank Record

```
LOCUS       HSU40282               1789 bp    mRNA    linear   PRI 21-MAY-1998
DEFINITION  Homo sapiens integrin-linked kinase (ILK) mRNA, complete cds.
ACCESSION   U40282
VERSION     U40282.1  GI:3150001
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1789)
  AUTHORS   Hannigan,G.E., Leung-Hagesteijn,C., Fitz-Gibbon,L., Coppolino,M.G.,
            Radeva,G., Filmus,J., Bell,J.C. and Dedhar,S.
  TITLE     Regulation of cell adhesion and anchorage-dependent growth by a new
            beta 1-integrin-linked protein kinase
  JOURNAL   Nature 379 (6560), 91-96 (1996)
  MEDLINE   96135142
REFERENCE   2  (bases 1 to 1789)
  AUTHORS   Dedhar,S. and Hannigan,G.E.
  TITLE     Direct Submission
  JOURNAL   Submitted (07-NOV-1995) Shoukat Dedhar, Cancer Biology Research,
            Sunnybrook Health Science Centre and University of Toronto, 2075
            Bayview Avenue, North York, Ont. M4N 3M5, Canada
REFERENCE   3  (bases 1 to 1789)
  AUTHORS   Dedhar,S. and Hannigan,G.E.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-MAY-1998) Shoukat Dedhar, Cancer Biology Research,
            Sunnybrook Health Science Centre and University of Toronto, 2075
            Bayview Avenue, North York, Ont. M4N 3M5, Canada
  REMARK    Sequence update by submitter
COMMENT     On May 21, 1998 this sequence version replaced gi:2648173.
```

24

GenBank Flat File (GBFF)



25

Sample GenBank Record

```
LOCUS       HSU40282                1789 bp    mRNA    linear   PRI 21-MAY-1998
DEFINITION  Homo sapiens integrin-linked kinase (ILK) mRNA, complete cds.
ACCESSION   U40282
VERSION     U40282.1  GI:3150001
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1789)
  AUTHORS   Hannigan,G.E., Leung-Hagesteijn,C., Fitz-Gibbon,L., Coppolino,M.G.,
            Radeva,G., Filmus,J., Bell,J.C. and Dedhar,S.
  TITLE     Regulation of cell adhesion and anchorage-dependent growth by a new
            beta 1-integrin-linked protein kinase
  JOURNAL   Nature 379 (6560), 91-96 (1996)
  MEDLINE   96135142
REFERENCE   2  (bases 1 to 1789)
  AUTHORS   Dedhar,S. and Hannigan,G.E.
  TITLE     Direct Submission
  JOURNAL   Submitted (07-NOV-1995) Shoukat Dedhar, Cancer Biology Research,
            Sunnybrook Health Science Centre and University of Toronto, 2075
            Bayview Avenue, North York, Ont. M4N 3M5, Canada
```

26

Sample GenBank Record

```
FEATURES             Location/Qualifiers
     source          1..1789
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="11"
                     /map="11p15"
                     /cell_line="HeLa"
     gene            1..1789
                     /gene="ILK"
     CDS             157..1515
                     /gene="ILK"
                     /note="protein serine/threonine kinase"
                     /codon_start=1
                     /product="integrin-linked kinase"
                     /protein_id="AAC16892.1"
                     /db_xref="GI:3150002"
                     /translation="MDDIFTQCREGNAVAVRLWLDNTENDLNQGDDHGFSPLHWACRE
                     GRSAVVEMLIMRGARINVMNRGDDTPLHLAASHGHRDIVQKLLQYKADINAVNEHGNV
                     PLHYACFWGQDQVAEDLVANGALVSICNKYGEMPVDKAKAPLRELLRERAEKMGQNLN
                     RIPYKDTFWKGTTRTRPRNGTLNKHSGIDFKQLNFLTKLNENHSGELWKGRWQGNDIV
                     VKVLKVRDWSTRKSRDFNEECPRLRIFSHPNVLPVLGACQSPPAPHPTLITHWMPYGS
                     LYNVLHEGTNFVVDQSQAVKFALDMARGMAFLHTLEPLIPRHALNSRSVMIDEDMTAR
                     ISMADVKFSFQCPGRMYAPAWVAPEALQKKPEDTNRRSADMWSFAVLLWELVTREVPF
                     ADLSNMEIGMKVALEGLRPTIPPGISPHVCKLMKICMNEDPAKRPKFDMIVPILEKMQ
                     DK"
BASE COUNT      443 a    488 c    480 g    378 t
ORIGIN
        1 gaattcatct gtcgactgct accacgggag ttccccggag aaggatcctg cagcccgagt
       61 cccgaggata aagcttgg
              ...
     1741 ccgcctgtca caataaagtt tattatgaaa aaaaaaaaaa aaaaaaaaa
```

27

LOCUS, Accession, Accession.version & gi

```
LOCUS       HSU40282     1786 bp    mRNA            PRI       28-NOV-1997
DEFINITION  Homo sapiens integrin-linked kinase (ILK) mRNA, complete cds.
ACCESSION   U40282
  VERSION:  U40282.1  GI: 3150001
```

LOCUS: HSU40282
ACCESSION: U40282
Nucleotide gi: 3150001
VERSION: U40282.1 GI: 3150001
Protein gi: 3150002
protein_id: AAC16892.1

```
     CDS             157..1515
                     /gene="ILK"
                     /note="protein serine/threonine kinase"
                     /codon_start=1
                     /product="integrin-linked kinase"
                     /db_xref="GI:3150002"
                     /protein_id="AAC16892.1"
```

28

GenBank Organismal divisions:

- **PRI** - Primate
- **ROD** - Rodent
- **MAM** - Mammalian
- **VRT** - Vertebrate
- **INV** - Invertebrate
- **PLN** - Plant
- **BCT** - Bacterial
- **VRL** - Viral
- **PHG** - Phage
- **SYN** - Synthetic
- **UNA** - Unannotated

29

Predictive Methods using DNA and Protein Sequences

Part I

30

BLAST

- Seeks high-scoring segment pairs (HSP)
  - pair of sequences that can be aligned without gaps
  - when aligned, have maximal aggregate score (score cannot be improved by extension or trimming)
  - score must be above score threshhold *S*
  - gapped (2.0) or ungapped (1.4)
- Search engines
  - WWW search form
    http://www.ncbi.nlm.nih.gov/BLAST
  - Unix command line
    blastall -p progname -d -db -i query > outfile
  - E-mail server E-mail server
    blast@ncbi.nlm.nih.gov

31

BLAST Algorithms

Program	*Query Sequence*	*Target Sequence*
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide, six-frame translation	Protein
TBLASTN	Protein	Nucleotide, six-frame translation
TBLASTX	Nucleotide, six-frame translation	Nucleotide, six-frame translation

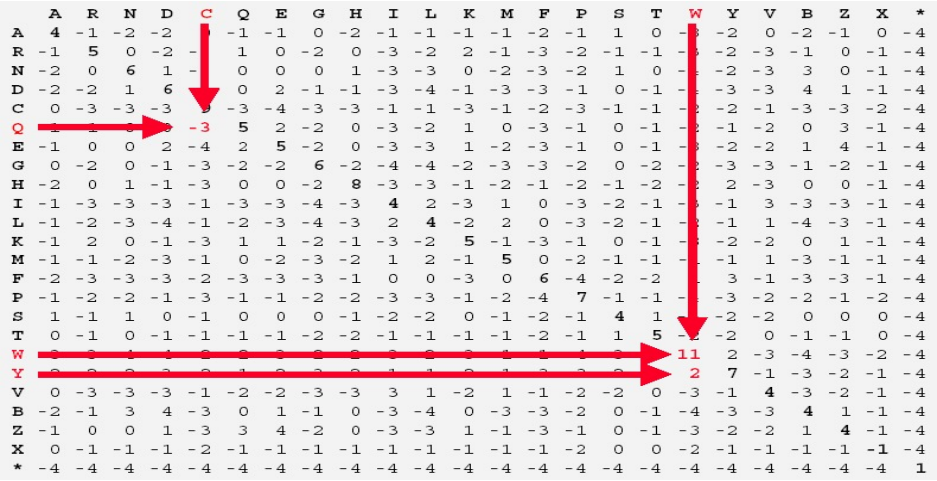
32

Scoring Matrices

- Empirical weighting scheme to represent biology
  - Cys/Pro important for structure and function
  - Trp has bulky side chain
  - Lys/Arg have positively-charged side chains
- Importance of understanding scoring matrices
  - Appear in all analyses involving sequence comparison
  - Implicitly represent a particular theory of evolution
  - Choice of matrix can strongly influence outcomes

33

Matrix Structure



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

34

PAM Matrices

- Margaret Dayhoff, 1978
- Point Accepted Mutation (PAM)
  - Look at patterns of substitutions in related proteins
  - The new side chain must function the same way as the old one ("acceptance")
  - On average, 1 PAM corresponds to 1 amino acid change per 100 residues
  - 1 PAM ~ 1% divergence
  - Extrapolate to predict patterns at longer distances

35

PAM Matrices

- Assumptions
  - Replacement is independent of surrounding residues
  - Sequences being compared are of average composition
  - All sites are equally mutable
- Sources of error
  - Small, globular proteins used to derive matrices (departure from average composition)
  - Errors in PAM 1 are magnified up to PAM 250
  - Does not account for conserved blocks or motifs

36

BLOSUM Matrices

- Henikoff and Henikoff, 1992
- Blocks Substitution Matrix (BLOSUM)
  - Look only for differences in conserved, ungapped regions of a protein family
  - More sensitive to structural or functional substitutions
  - BLOSUM *n*
    - Contribution of sequences > *n*% identical weighted to 1
    - Substitution frequencies are more heavily-influenced by sequences that are more divergent than this cutoff
    - Clustering reduces contribution of closely-related sequences
    - Reducing *n* yields more distantly-related sequences

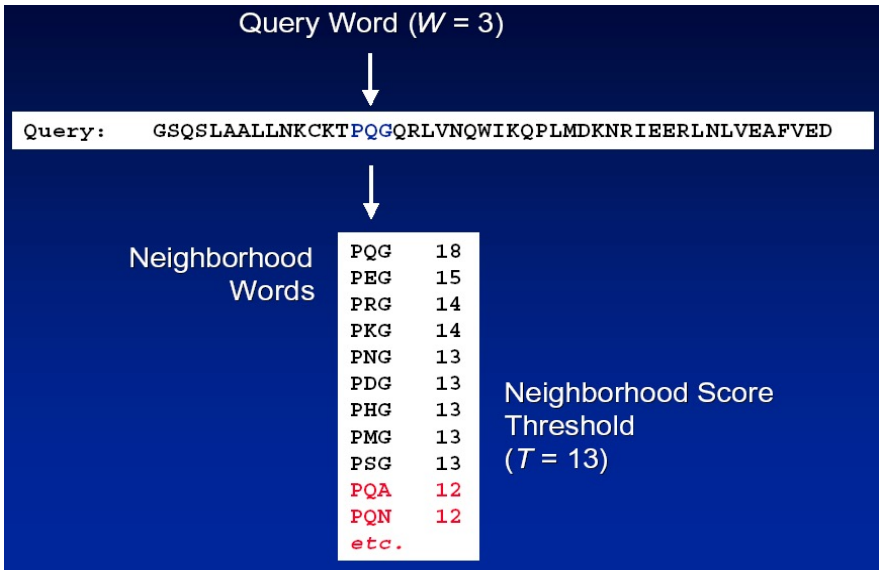
37

So many matrices...

- Triple-PAM strategy *(Altschul, 1991)*
  - PAM 40 Short alignments, highly similar
  - PAM 120
  - PAM 250 Longer, weaker local alignments
- BLOSUM *(Henikoff, 1993)*
  - BLOSUM 90 Short alignments, highly similar
  - BLOSUM 62 Most effective in detecting known members of a protein family
  - BLOSUM 30 Longer, weaker local alignments
- No single matrix is the complete answer for all sequence comparisons

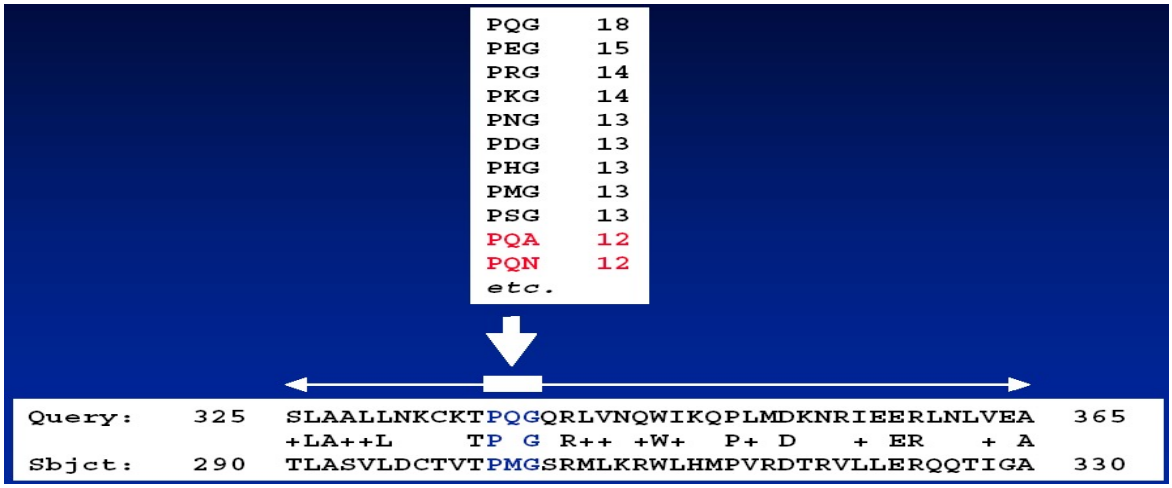
38

Neighborhood Words



39

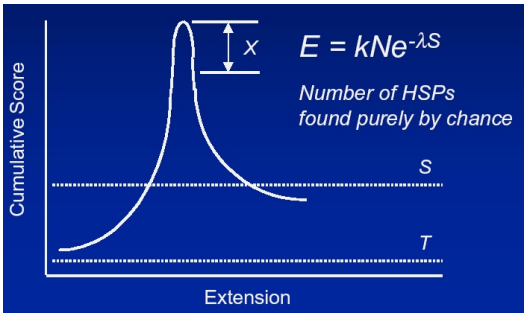
High-Scoring Segment Paris



40

BLAST Search Requirements

- A query sequence, in FASTA format
- Which BLAST program to use
- Which database to search
- Parameter values



41

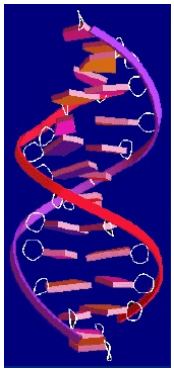
Predictive Methods using DNA and Protein Sequences

Part II

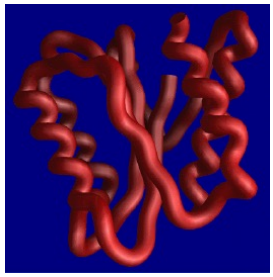
42

The Flow of Biotechnology Information

Gene



Function



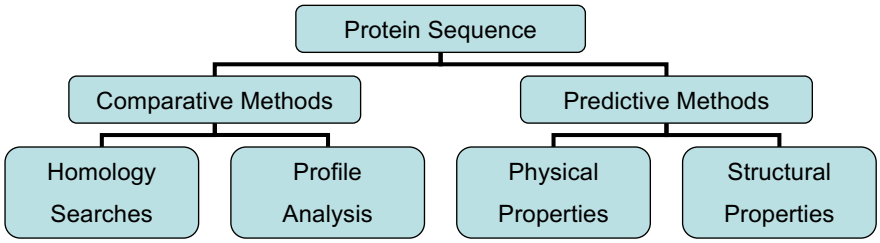
43

Protein Conformation

- Christian Anfinsen
  Studies on reversible denaturation
  "Sequence specifies conformation"
- Chaperones and disulfide interchange enzymes:
  involved but not controlling final state
- "Starting with a newly-determined sequence, what can be determined computationally about its possible function and structure?"

44

Protein Sequence Analysis



- *Shared ancestry?*
- *Similar function?*
- *Domain or complete sequence?*

45

BLAST Algorithms

Program	*Query Sequence*	*Target Sequence*
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide, six-frame translation	Protein
TBLASTN	Protein	Nucleotide, six-frame translation
TBLASTX	Nucleotide, six-frame translation	Nucleotide, six-frame translation

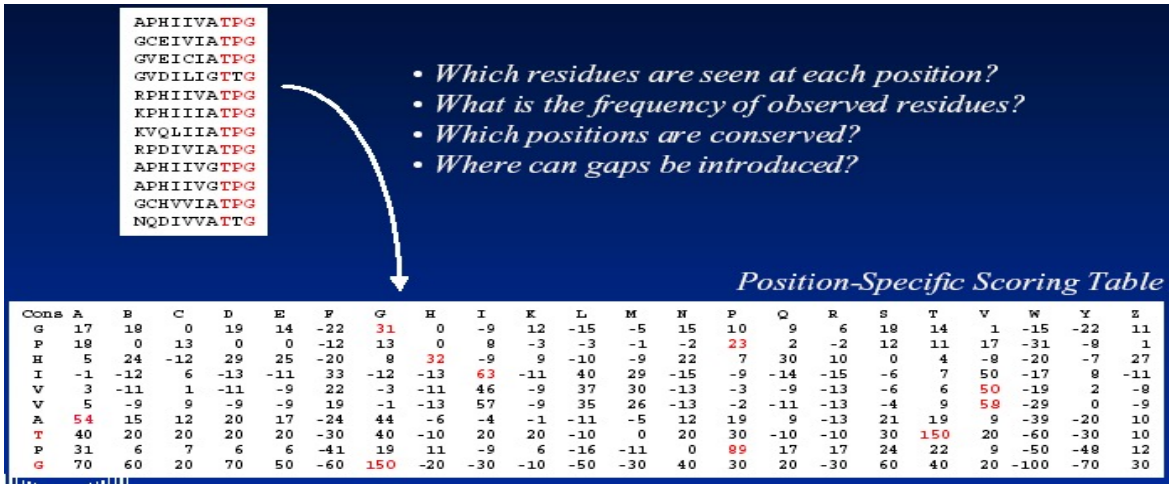
46

Profiles

- Numerical representations of multiple sequence alignments
- Depend upon *patterns* or *motifs* containing conserved residues
- Represent the common characteristics of a protein family
- Can find similarities between sequences with little or no sequence identity
- Allow for the analysis of distantly-related proteins

47

Profile Construction



Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
G	17	18	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22	11
P	18	0	13	0	0	-12	13	0	8	-3	-3	-1	-2	23	2	-2	12	11	17	-31	-8	1
H	5	24	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7	27
I	-1	-12	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-6	7	50	-17	8	-11
V	3	-11	1	-11	-9	22	-3	-11	46	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2	-8
V	5	-9	9	-9	-9	19	-1	-13	57	-9	35	26	-13	-2	-11	-13	-4	9	58	-29	0	-9
A	54	15	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20	10
T	40	20	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30	10
P	31	6	7	6	6	-41	19	11	-9	6	-16	-11	0	89	17	17	24	22	9	-50	-48	12
G	70	60	20	70	50	-60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30

48

Profile Scan

- Search sequence against a collection of profiles
- Databases available
  - PROSITE (prosite.expasy.org) 1634 entries
    - ScanProsite
  - Pfam (pfam.sanger.ac.uk) 13672 families

49

PSI-BLAST

- Position-Specific Iterated BLAST search
- Easy-to-use version of a profile-based search
  - Perform BLAST search against protein database
  - Use results to calculate a position-specific scoring matrix
  - PSSM replaces query for next round of searches
  - May be iterated until no new significant alignments are found
    - Convergence – all related sequences deemed found
    - Divergence – query is too broad, make cutoffs more stringent

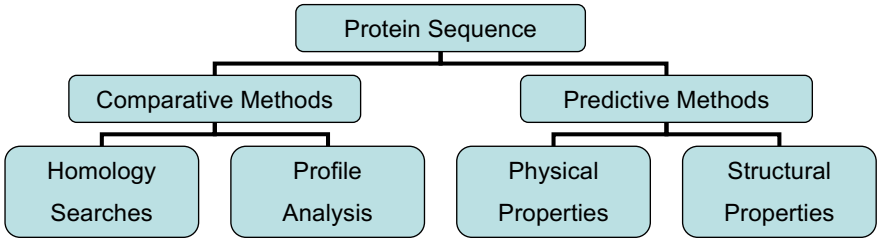
50

BLOCKS

- Steve Henikoff, Fred Hutchinson Cancer Research Center, Seattle
- Multiple alignments of conserved regions in protein families
  - 1 “block” = 1 short, ungapped multiple alignment
  - Families can be defined by one or more blocks
  - Searches allow detection of one or more blocks representing a family
- Search engines
  - Web http://blocks.fhcrc.org/

51

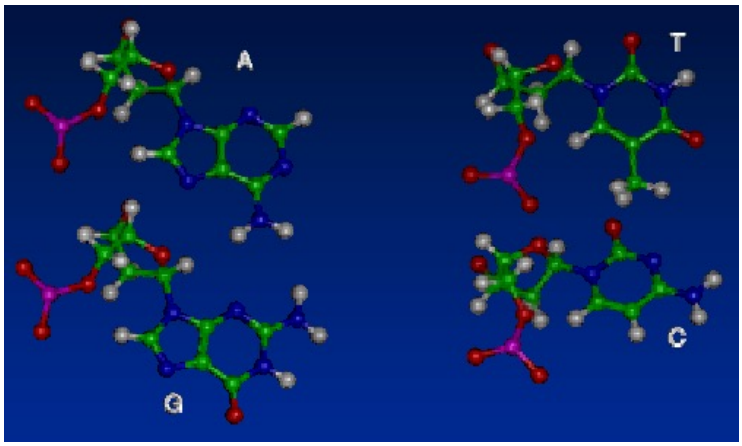
Protein Sequence Analysis



- *Composition*
- *Hydrophobicity*
- *Secondary structure*
- *Specialized structures*
- *Tertiary structure*

52

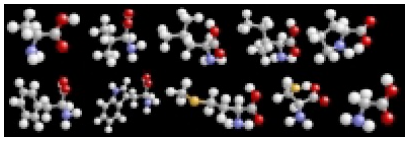
Information Landscape



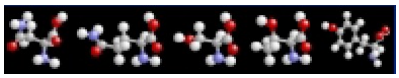
53

Information Landscape

Nonpolar



Polar *Neutral*



Polar *Basic*

Polar *Acidic*

54

ProtParam

- Computes physicochemical parameters
  - Molecular weight
  - Theoretical pI
  - Amino acid composition
  - Extinction coefficient
- Simple query
  - SWISS-PROT accession number
  - User-entered sequence, in single-letter format
- *http://web.expasy.org/protparam*

55

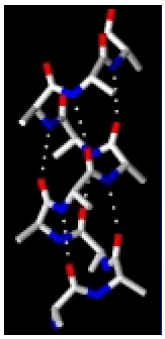
Secondary Structure Prediction

- Deduce the most likely position of alpha-helices and beta-strands
- Confirm structural or functional relationships when sequence similarity is weak
- Determine guidelines for rational selection of specific mutants for further laboratory study
- Basis for further structure-based studies

56

Alpha-helix

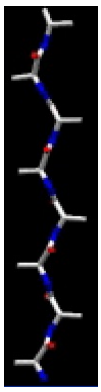
- Corkscrew
- Main chain forms backbone, side chains project out
- Hydrogen bonds between CO group at *n* and NH group at *n+4*
- Helix-formers:
- Ala, Glu, Leu, Met ,
- Helix-breaker: Pro



57

Beta-strand

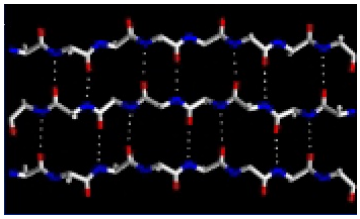
- Extended structure ("pleated")
- Peptide bonds point in opposite directions
- Side chains point in opposite directions
- No hydrogen bonding *within* strand



58

Beta-sheet

- Stabilization through hydrogen bonding
- Parallel or antiparallel
- Variant: beta-turn



59

nnpredict

- Neural network approach to making predictions *(Kneller et al., 1990)*
- Best-case accuracy > 65%
- Search engines
  - Web http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html

60

nnpredict Query

```
option: a/b
>flavodoxin - Anacystis nidulans
AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVGELQSDWEGIY
DDLDSVNFQGKKVAYFGAGDQVGYSDNFQDAMGILEEKISSLGSQTVGYWPIEGYDFNESKAVRNNQFVG
LAIDEDNQPDLTKNRIKTWVSQLKSEFGL
```

α/β folding class

```
Tertiary structure class: alpha/beta

Sequence:
AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVG
ELQSDWEGIYDDLDSVNFQGKKVAYFGAGDQVGYSDNFQDAMGILEEKISSLGSQTVGYW
PIEGYDFNESKAVRNNQFVGLAIDEDNQPDLTKNRIKTWVSQLKSEFGL

Secondary structure prediction (H = helix, E = strand, - = no prediction):
----EEE------EEEHHHHHHH------EEEH--------------EEEE--------
---------------HHHH---EEEE-----------H--HHHHHHHH------E--E-
-E--------------HH--E----------------EHHHHH------
```

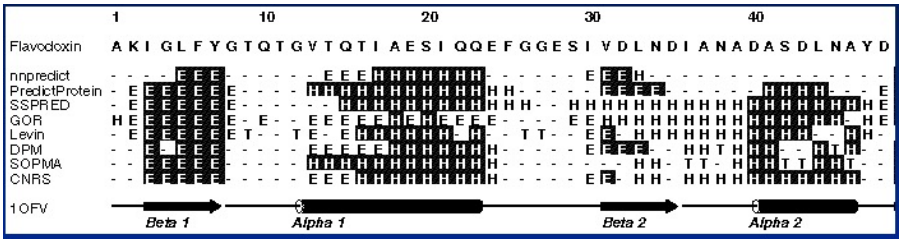
61

PredictProtein

- Multi-step predictive algorithm *(Rost et al., 1994)*
  - Protein sequence queried against SWISS-PROT
  - MaxHom used to generate iterative, profile-based multiple sequence alignment *(Sander and Schneider, 1991)*
  - Multiple alignment fed into neural network (PHDsec)
- Accuracy
  - Average > 70%
  - Best-case > 90%
- Search engines
  - Web http://www.predictprotein.org/

62

Accuracy of Predictions



63

SignalP

- Neural network trained based on phylogeny
  - Gram-negative prokaryotic
  - Gram-positive prokaryotic
  - Eukaryotic
- Predicts secretory signal peptides
  (*not* those involved in intracellular signal transduction)
- http://www.cbs.dtu.dk/services/SignalP/

64

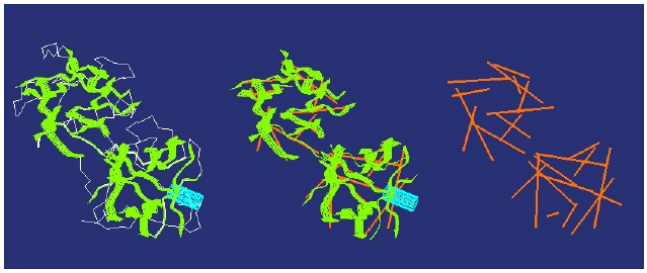
Predicting Tertiary Structure

- Sequence specifies conformation, *but* conformation does *not* specify sequence
- Structure is conserved to a much greater extent than sequence
  - Limited number of protein folds
- Similarities between proteins may not necessarily be detected through "traditional" methods

65

VAST Structure Comparison

Step 1: Construct vectors for secondary structure elements

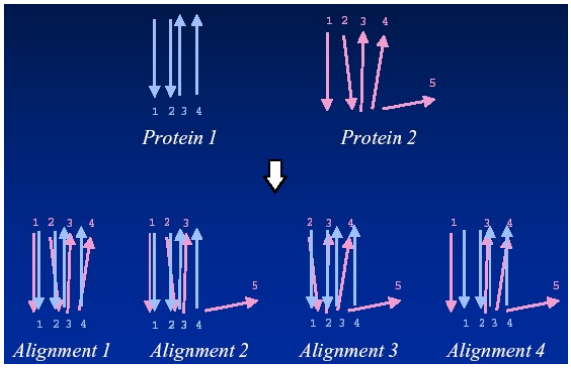


Ricin Chain B

66

VAST Structure Comparison

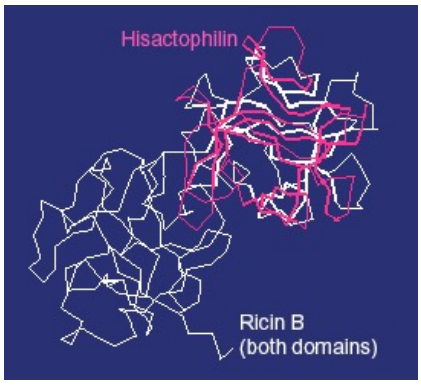
Step 2: Optimally align structure element vectors



67

VAST Structure Comparison

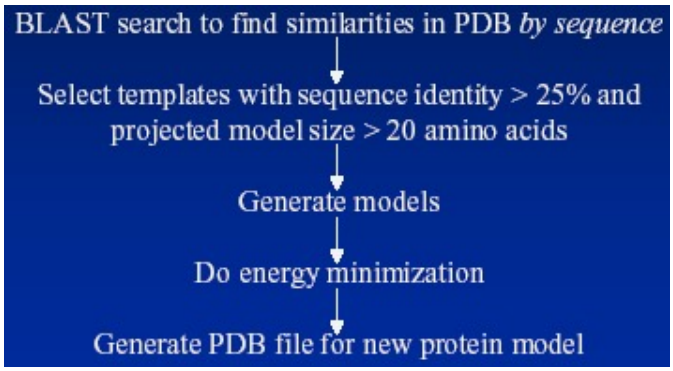
Step 3: Refine residue-by-residue alignment using Monte Carlo



68

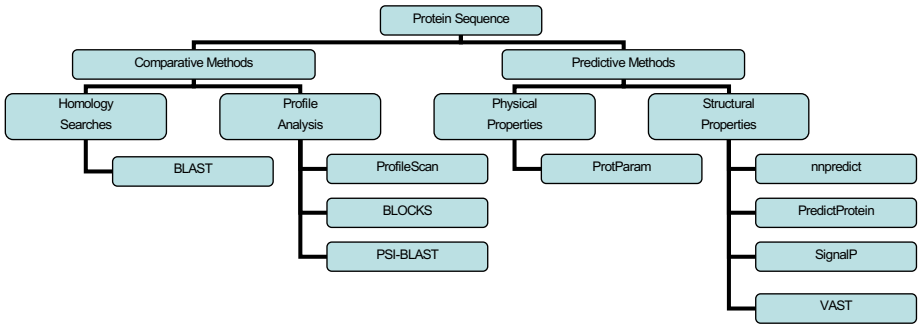
SWISS-MODEL

- Automated comparative protein modelling server
- http://swissmodel.expasy.ch/



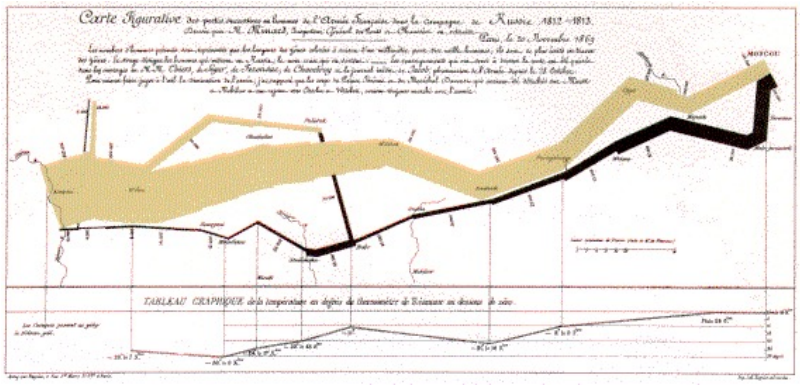
69

Protein Sequence Analysis



70

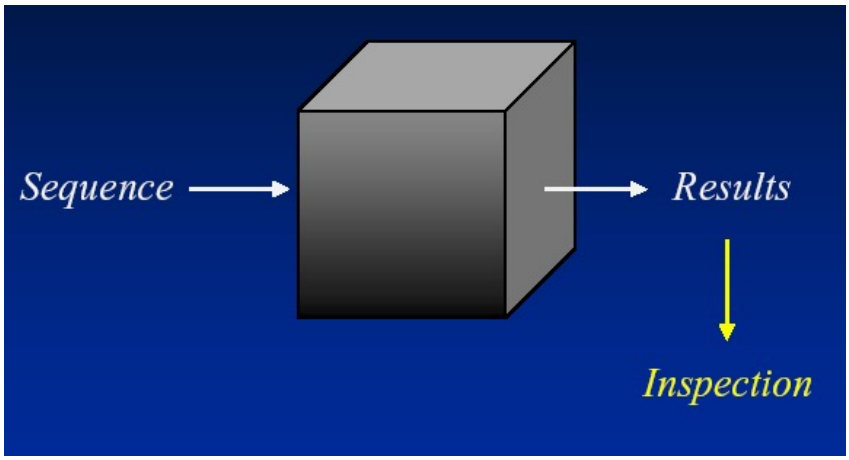
Napoleon's Invasion of Russia, 1812
Chart by Minard



http://uts.cc.utexas.edu/~jrubarth/gslis/lis385t.16/Napoleon/

71

Understanding Analysis



72

Some lessons learned by bioinformaticians –
sometimes, the hard way

73

"Short Motif Pitfall"

- The level of sequence identity required for significant homology is much higher for smaller regions
- Two proteins may share a common domain while still being dissimilar elsewhere
- For very short motifs, homology *cannot* be inferred by sequence identity

short motifs may not be helpful in describing what a protein does

➡

74

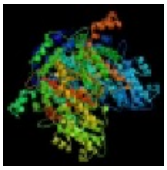
Immunoglobulin Signature

- Signature defined: [FY]-x-C-x-[VA]-x-H
- Precision
  - Total: 456 hits in 412 sequences
  - True positives: 385 hits in 341 sequences
  - False positives: 71 hits in 71 sequences

Acyl-CoA dehydrogenase	Aminoadipate-semialdehyde dehydrogenase
Acyl-amino acid-releasing enzyme	DNA replication licensing factor
Alpha-adaptin A	Neprin A
GDP-mannose 6-dehydrogenase	Cytochrome C-522
Membrane alanyl aminopeptidase	Phosphatidylinositol 3-kinase
Phosphatidyl cytidylyl transferase	Origin recognition complex subunit 2
D-lactate dehydrogenase	Para-aminobenzoate synthase
DNA polymerase B	Alpha-platelet-derived growth factor
Hemerythyrin	Serine-threonine protein kinase
Anterior-restricted homeobox protein	Photosystem II 44 kDa reaction center protein
Mast-stem cell growth factor	DNA-directed RNA polymerase II (subunits)
Limulus clotting factor C	Chloroplast 30S ribosomal protein S4
Arachidonate 12-lipoxygenase	Titin

75

100% identity, but...



- Phosphoglucose isomerase
  catalyzes interconversion of D-glucose-6-phosphate and D-fructose-6-phosphate
- Neuroleukin
  secreted by T-cells, promotes survival of some embryonic spinal neurons and sensory nerves; B-cell maturation
- Autocrine motility factor
  tumor cell product that stimulates cancer cell migration (metastasis?)
- Differentiation and maturation mediator
  In vitro *differentiation of human myeloid leukemia HL-60 cells to terminal monocytes*

76

Proteins with Multiple Functions

Thymidine phosphorylase	Endothelial cell growth factor
Thymidylate synthase	Translation inhibitor
birA biotin synthase	bir operon repressor
Cystic fibrosis transmembrane conductance regulator (CFTR)	Regulates other ion channels
Crystallin	Enolase Lactate dehydrogenase Heat shock protein

77

Does sequence similarity imply common function?

Maybe.

78

Structural Superfamilies: TIM Barrel

- Minimum 200 residues required for structure, with 160 residues structurally equivalent
- Structures mediate a wide variety of chemical reactions critical to biological survival
- May account for up to
  - 10% of all soluble enzymes
  - 10% of all proteins

Triose phosphate isomerase
Ribulose-phosphates
Thiamin phosphate synthase
FMN-linked oxidoreductases
NAD(P)-linked oxidoreductase
Glycosyltransferases
Metallo-dependent hydrolases
Aldolase
Enolase
Phosphenol pyruvate
Malate synthase G
RuBisCo
Xylose isomerase-like proteins
Bacterial luciferase-like proteins
Quinolinic acid phosphoribosyltransferases
Cobalamin (B12)-dependent enzymes
tRNA-guanine transglycosylase
Dihydropteroate synthetase
Uroporphyrinogen decarboxylase
Methylenetetradhydrofolate reductase
Phosphoenolpyruvate mutase

79

Does structural similarity imply common function?

It depends.

80

Predicting Function: Considerations

- Assure that the database hits and predictive methods based on sequence yield information that ***make biological sense***
  - Predicted motifs or features biologically correct
  - Consistency with findings at the bench
- ***Even if one is able to predict function, the prediction can indeed turn out to be incorrect – experimental proof is absolutely essential!***

81