Bio 2304
Introduction to bioinformatics

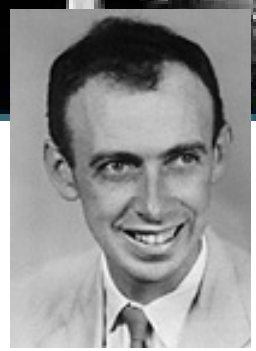
Prof. Nils Peder Willassen
Dept. of molecular biotechnology
Univ. of Tromsø

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø



”We have discovered the secret of life”





James Dewey Watson



Francis Harry Compton Crick

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Molecular structure of DNA

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.



We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

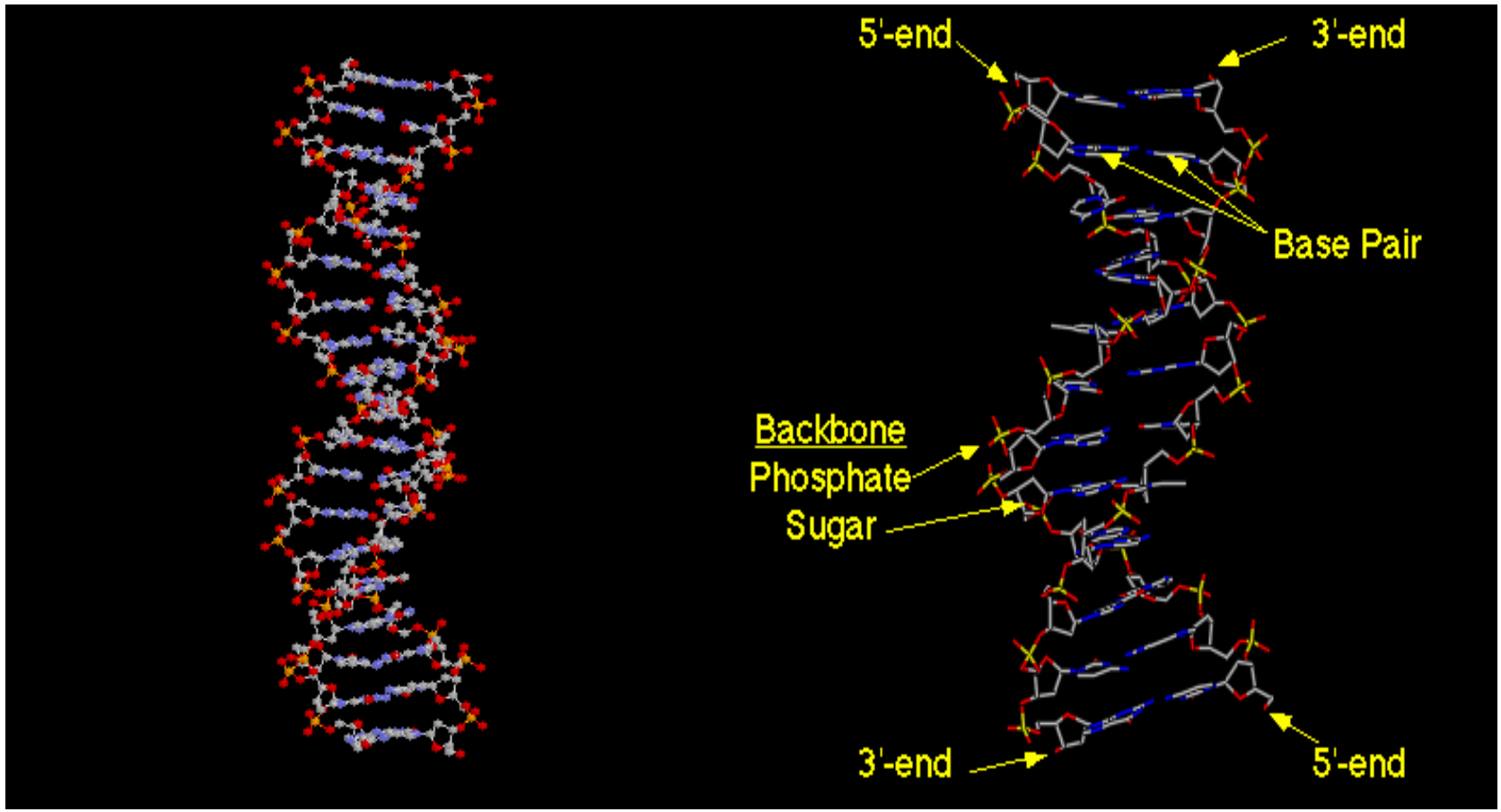
It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given

Nature
25 april 1953

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Molecular structure of DNA



Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

What was the relationship between DNA and proteins?

”Cracking" the genetic code

- Nucleotide to amino acid?

  4 DNA bases – 20 amino acids

- Doublet code?

  GT, GG, GA etc. (16 possible codons)

- Triplet code?

  GAT, CGA, AGT etc. (64 possible codons)

- Intermediate device?

  Adaptor hypothesis

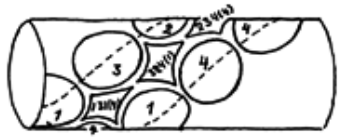
How to solve the problem?

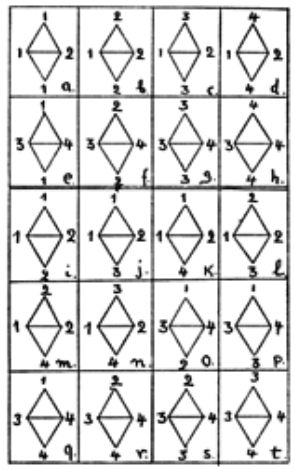
Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø



Attempted to solve this problem by combinatorics

- **George Gamow (Big bang theory)**
- **Francis Crick**





Coding scheme

- The Diamond Code
- Composition Code
- Comma-free Code



George Gamow

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Attempted to solve it biochemically

- **Marshall Nirenberg**
- **J. Heinrich Matthaei**

Synthetic RNA and the Poly-U experiments, 1959-1962

Artificial RNA could stimulate protein synthesis in a cell-free system



Marshall Nirenberg

January 15, 1962

Dr. Francis H. C. Crick
University of Cambridge
Department of Physics
Cavendish Laboratory
Free School Lane
Cambridge, England

I haven't seen the English newspapers but the American press has been saying that this type of work may result in (1) the cure of cancer and allied diseases (2) the cause of cancer and the end of mankind, and (3) a better knowledge of the molecular structure of God. Well, it's all in a day's work.

Sincerely yours,

Marshall W. Nirenberg

MWN:lh

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

”Cracking” the genetic code



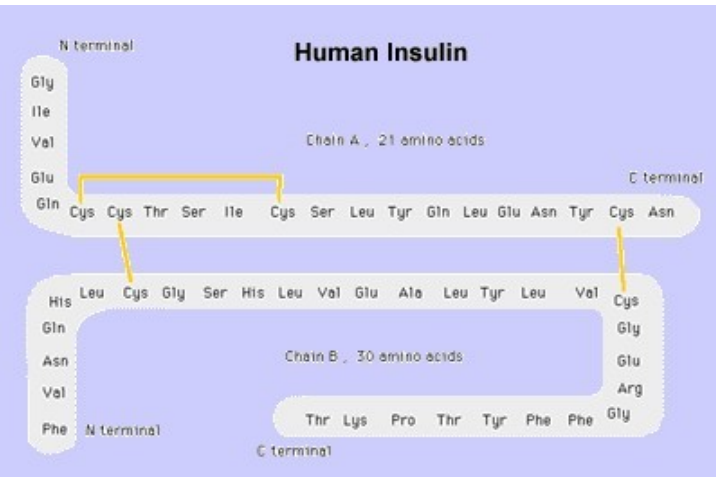
By 1966, Nirenberg announced that he had deciphered the sixty-four RNA codons for all twenty amino acids.

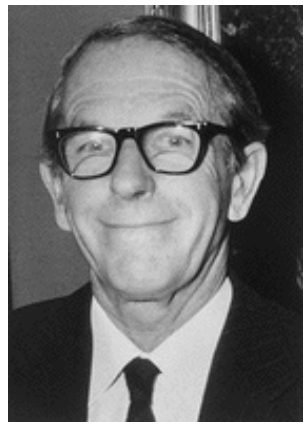
Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Amino acid sequencing

Sanger developed a new chromatographic method for determining end-groups and his results on the free amino groups of insulin were published in 1945.

Exploiting the techniques further he determined the N-terminal sequences of the two insulin chains in 1949. The complete sequence of insulin was published in 1955.



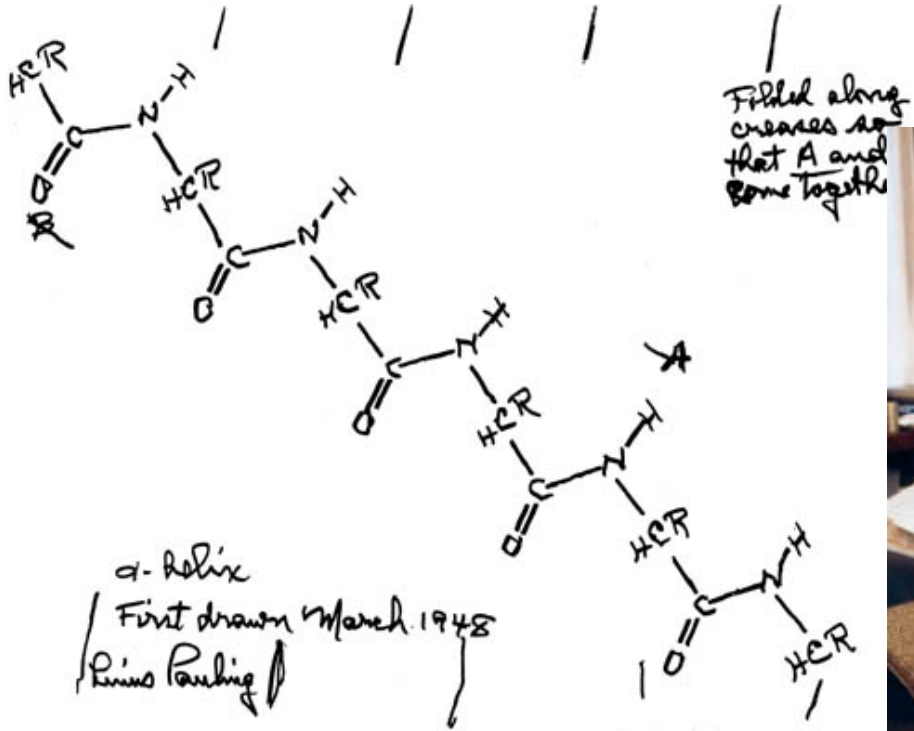


Frederick Sanger

Year	Protein	Number of residues
1935	Insulin	1
1945	Insulin	2
1947	Gramicidin S	5
1949	Insulin	9
1955	Insulin	51
1960	Ribonuclease	120

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Protein structure determination



Linus Pauling



1951: The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain

1953: Stable Configurations of Polypeptide Chains

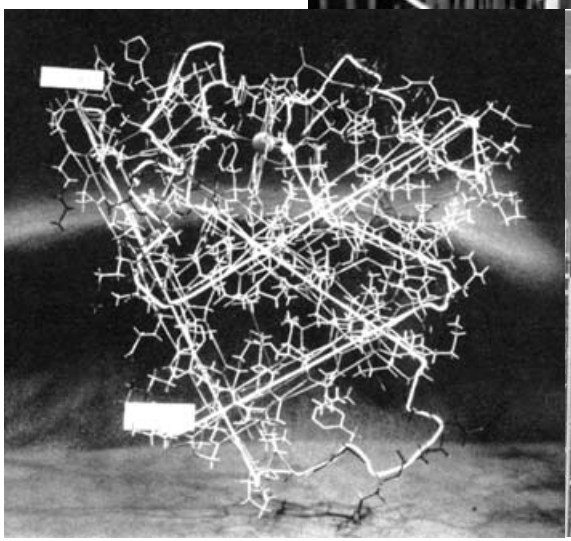
Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Protein structure determination





Max Ferdinand Perutz







John Cowdery Kendrew

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Protein structure determination

By 1959, Max was able to obtain the structure of haemoglobinat 5.5 Å resolution.

Kendrew determined the myoglobin structure at a resolution of 2 Å



Fig. 12. Complete haemoglobin molecule. The haem groups are indicated by grey disks.

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Protein structure database

The first protein structure databank, Brookhaven Protein Data Bank, was announced in 1973, established in 1977.

PDB Current Holdings Breakdown

		Molecule Type				
		Proteins	Nucleic Acids	Protein/NA Complexes	Other	**Total**
Exp. Method	X-ray	35898	981	1677	24	**38580**
	NMR	5625	779	134	7	**6545**
	Electron Microscopy	105	10	38	0	**153**
	Other	80	4	4	2	**90**
	Total	**41708**	**1774**	**1853**	**33**	**45368**

1993

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

DNA sequencing

In 1977, the first techniques for DNA sequencing were reported independently by A.M. Maxam and W. Gilbert, and F. Sanger and co-workers.

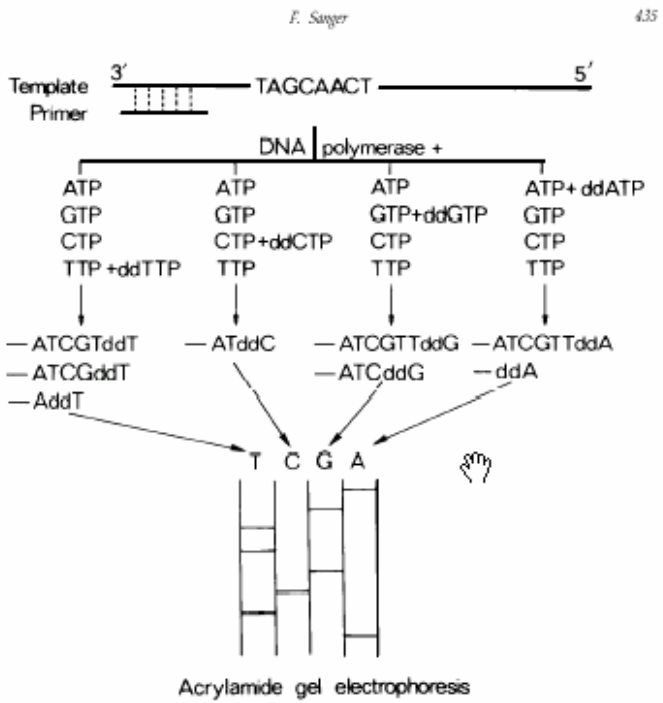


Fig. 3. Principle of the chain-terminating method.

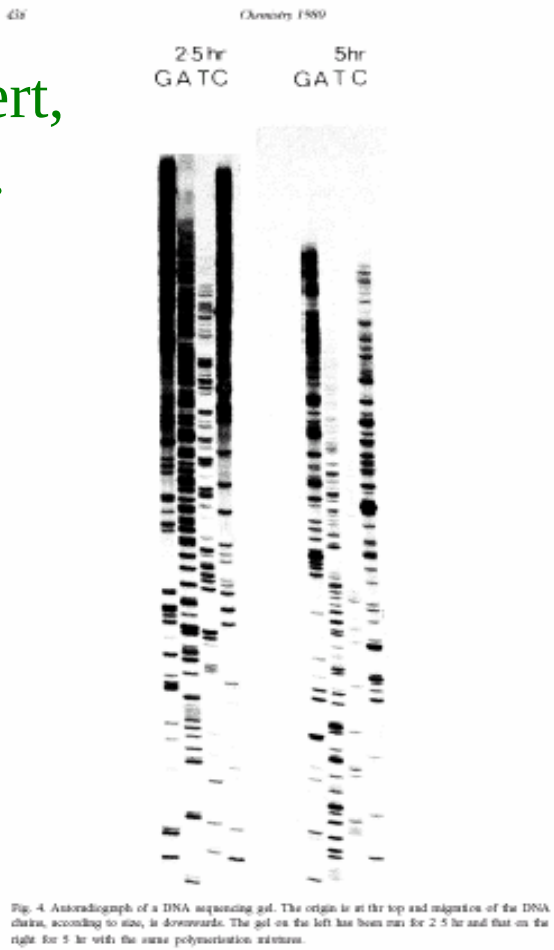
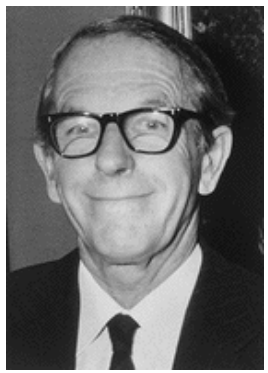


Fig. 4. Autoradiograph of a DNA sequencing gel. The origin is at the top and migration of the DNA chains, according to size, is downwards. The gel on the left has been run for 2·5 hr and that on the right for 5 hr with the same polymerization mixtures.



Walter Gilbert



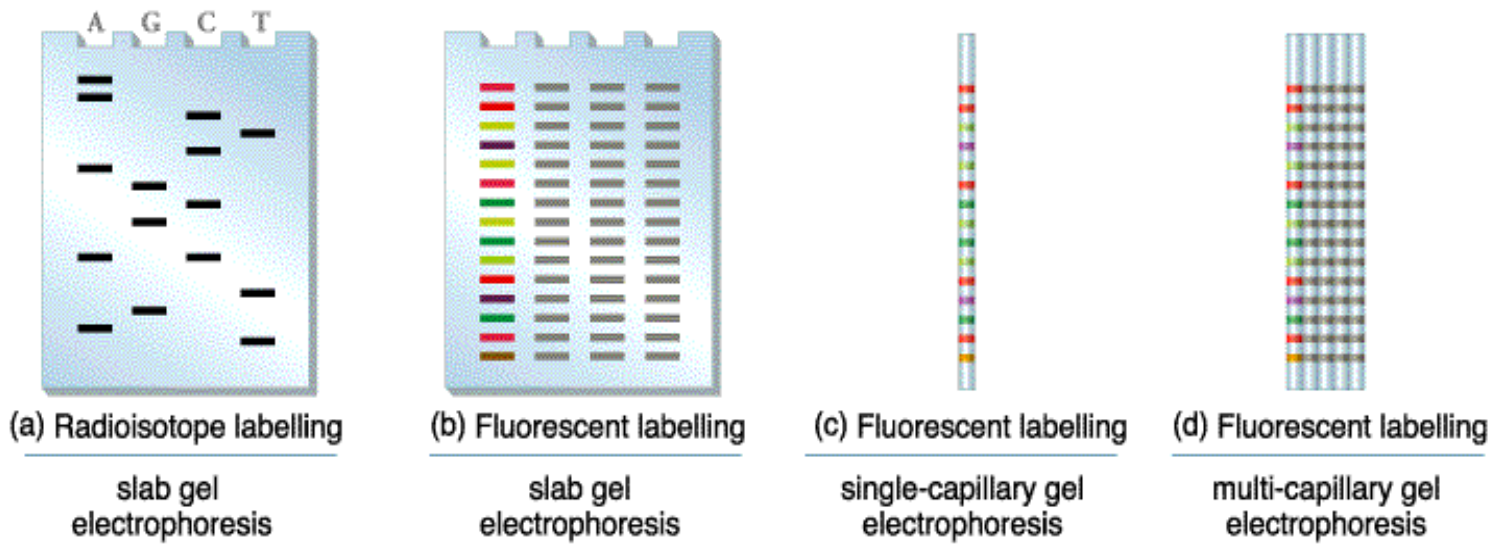
Frederick Sanger

otein Research group
ept. of Molecular Biotechnology
University of Tromsø

DNA sequencing

An advance in sequencing technology occurred in 1986 and 1987 when L. E. Hood and coworkers at Caltech, W. Ansorge and coworkers at EMBL, and J. M. Prober and coworkers at DuPont independently reported laser-induced fluorescence detection methods for DNA bases

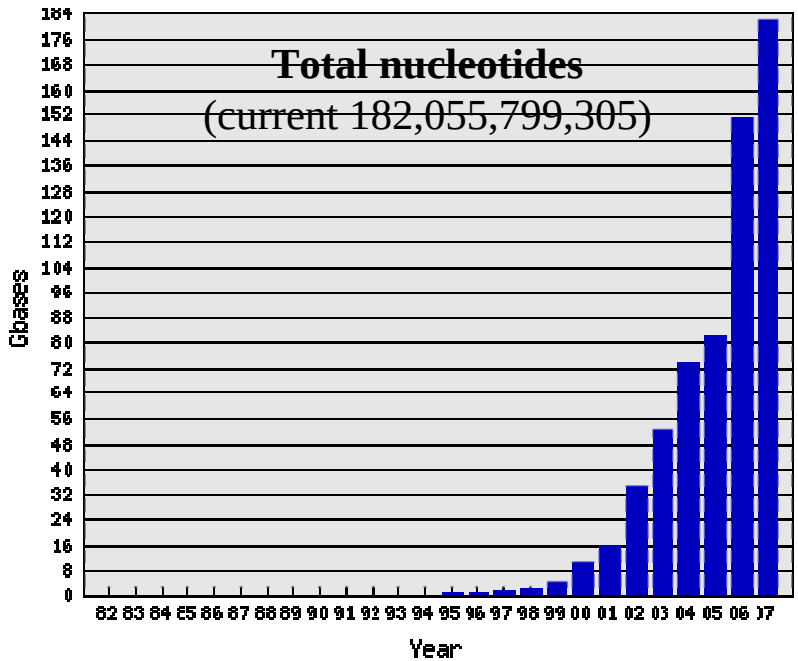
Figure 1. Evolution of DNA sequencers: from radio-isotope labeled slab-gel to fluorescent labeled capillary electrophoresis

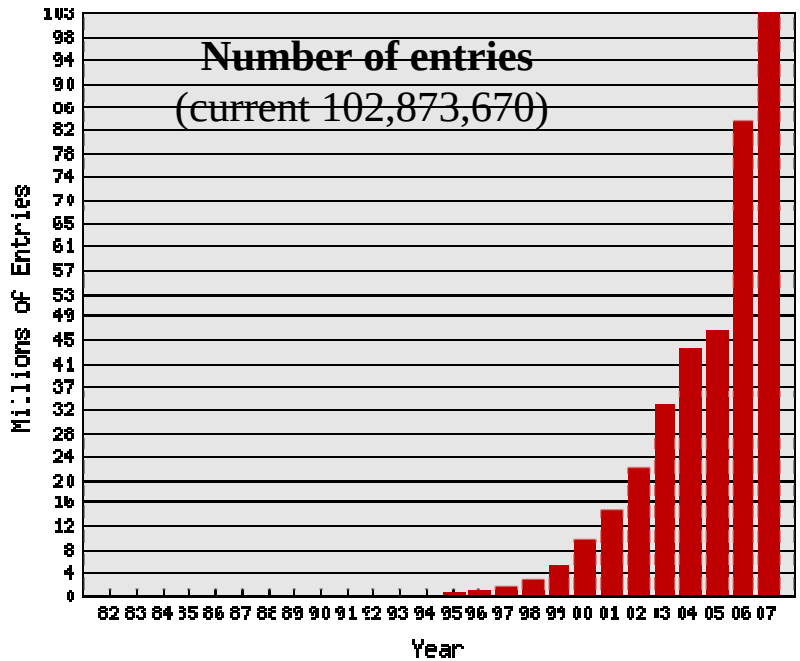


Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Sequence databases

The European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database established in 1980





EMBL Database contained **182,055,799,305** bases
in **102,873,570** records (26.08.07)

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Sequence alignment

Do the sequences share a common origin
- have a common ancestor sequence.

Have the sequences the same or related function ?

Have the sequences the same or related structure ?

How to search and retrieve related sequences ?

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø



Sequence alignment

```
T  C  A  T  G
C  A  T  T  G
```

2 matches, 0 gaps

```
T  C  A  T  G
      |  |
C  A  T  T  G
```

4 matches, 1 insertion

```
T  C  A  -  T  G
   |  |     |  |
.  C  A  T  T  G
```

3 matches (2 end gaps)

```
T  C  A  T  G  .
   |  |  |
.  C  A  T  T  G
```

4 matches, 1 insertion

```
T  C  A  T  -  G
   |  |  |     |
.  C  A  T  T  G
```

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø



Sequence alignment (pairwise)

- What to do for bigger string?

```
SSDSEREEHVKRFRQALDDTGMKVPMATTNLFTHPVFKDGGFTANDRDVRRYALRKTIRNIDLAVE
KPNEPRGDILLPTVGHALAFIERLERPELYGVNPEVGHEQMAGLNFPHGIAQALWAGKLFHIDLNG
```

- Needleman-Wunsch (1970) provided first automatic method

```
LGPSTKQFGKGSSSRIWDN
|       ||||    |   |
LNQIERSFGKGAIMRLGDA
```

Global alignment:

best alignment of entirety of both sequences

- Smith-Waterman (1981) algorithm for local alignment

```
-------FGKG--------
       ||||
-------FGKG--------
```

Local alignment:

best alignment of segments , without regard to rest of sequence

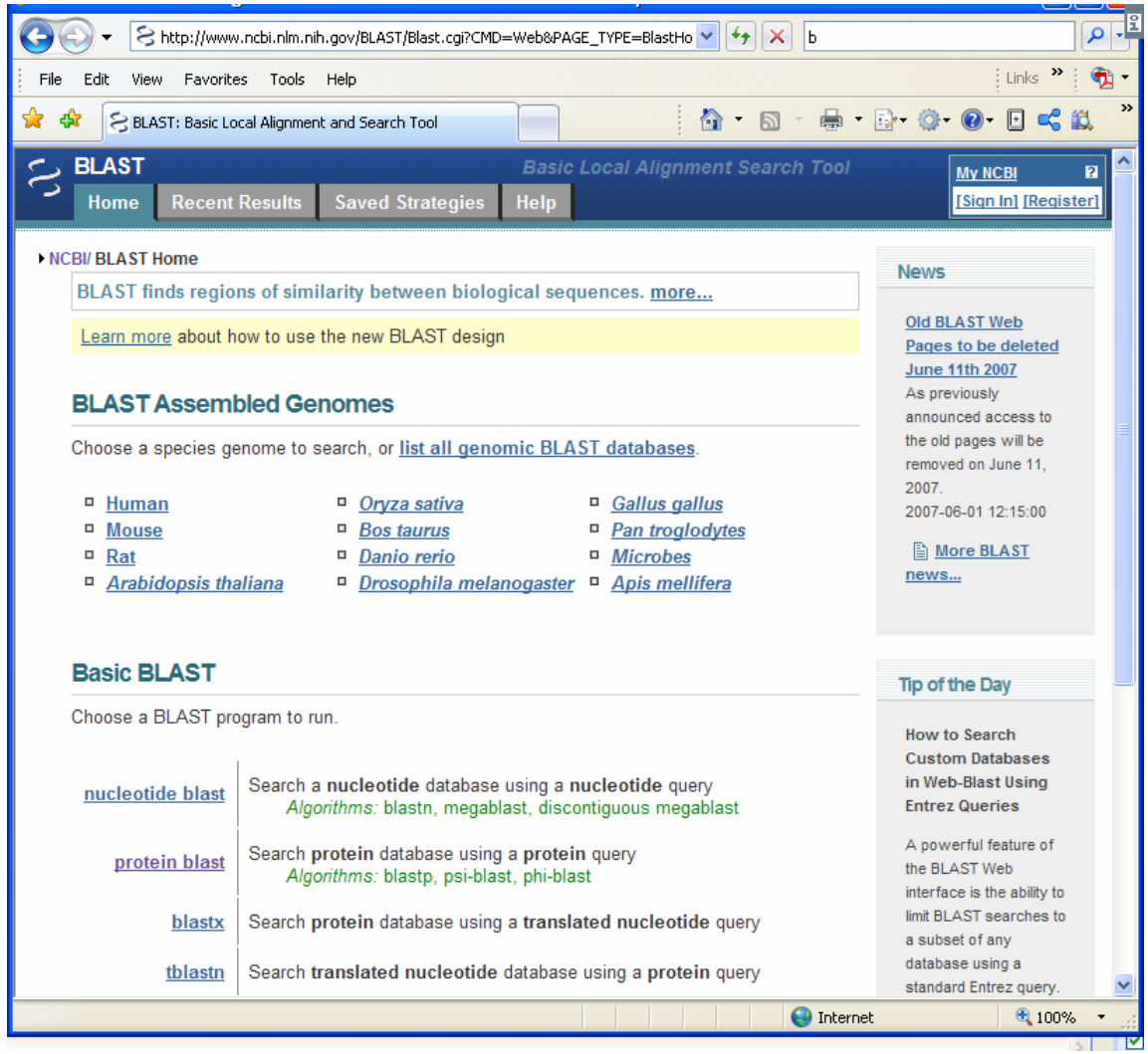
Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø



Database search

The FASTA algorithm for sequence comparison is published by Pearson and Lipman in 1988

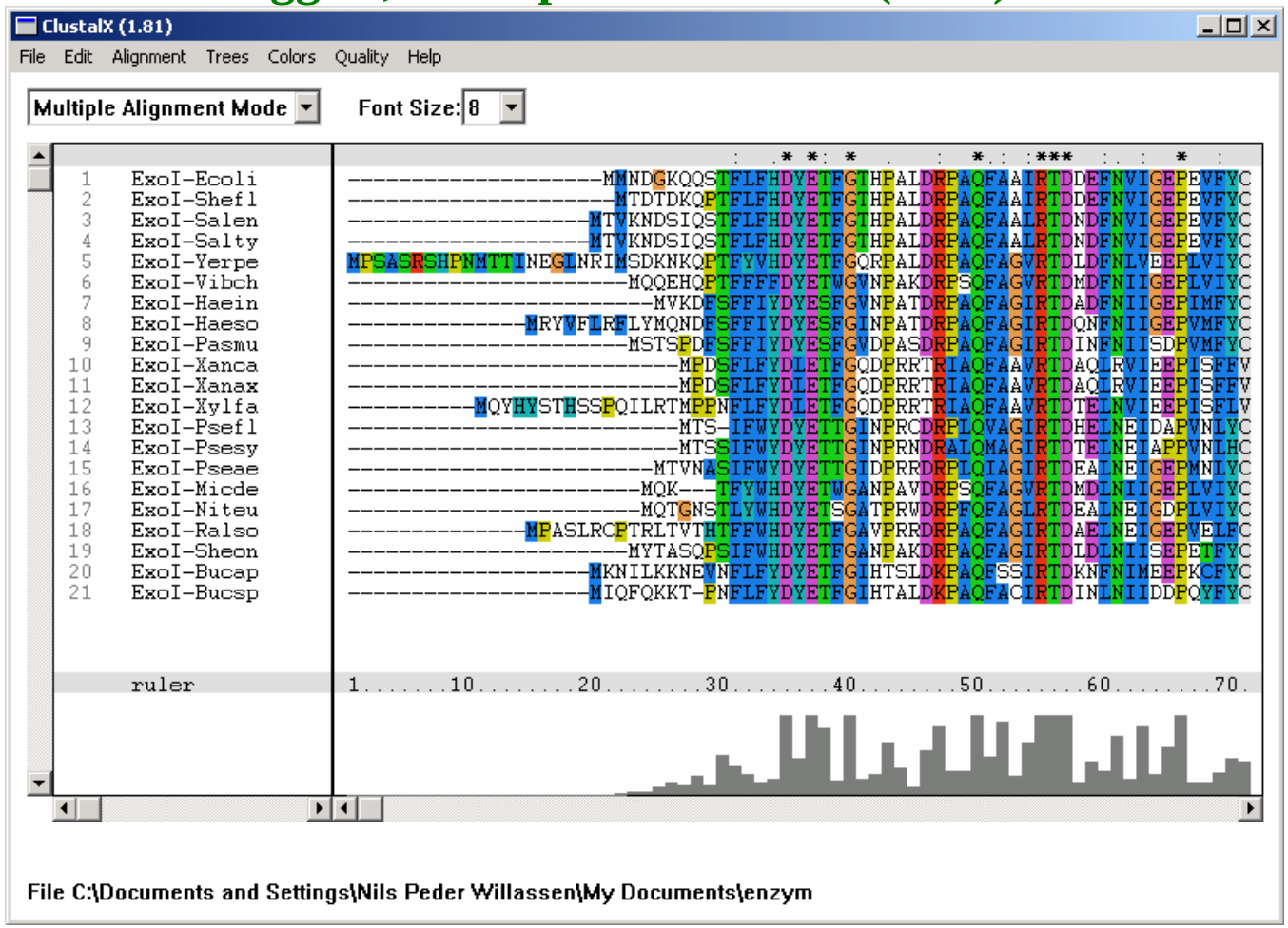
The BLAST program by Altschul, *et. al.* is implemented in 1990



Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Sequence alignment Multiple alignment

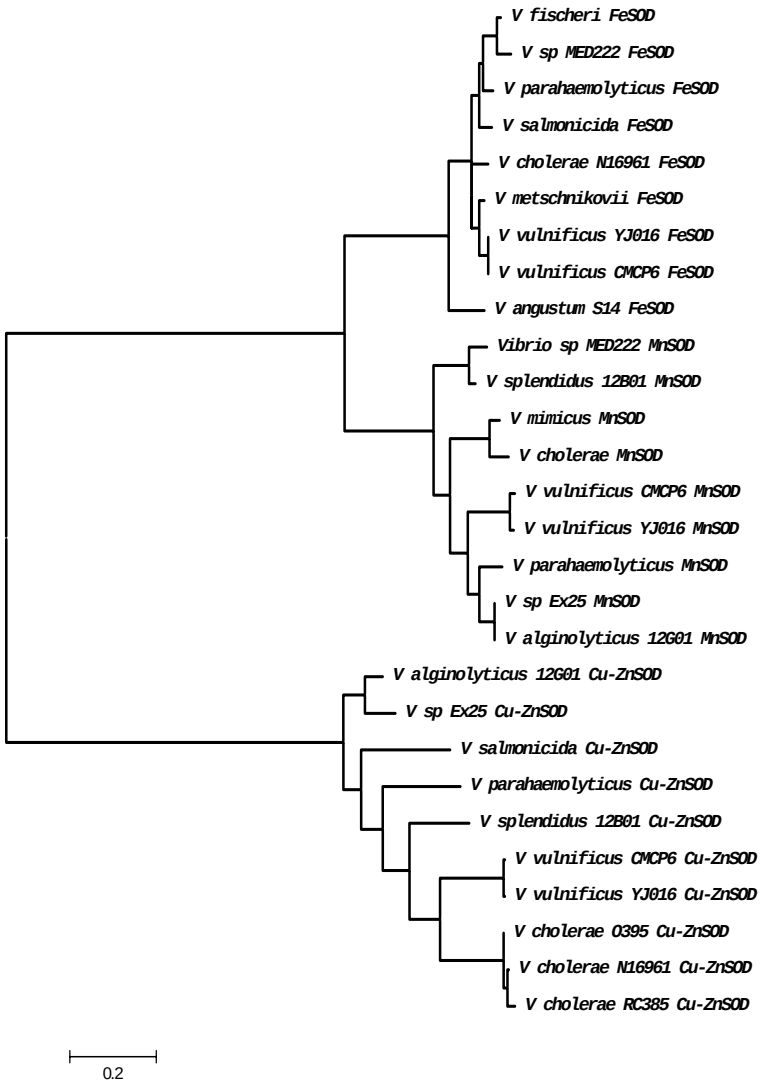
CLUSTAL W: Higgins, Thompson & Gibson (1994)



Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Phylogeny

MEGA: Higgins, Kumar, Kamura & Nei (1993)



Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Development of computers and www



Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

What is bioinformatics?

(Molecular) **Bio - informatics**

Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, computer science, and statistics) to understand and **organise the information associated** with these molecules, **on a large-scale.**

In short, bioinformatics is management information system for molecular biology and has many practical **applications**.

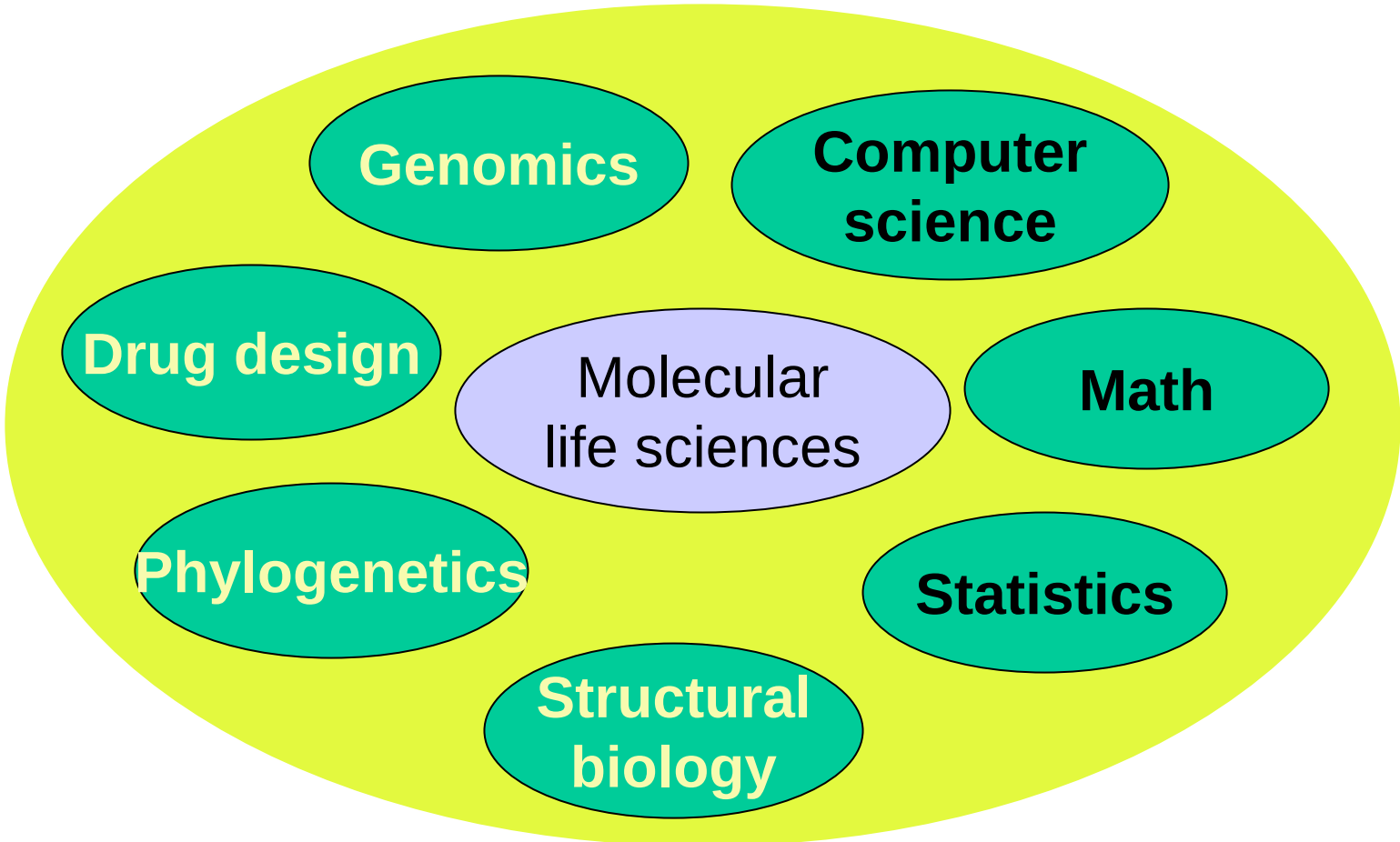
Mark Gerstein, Yale Univ.

Submitted to the Oxford English Dictionary

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

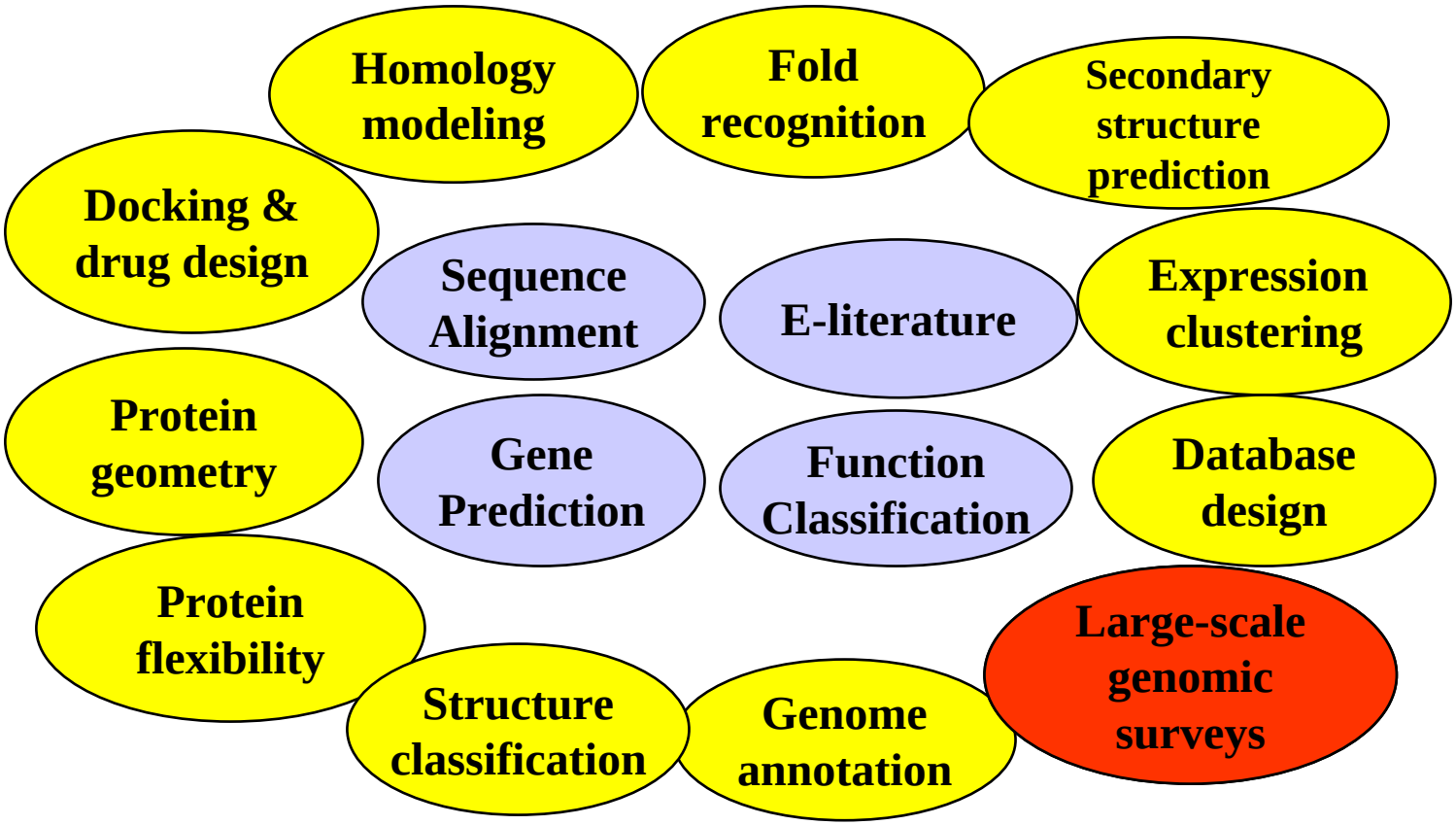


Bioinformatics



Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

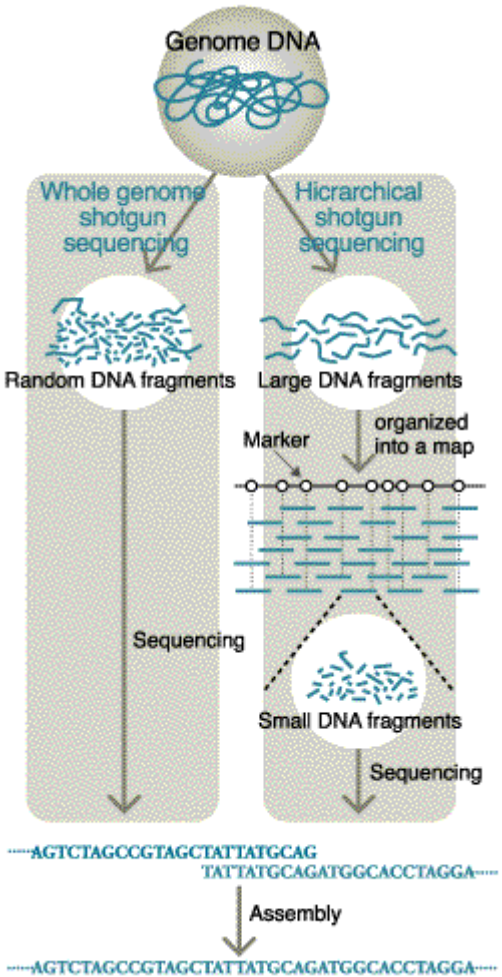
Bioinformatics subtopics

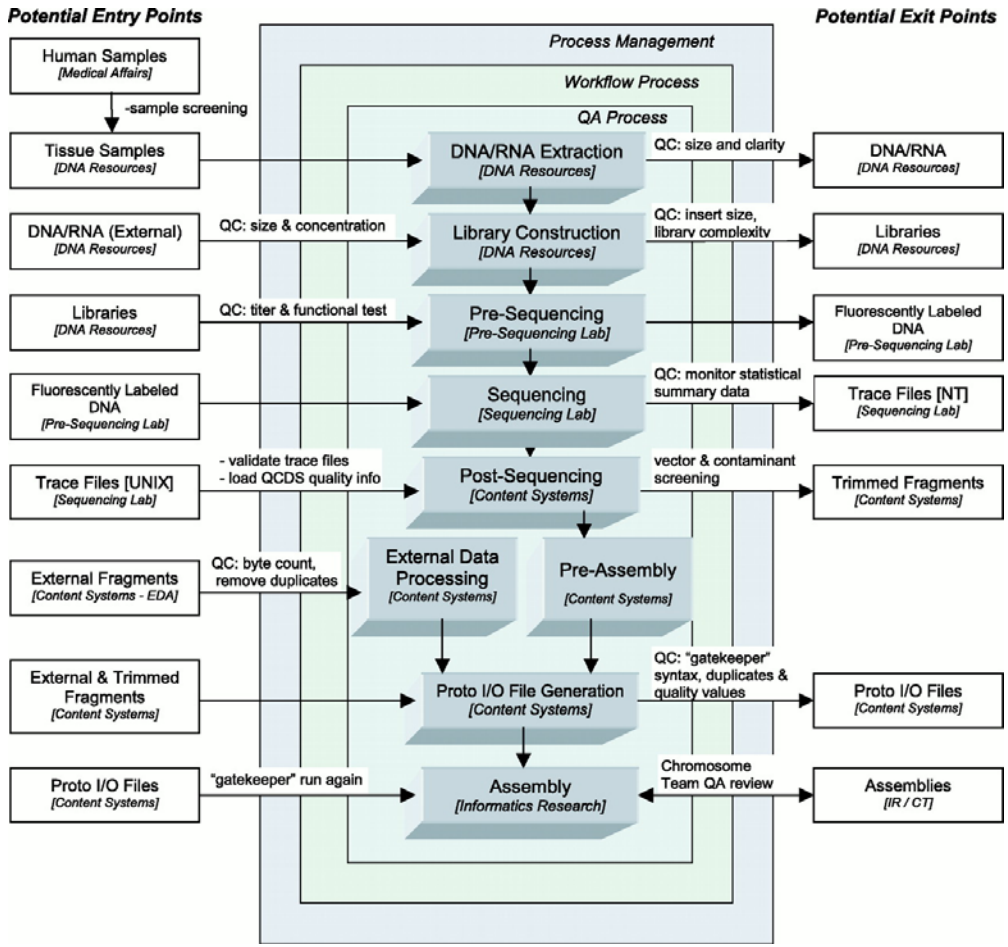


Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Genomics – genome sequencing

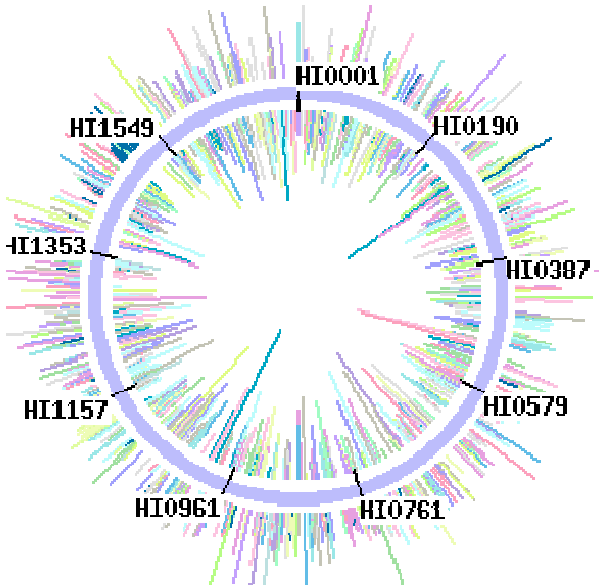
Figure 3. Schematic diagram of genome sequencing

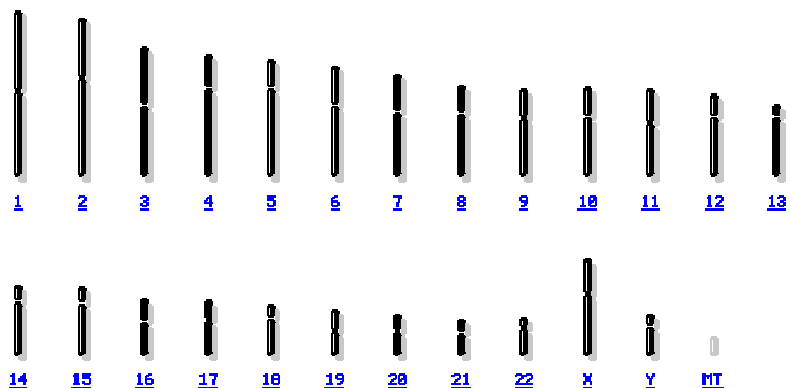




Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Genomics



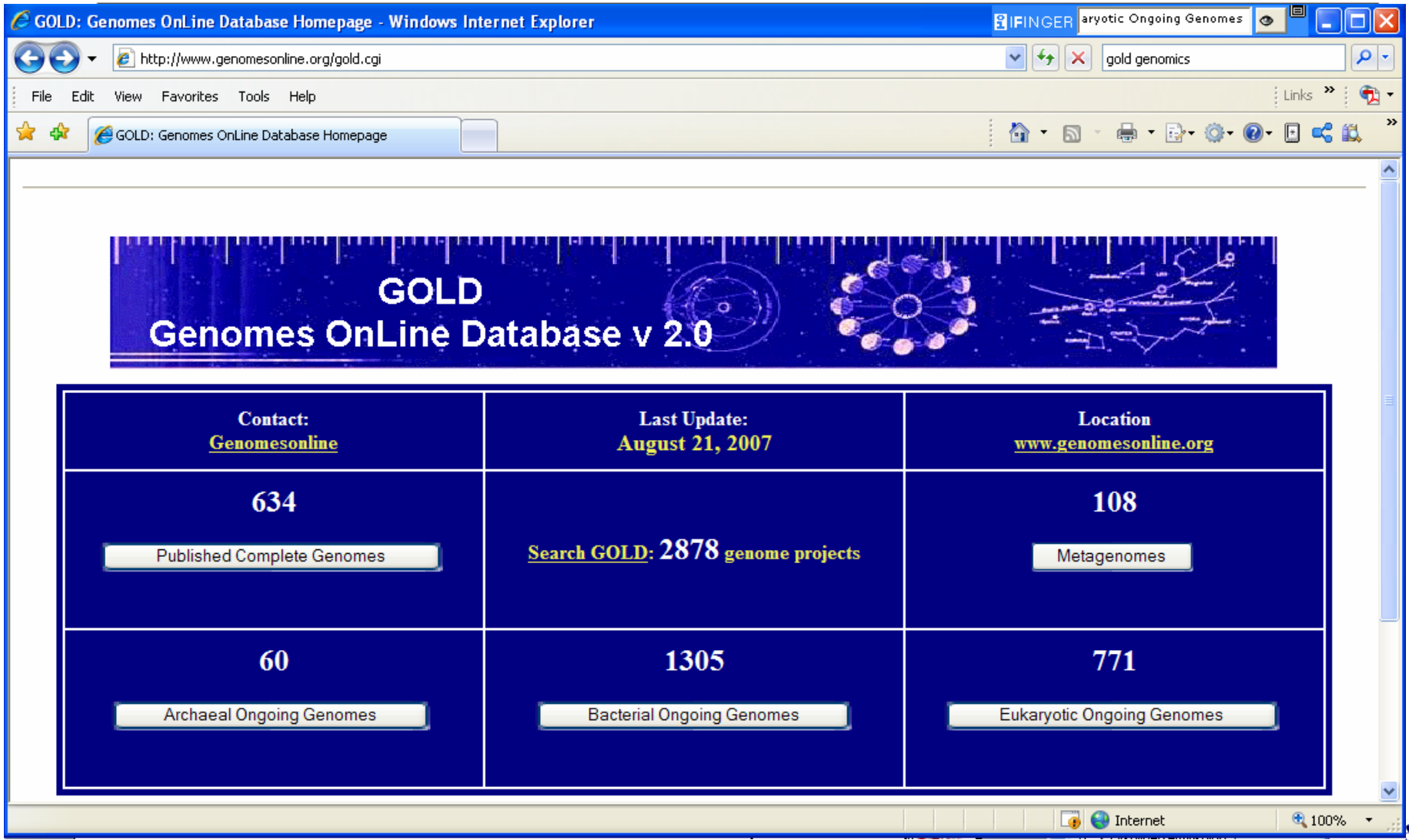


Whole-genome random sequencing and assembly of Haemophilus influenzae Rd
Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.-F., Dougherty,B.A., Merrick,J.M., McKenney,K., Sutton,G.G., FitzHugh,W., Fields,C.A., Gocayne,J.D., Scott,J.D., Shirley,R., Liu,L.I., Glodek,A., Kelley,J.M., Weidman,J.F., Phillips,C.A., Spriggs,T., Hedblom,E., Cotton,M.D., Utterback,T., Hanna,M.C., Nguyen,D.T., Saudek,D.M., Brandon,R.C., Fine,L.D., Fritchman,J.L., Fuhrmann,J.L., Geoghagen,N.S., Gnehm,C.L., McDonald,L.A., Small,K.V., Fraser,C.M., Smith,H.O. and Venter,J.C.
Science 269 (5223), 496-512 (1995)

Initial sequencing and analysis of the human genome
THE GENOME INTERNATIONAL SEQUENCING CONSORTIUM
Nature **409,** 860-921 (15 February 2001)

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Genomics



Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

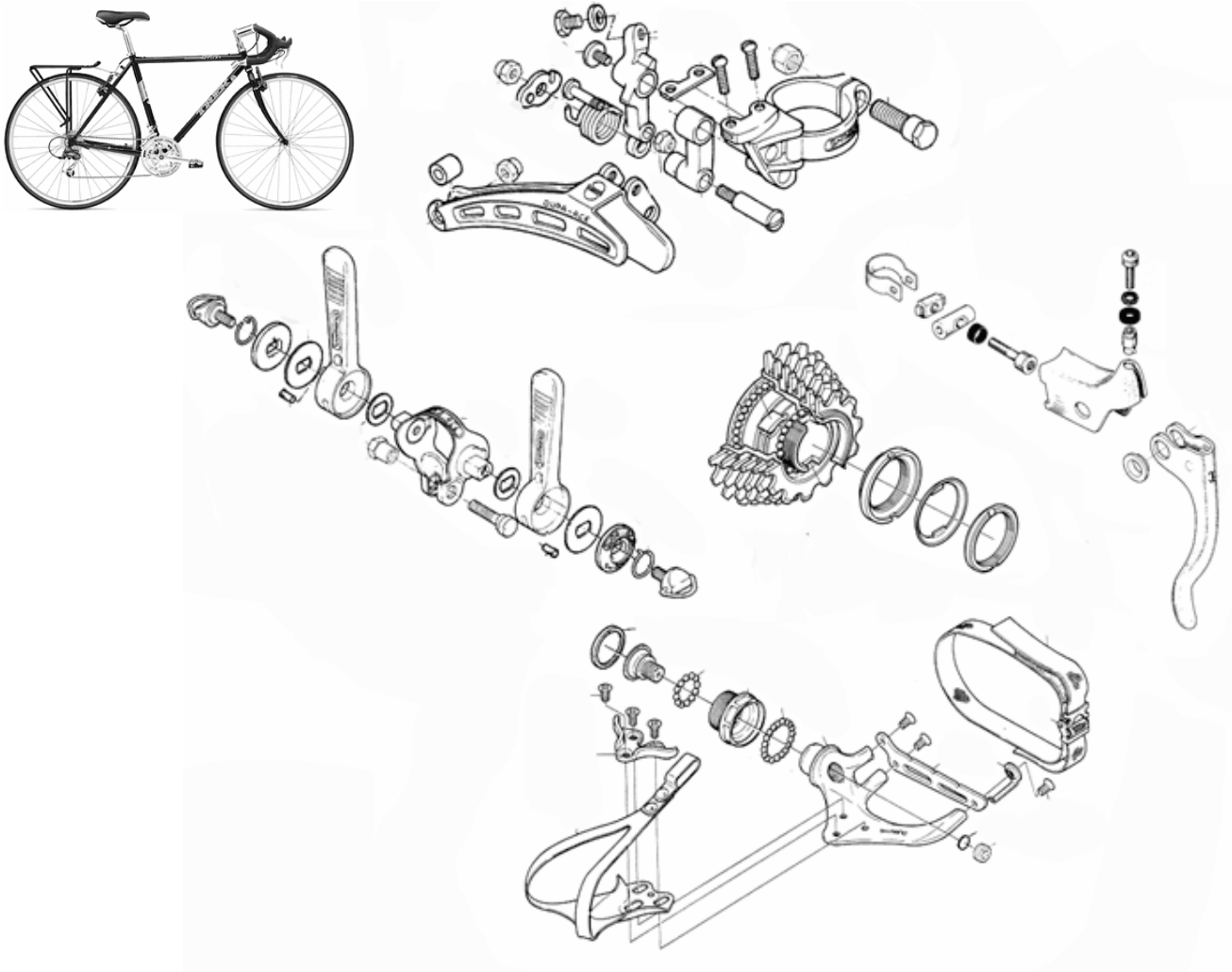
Genomics

Organism	Genome Size (Mbp)	# haploid chromosomes	# of genes* (thousands)
E. coli	4.64	1	4
Yeast (cerevisaie)	12.1	16	6
Nematode	100	6	18
Arabadopsis	100	5	25
Fruit fly	140	4	14
Rice	565	9	**55**
Human	**3,000**	**23**	**35**
Mouse	3,300	20	35
Corn	5,000	12	
Wheat	17,000	7 (hexaploid)	
Salamander	95,000	12	35

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø



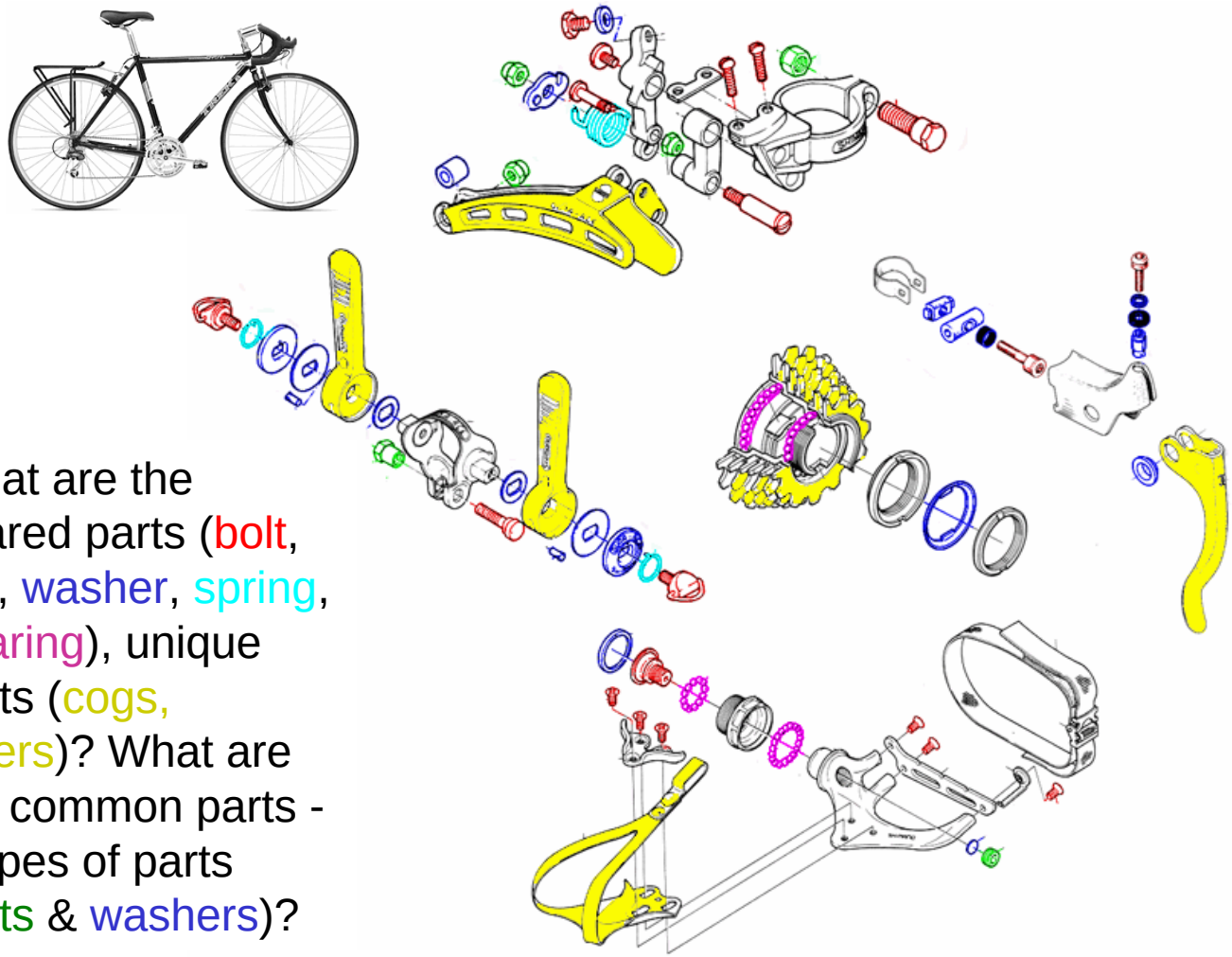
Genomics



Adapted from Mark Gerstein

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Genomics



How many roles can these play? How flexible and adaptable are they mechanically?

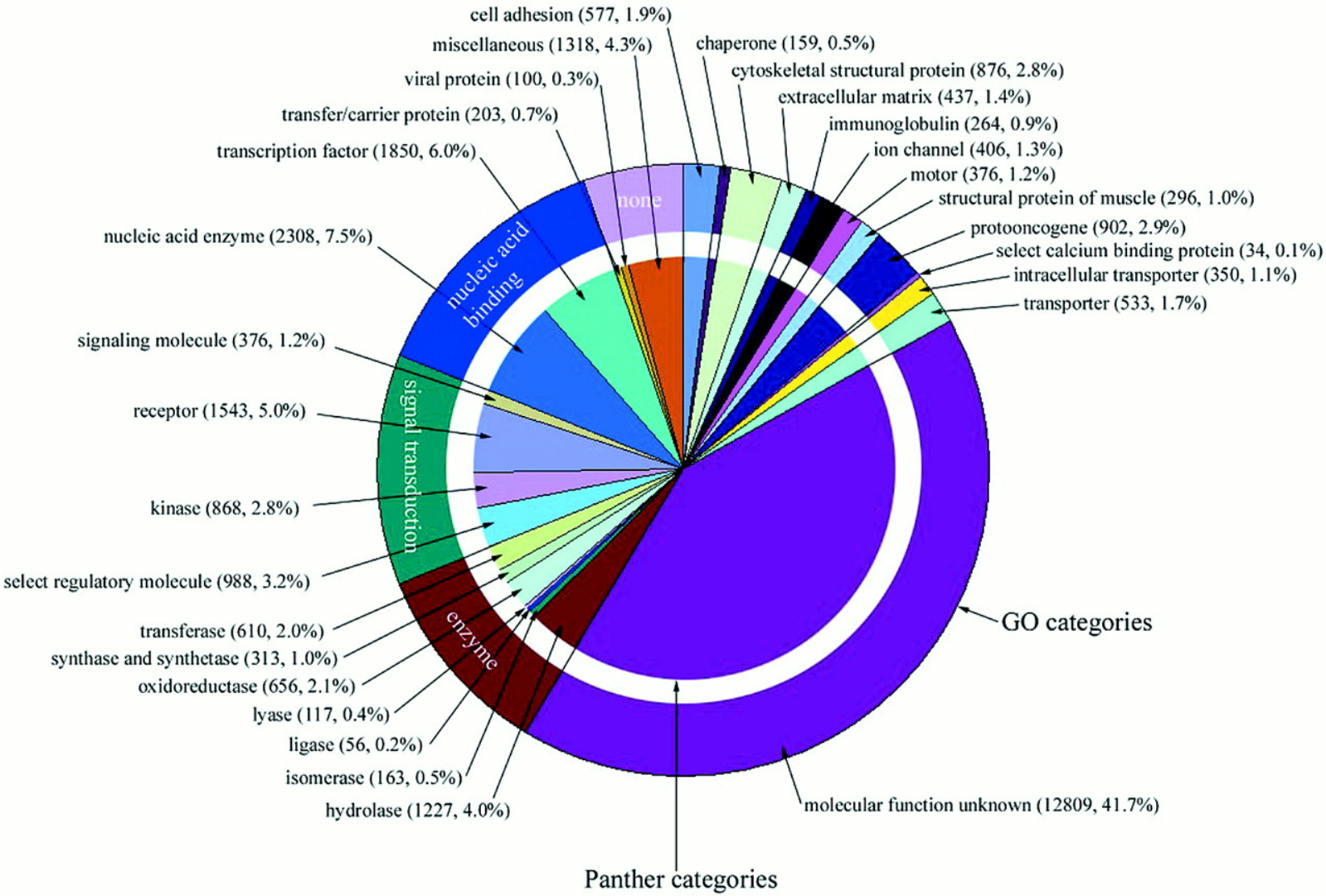
What are the shared parts (bolt, nut, washer, spring, bearing), unique parts (cogs, levers)? What are the common parts -- types of parts (nuts & washers)?

Where are the parts located?

Adapted from Mark Gerstein

Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Genomics



Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Genomics – beyond (post genome)

Whole genome:

Comparative genomics

Functional genomics (FUGE)

Structural genomics

Integrative genomics

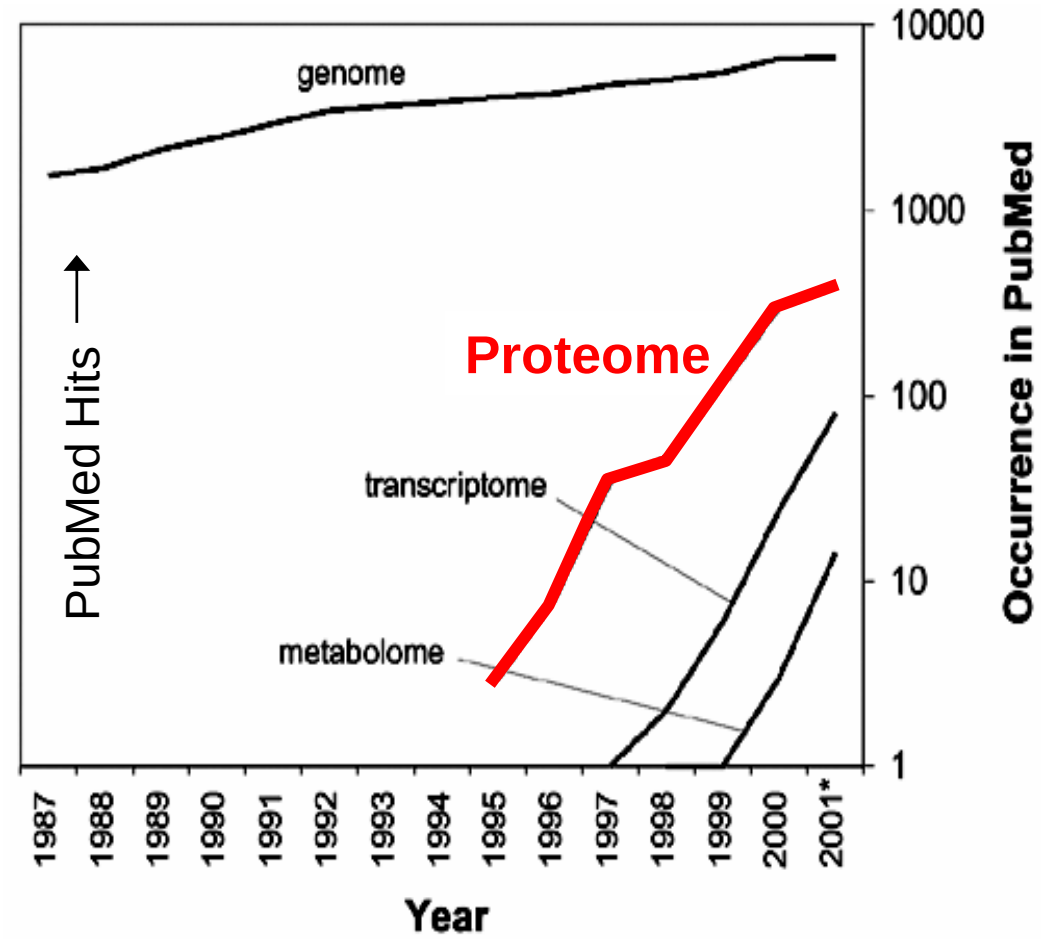
Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø



Genomics

The central post-genomic term

	Google Hits	Pub-Med Hits
Genome	~1880000	66171
Proteome	**~63,000**	**703**
Transcriptome	3520	72
Physiome	2980	15
Metabolome	349	12
Phenome	4980	6
Morphome	238	2
Interactome	56	2
Glycome	46	1
Secretome	21	1
Ribonome	1	1
Orfeome	42	-
Regulome	18	-
Cellome	17	-
Operome	8	-
Transportome	1	-
Functome	1	-



Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

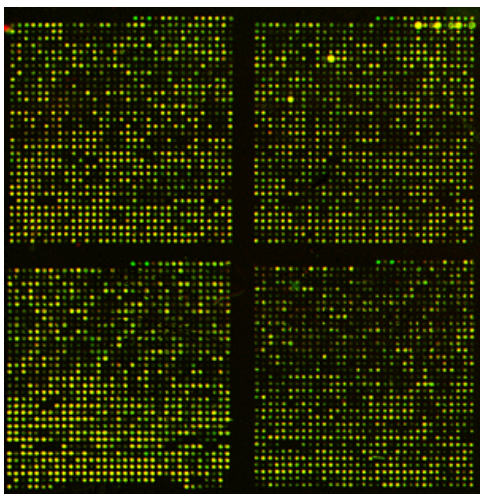
Genomics

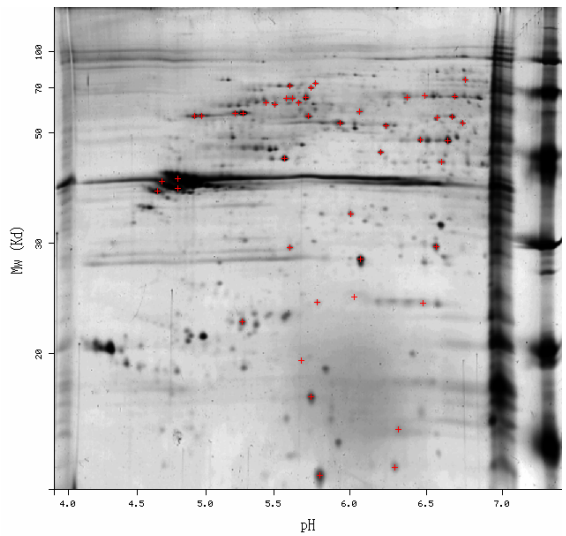
Functional genomics

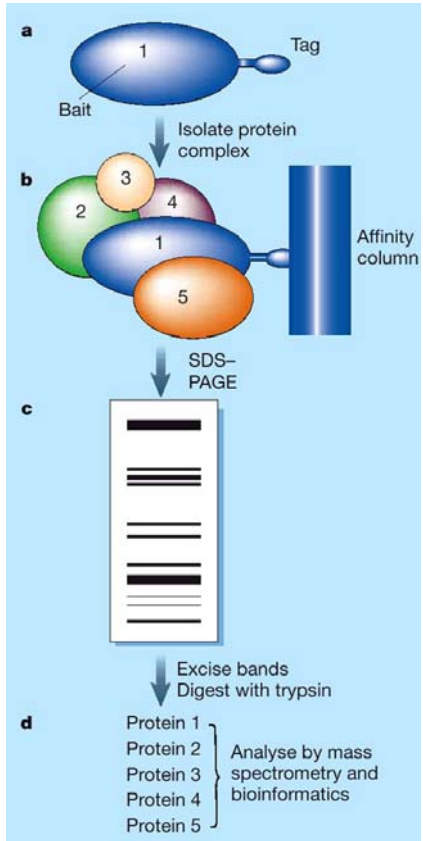
- Microarray (DNA/protein)
- Proteomics

    2D gel electrophoresis

    TAP (tandem affinity purification)





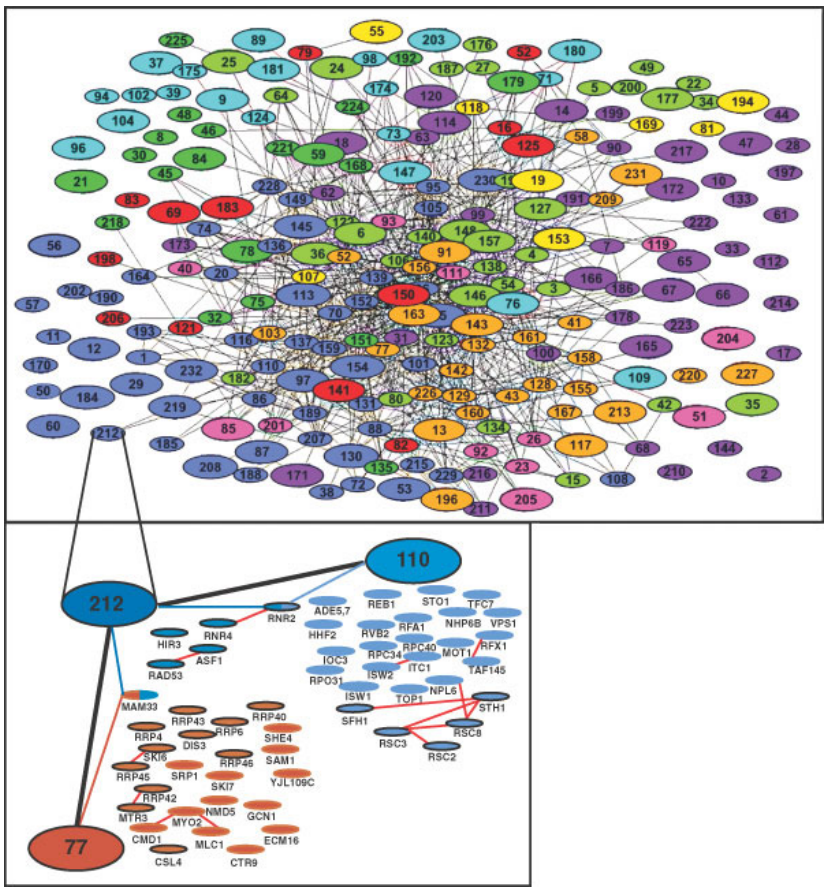


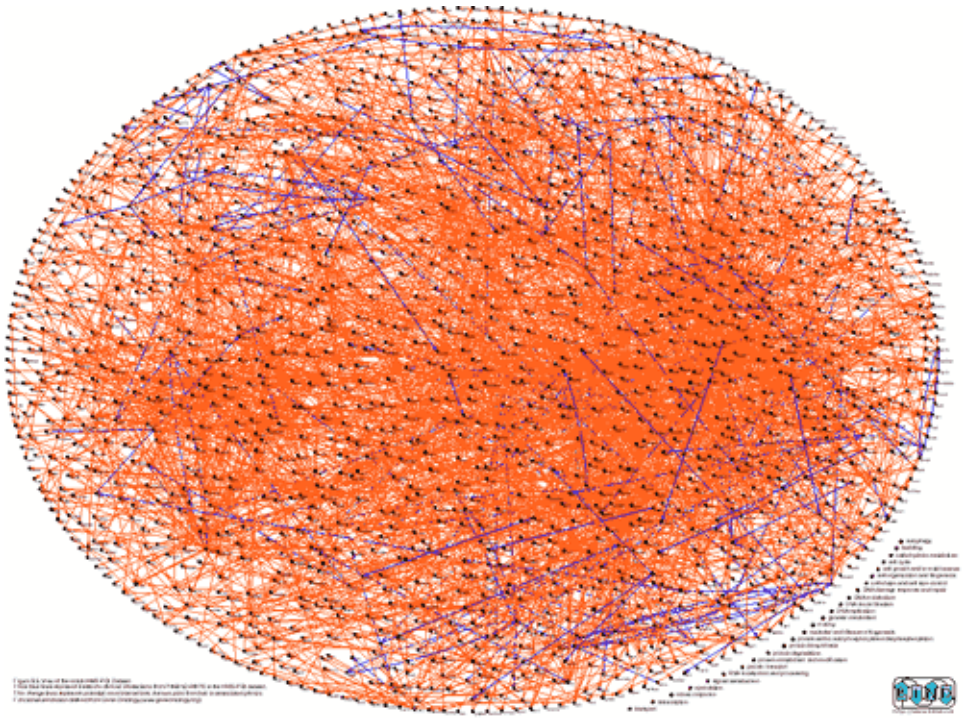
Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Genomics

Proteomics

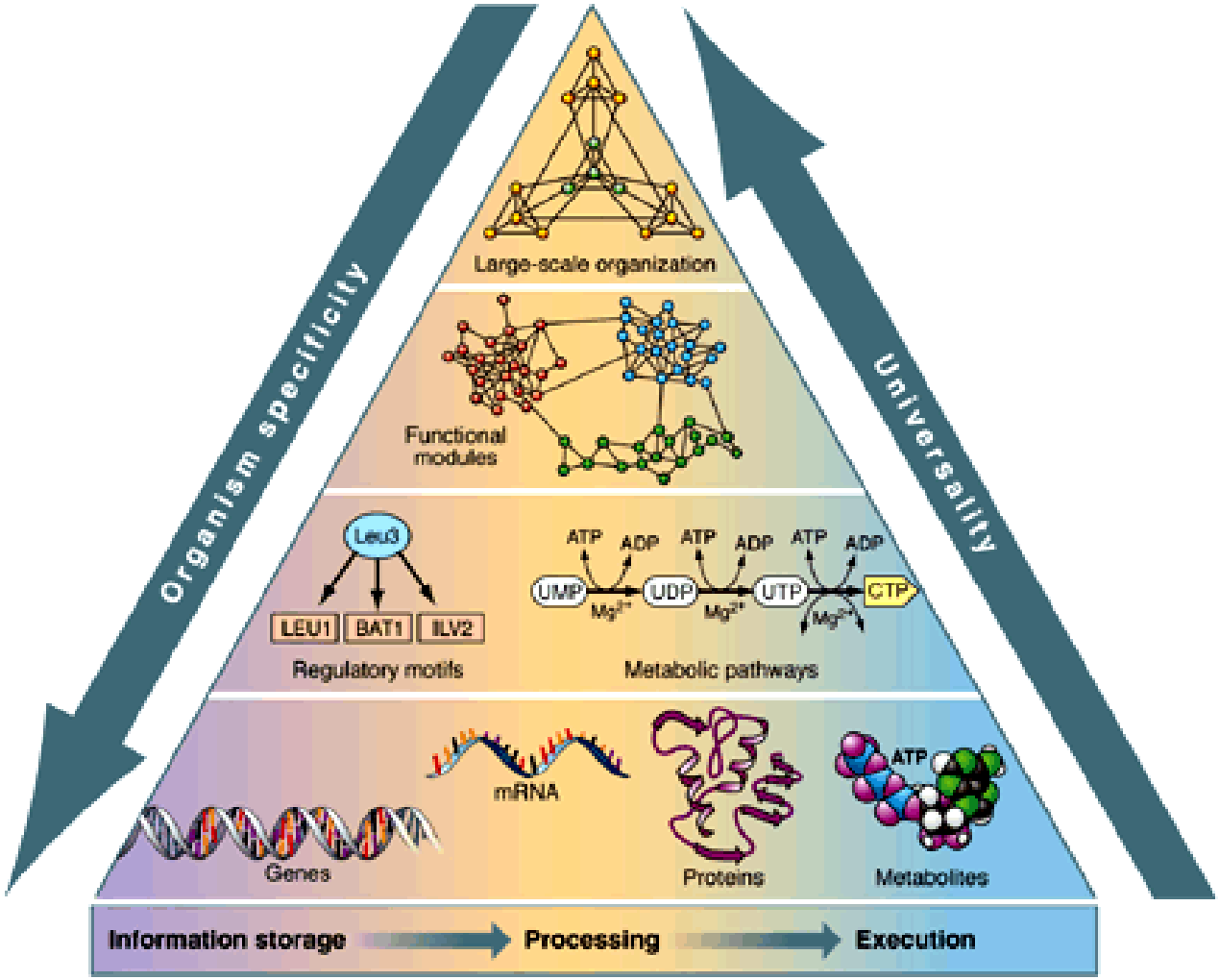
Understand Proteins, through analyzing populations





Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Genomics



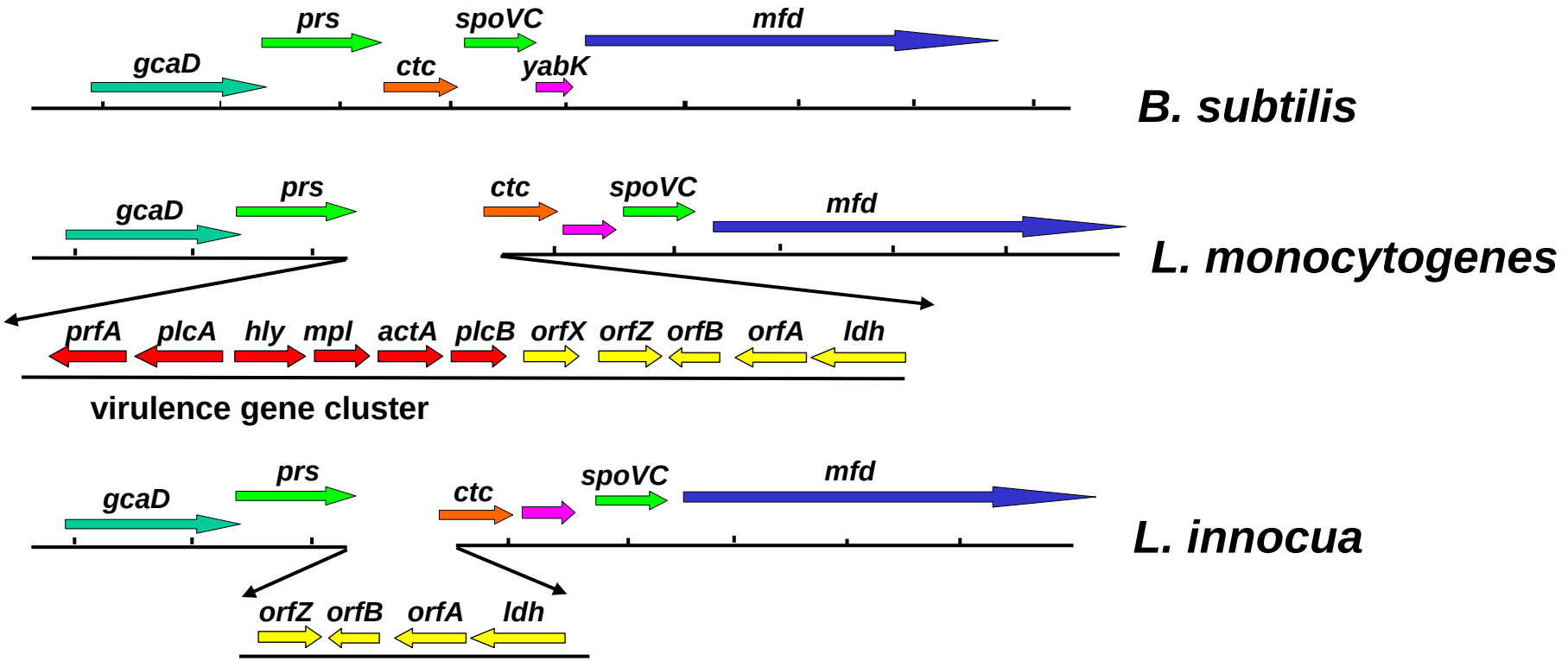
Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Major applications



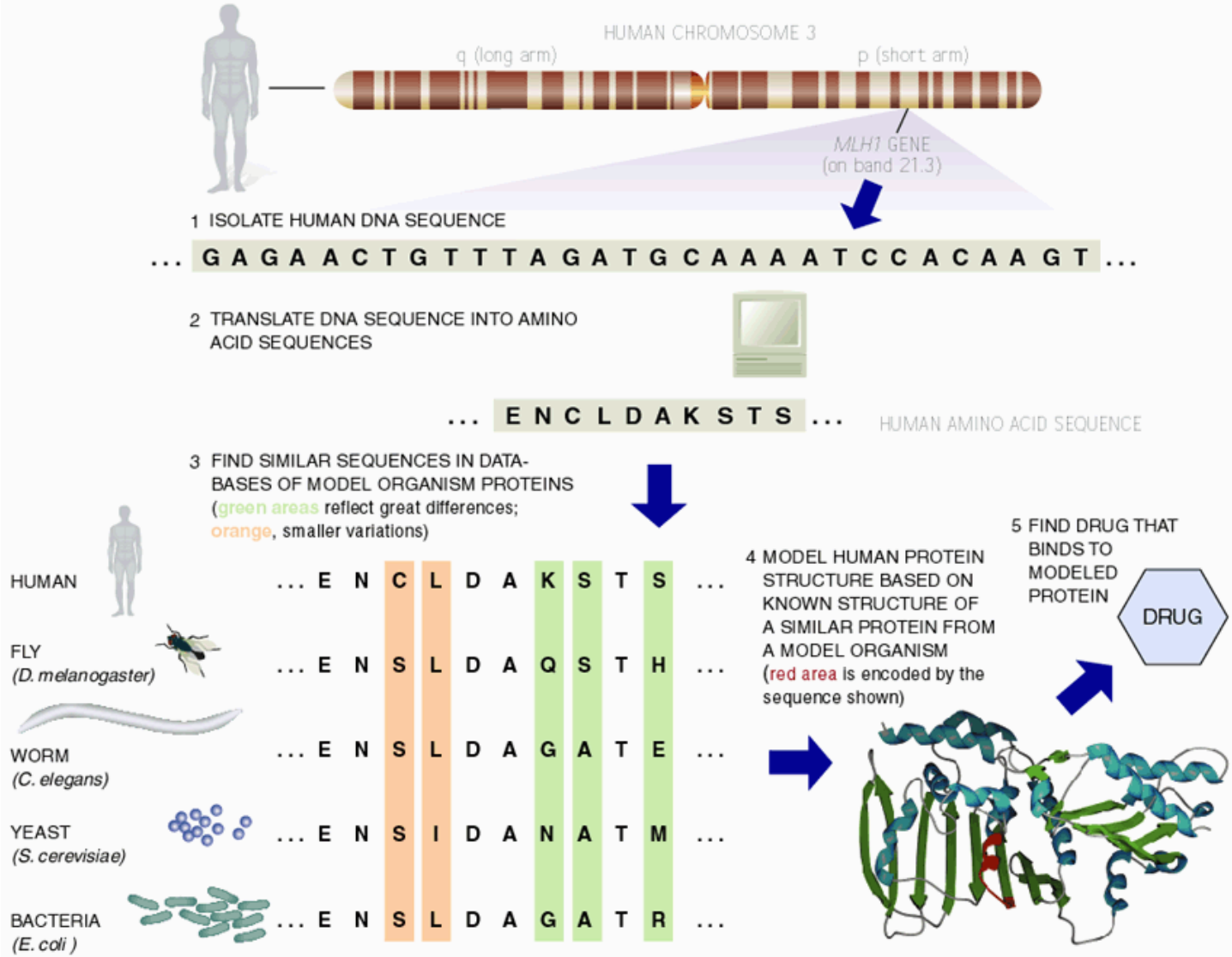
Major applications - vaccines

- **Identification of virulence genes**



Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

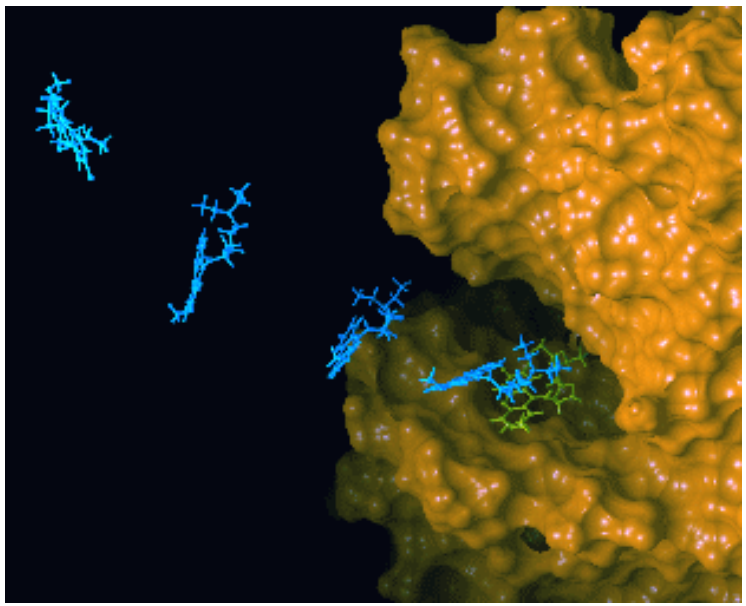
Major applications – finding homologs

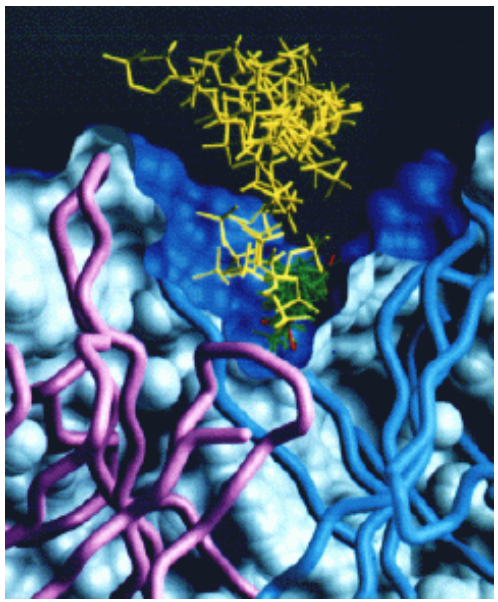


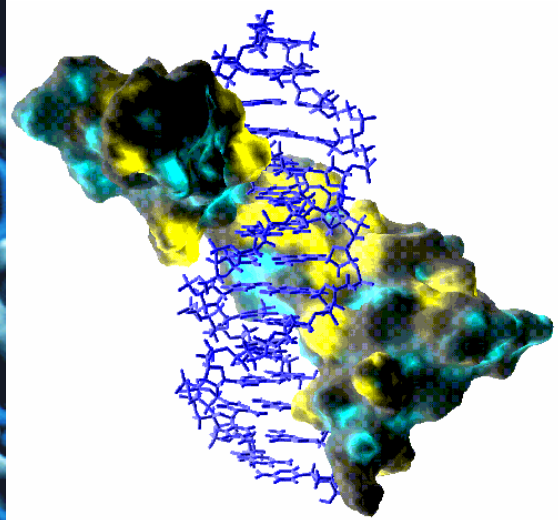
Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

Major applications - drug design

- Understanding how structures bind other molecules (function)
- Designing inhibitors
- Docking, structure modeling







From left to right, figures adapted from Olsen Group Docking Page at Scripps, Dyson NMR Group Web page at Scripps, and from Computational Chemistry Page at Cornell Theory Center.

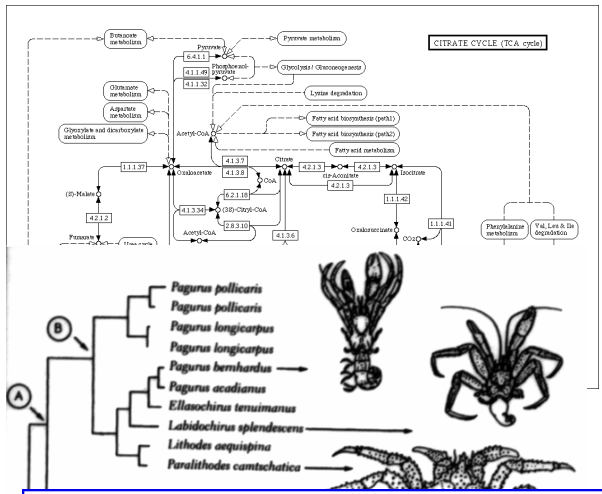
Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø

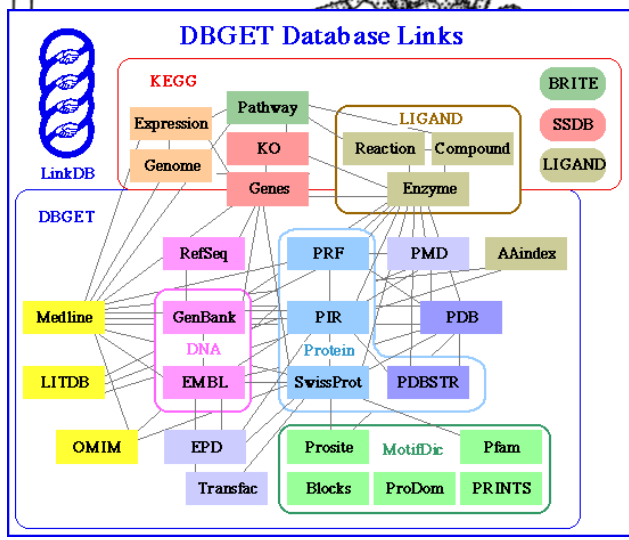
Major applications – integrate data

- Informationsystems to understand genomes
  - Metabolic pathways (glycolysis), traditional biochemistry
  - Regulatory networks
  - Whole organisms phylogeny, traditional zoology
  - Environments, habitats, ecology etc.
  - Literature (MEDLINE)

- The future

  ..... understand and create life





Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø



Protein Research group
Dept. of Molecular Biotechnology
University of Tromsø