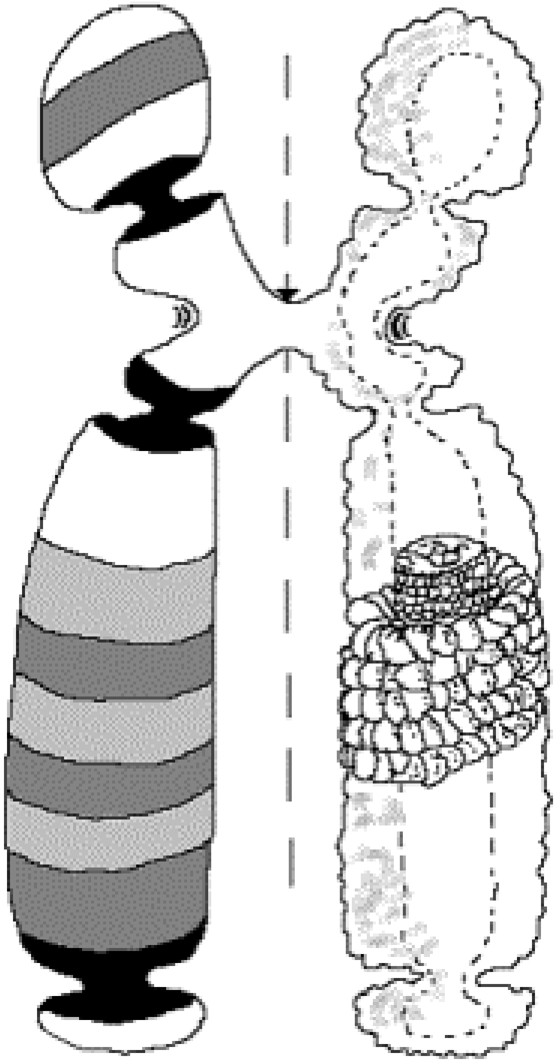
Open Source and the Human Genome Project

About this talk

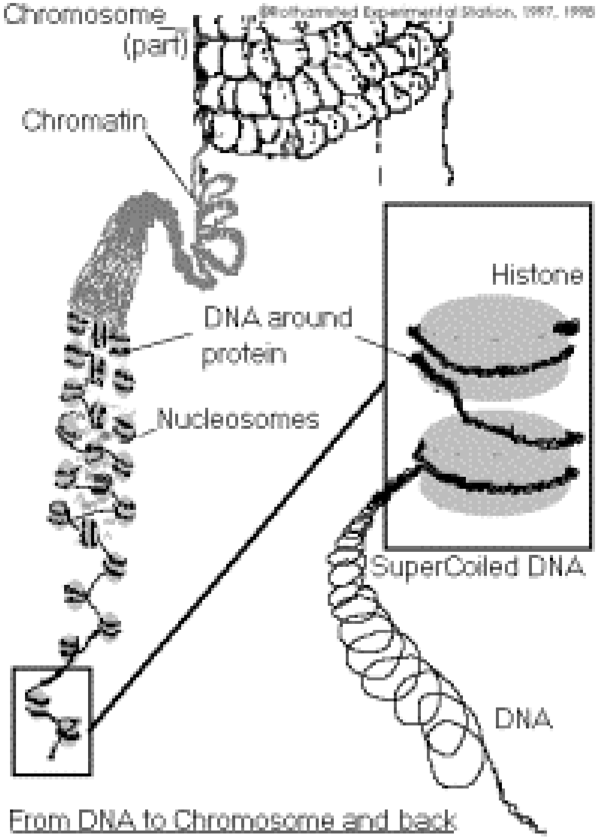
- who I am
- what I do
- overview»

- Introduction
- Crash course in Molecular biology
- The Human Genome Project
- “Gene patenting”
- Open Source bioinformatics
- The next wave

Molecular biology:
a crash course

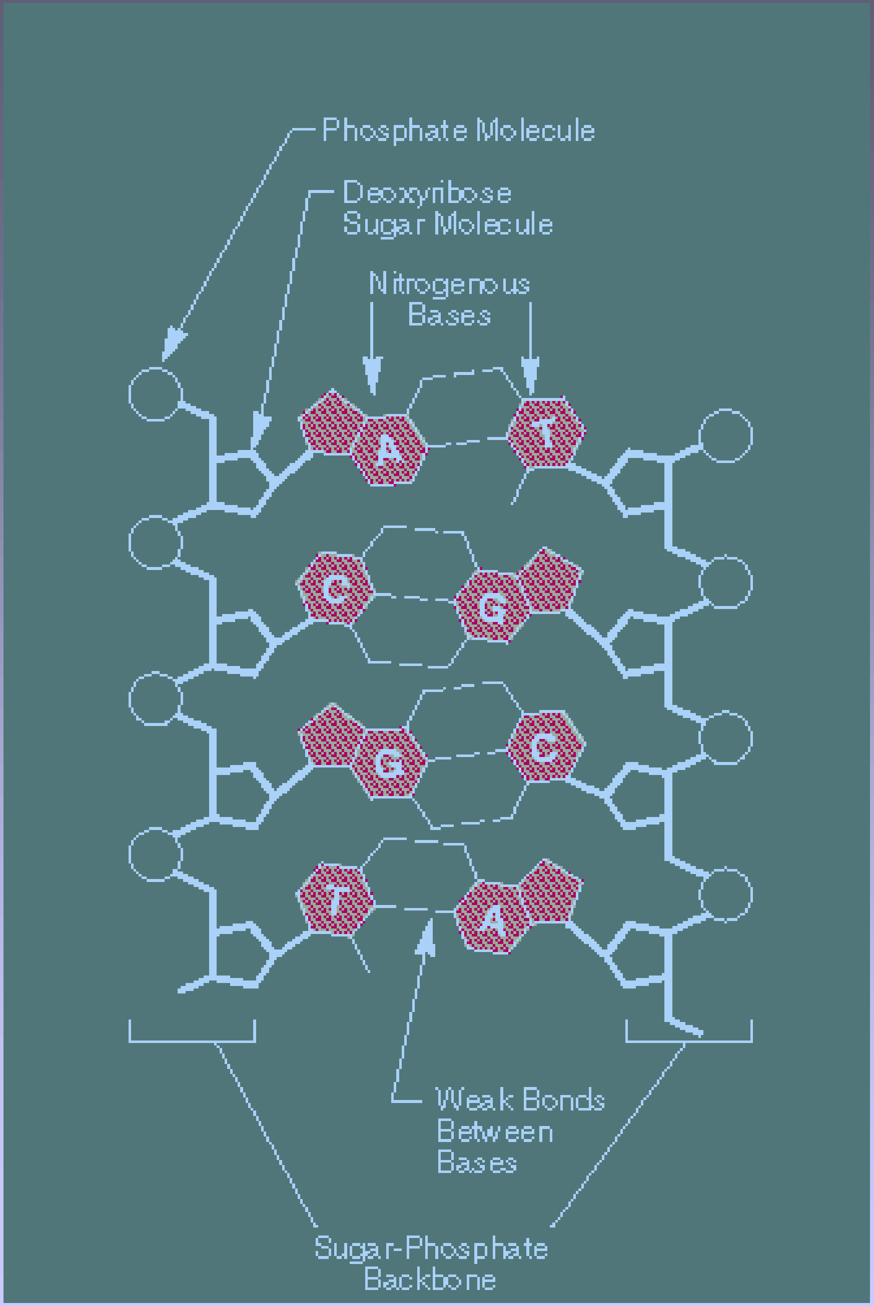


Metaphasic Chromosome

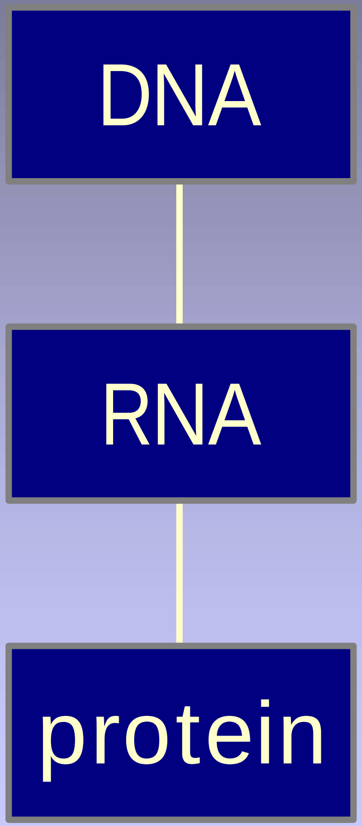


©Rothamsted Experimental Station, 1997, 1998

From DNA to Chromosome and back


Phosphate Molecule
Deoxyribose Sugar Molecule
Nitrogenous Bases
A
T
C
G
G
C
T
A
Weak Bonds Between Bases
Sugar-Phosphate Backbone

The central dogma

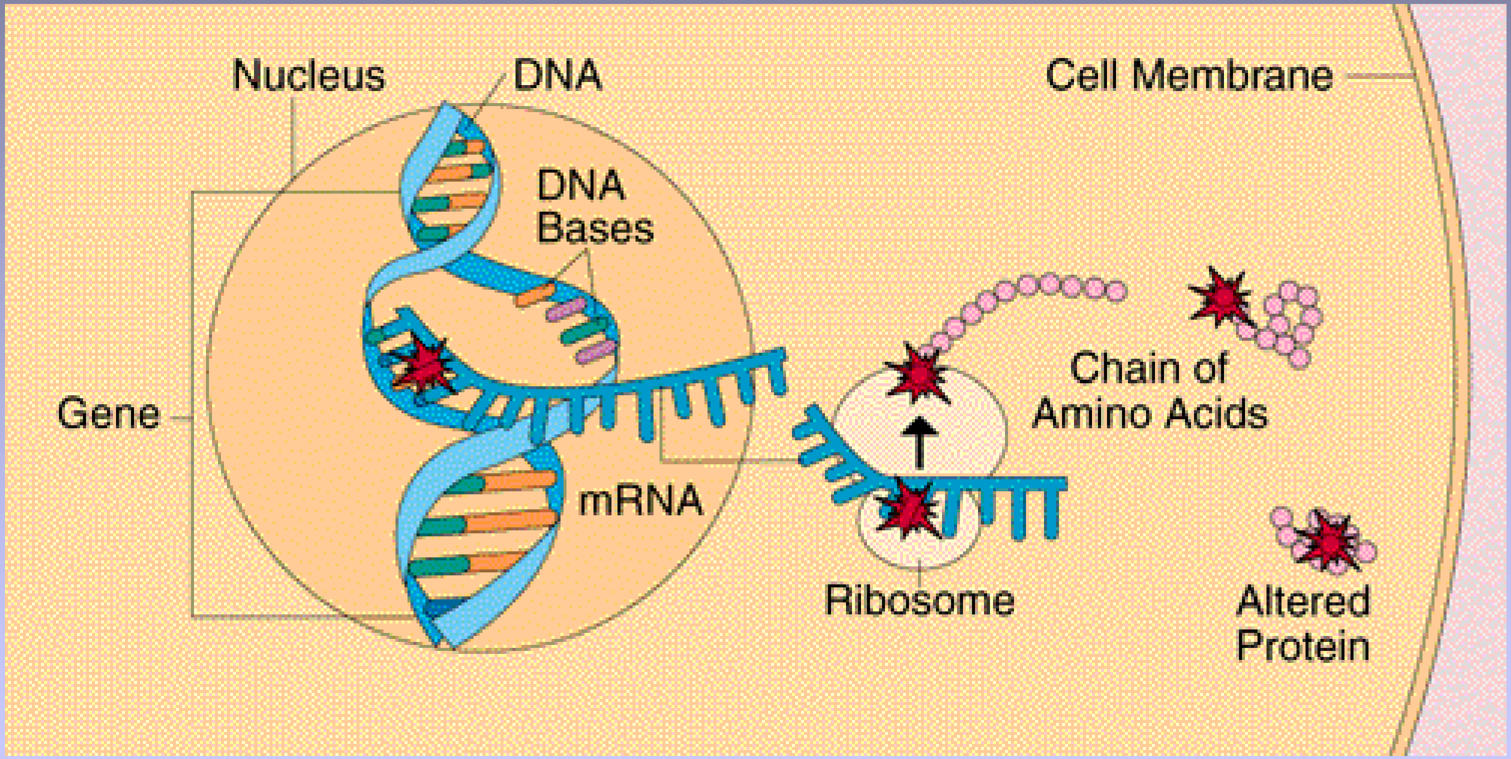


DNA

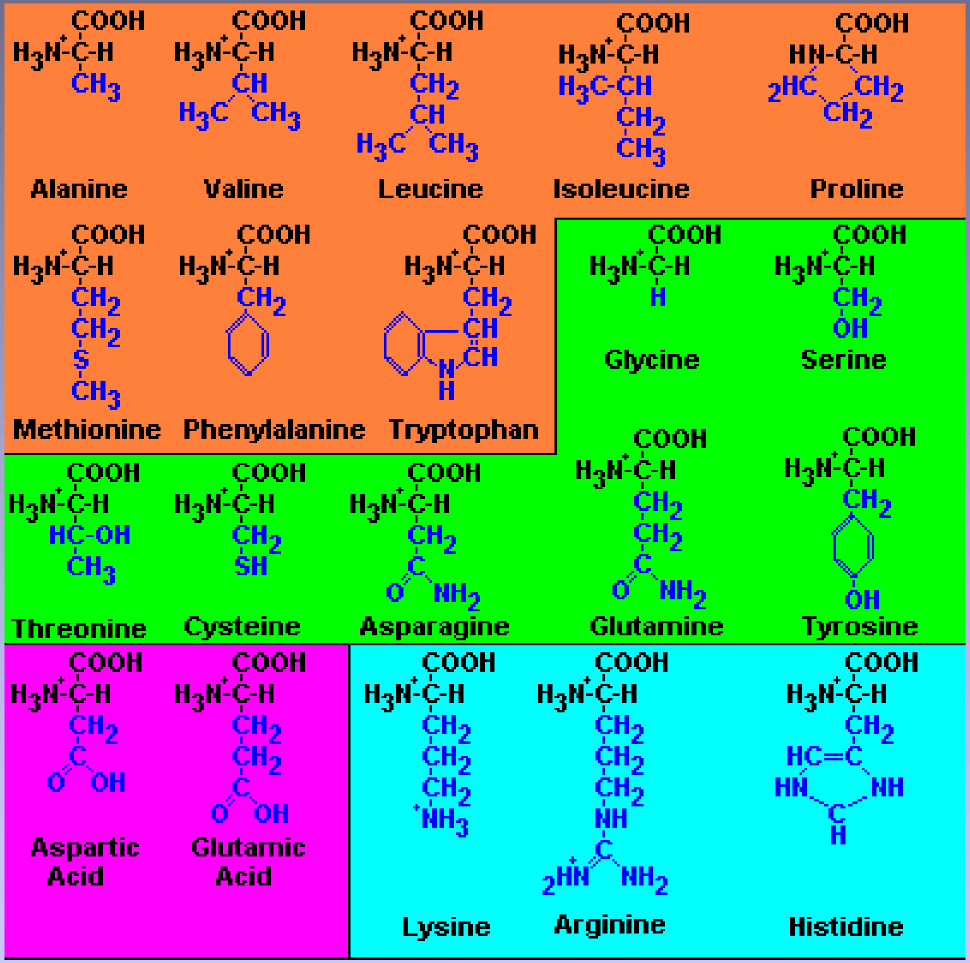
RNA

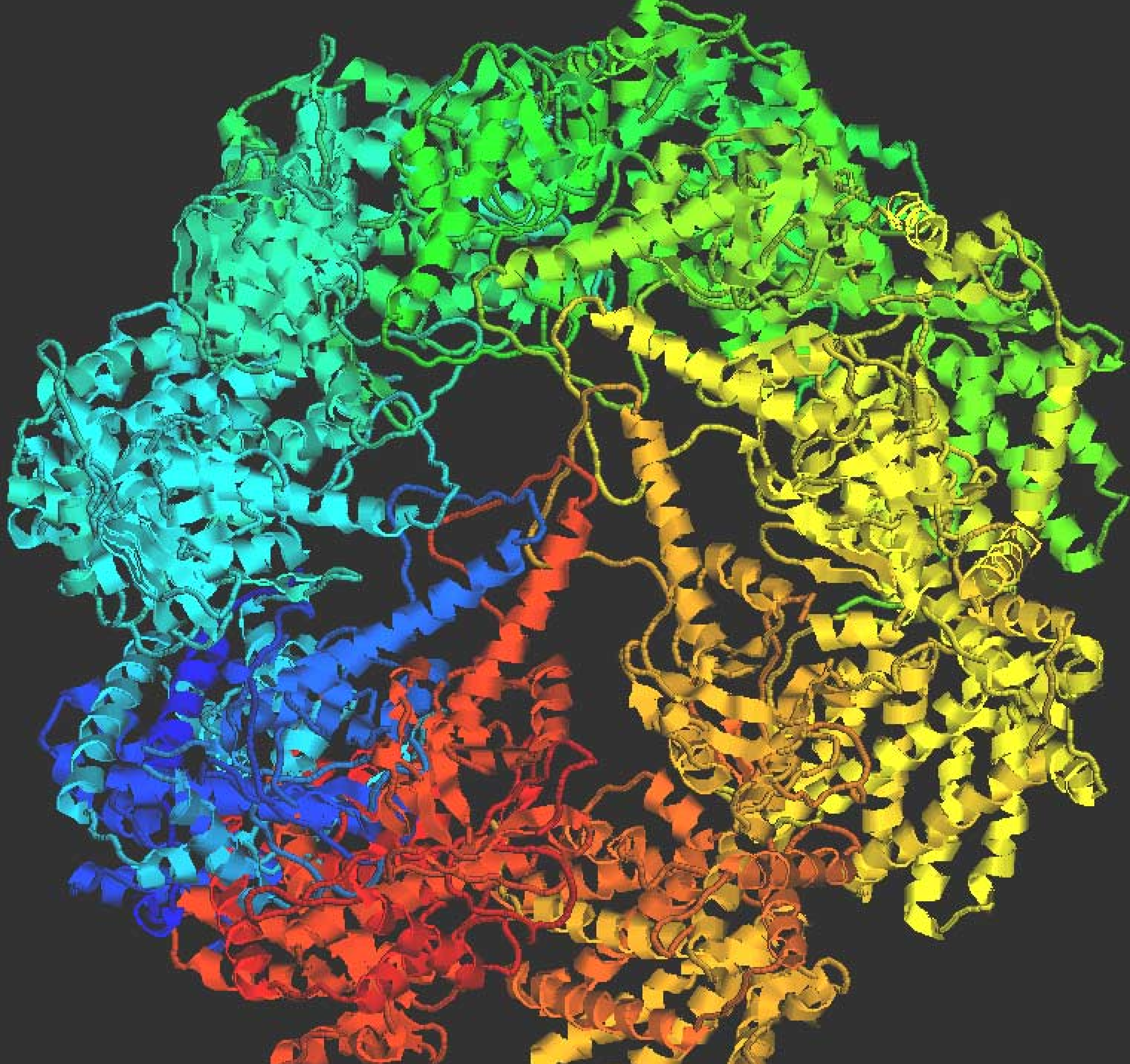
protein

Protein synthesis and disease

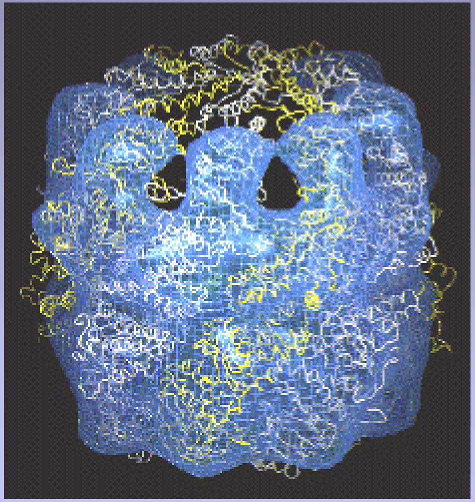


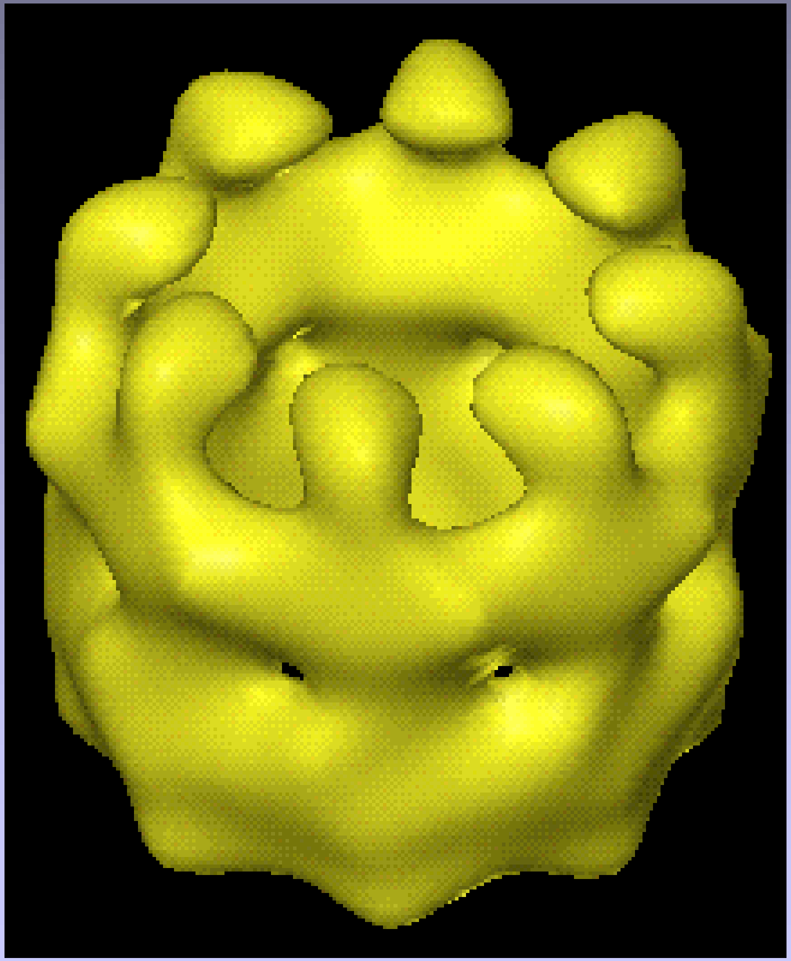
The beads of a protein chain



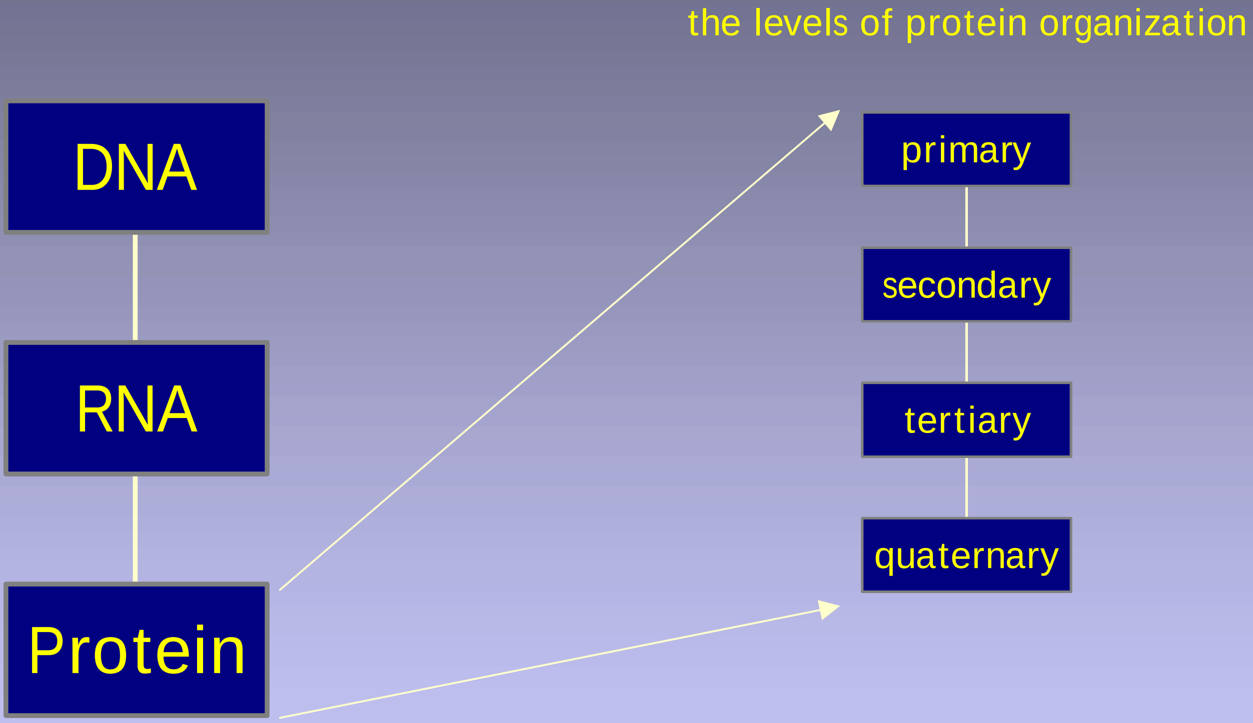


Cyro-electron micrograph of very large protein machine





the central dogma



the levels of protein organization

DNA

RNA

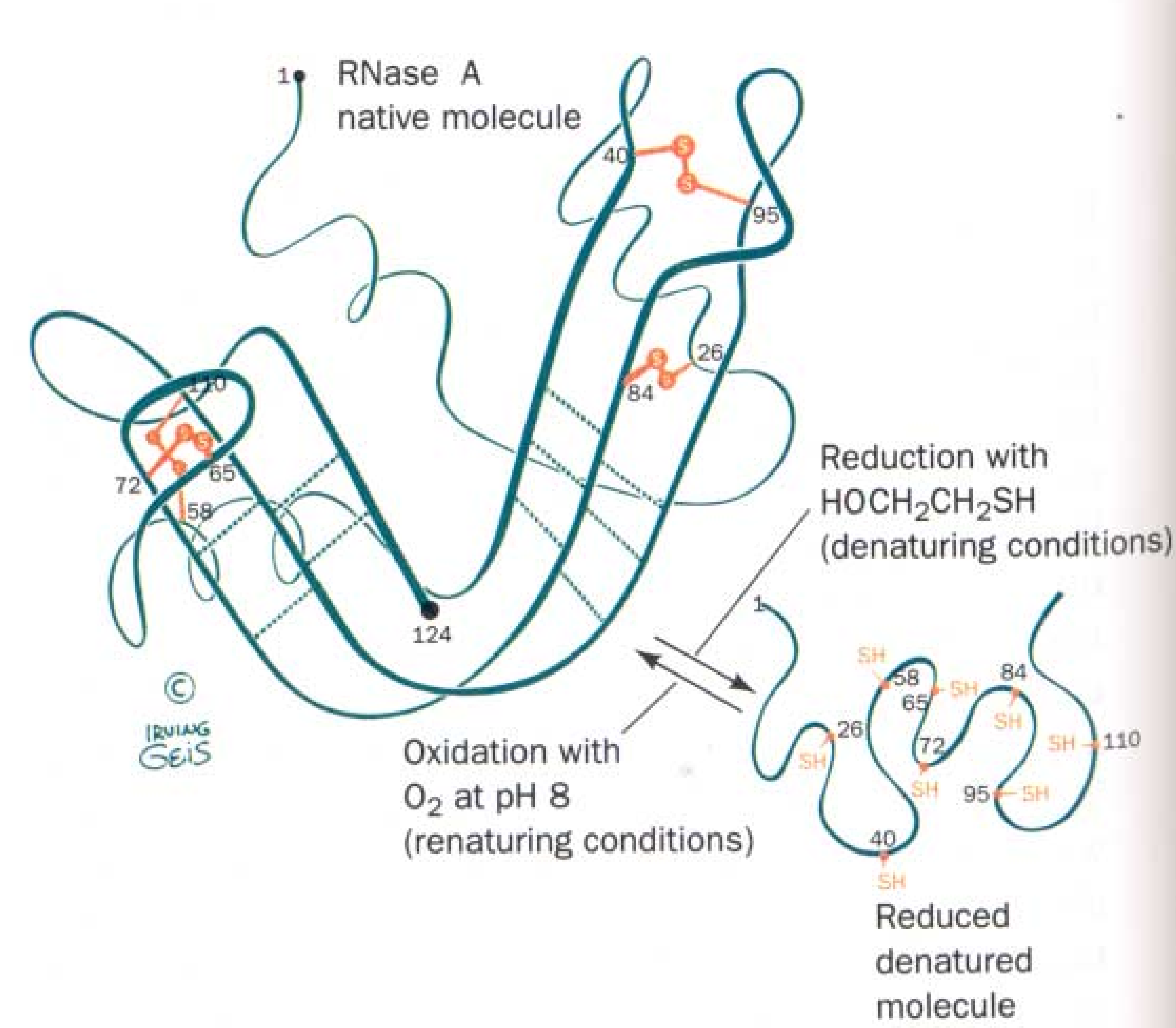
Protein

primary

secondary

tertiary

quaternary

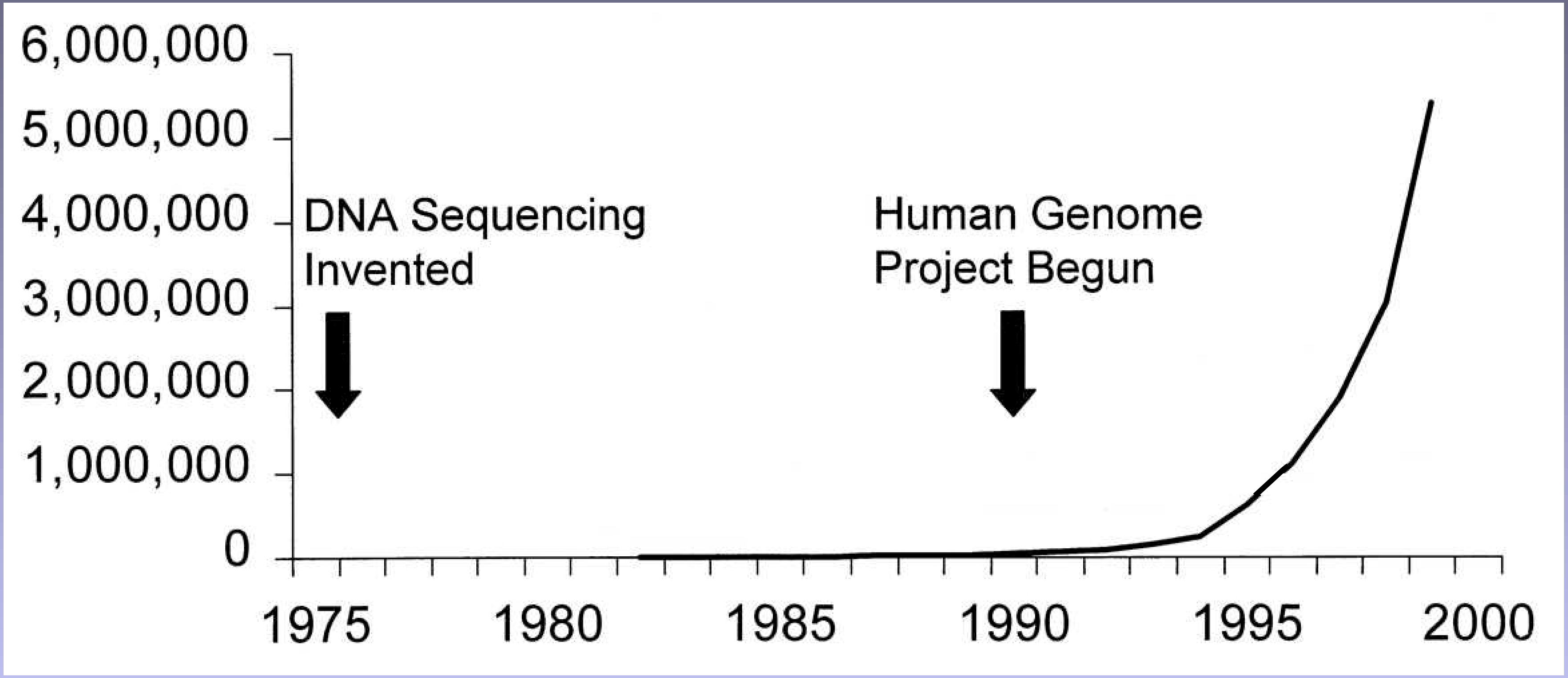

1
RNase A
native molecule
40
95
26
84
110
72
65
58
124
©
IRVING
GEIS
Reduction with
HOCH$_2$CH$_2$SH
(denaturing conditions)
Oxidation with
O$_2$ at pH 8
(renaturing conditions)
1
SH
58
65
SH
84
SH
26
SH
72
SH
SH
110
95
SH
40
SH
Reduced
denatured
molecule

The big three

- Sexual reproduction
- Consciousness
- Protein folding

The Human Genome Project

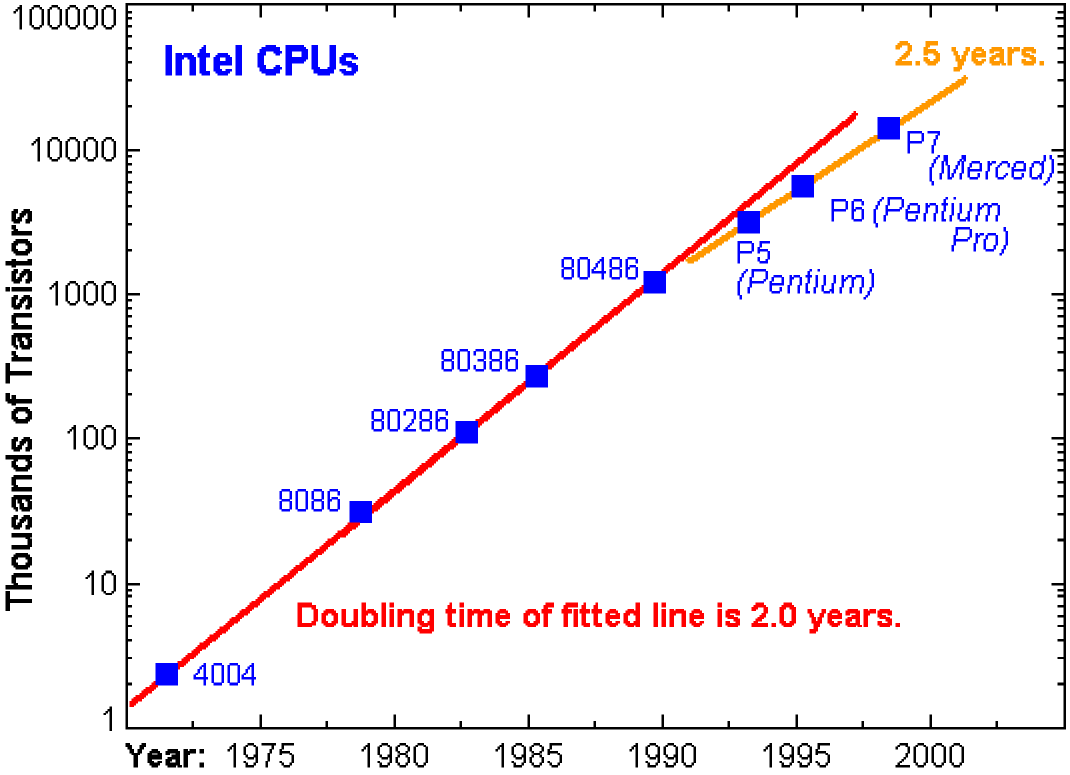
Growth in number of DNA sequences obtained



slightly modified from:

Boguski: *Science* (1999) vol. **286**, pp453-455

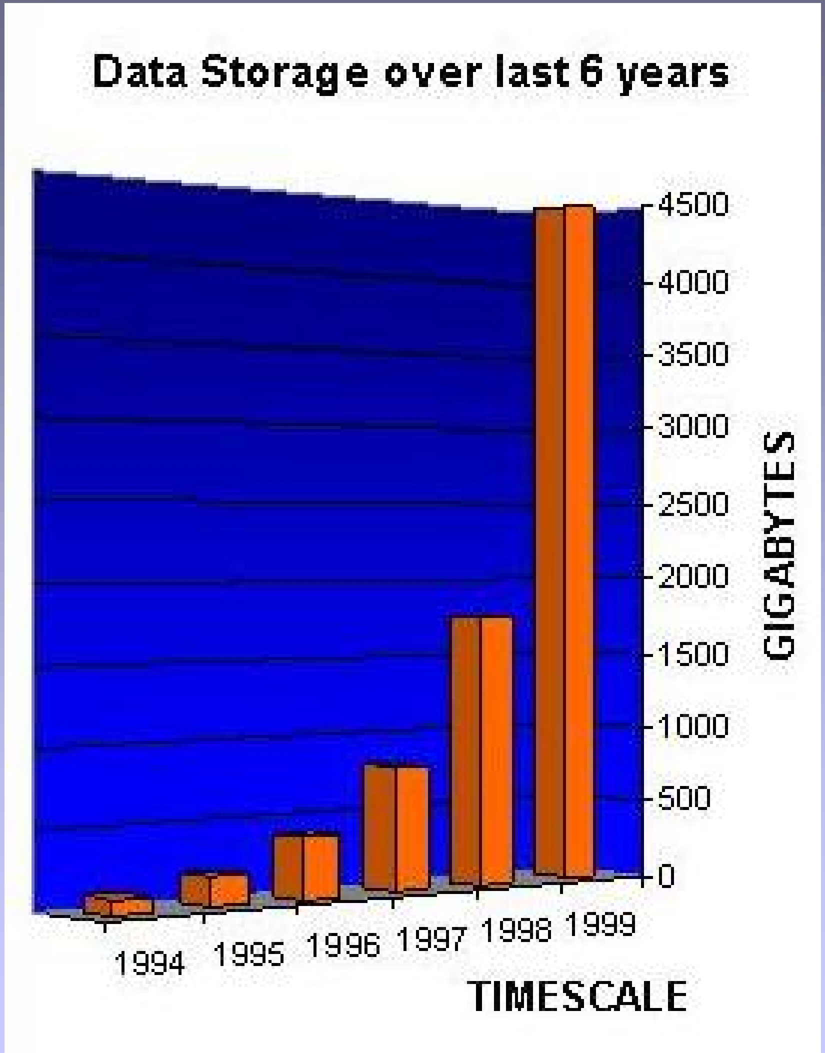
Moore’s Law as applied to Intel CPUs



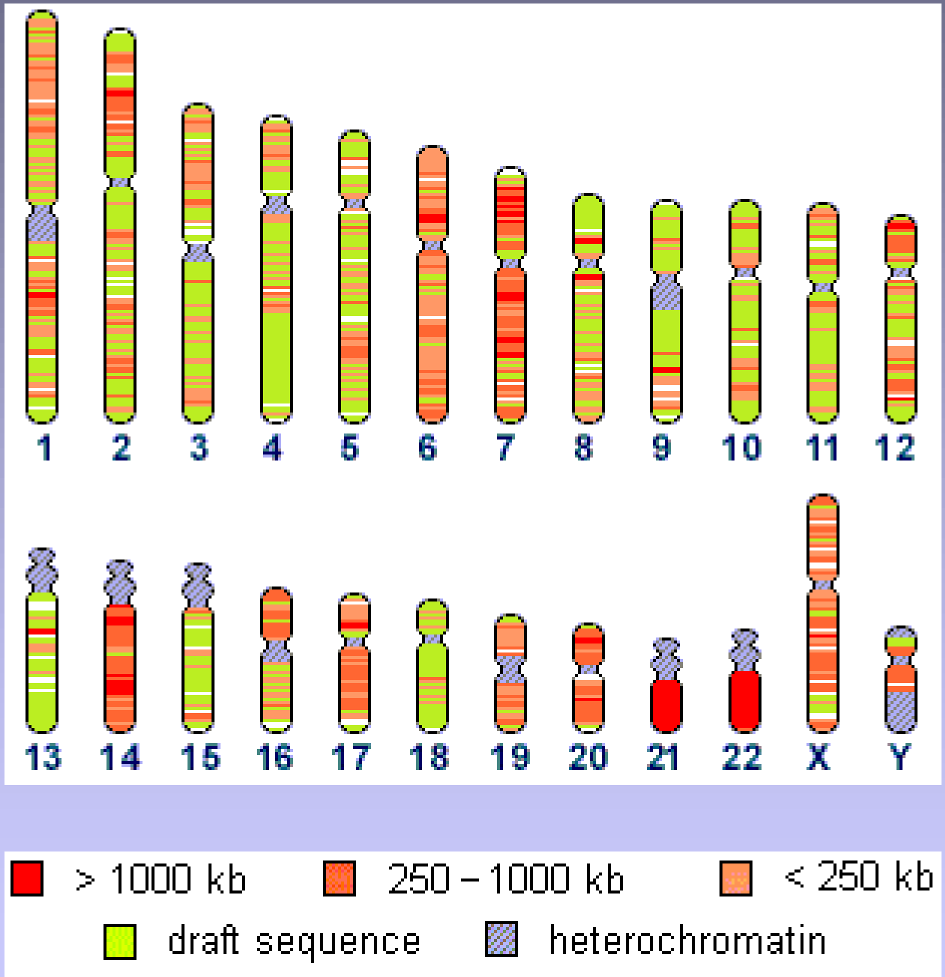
The Sanger Centre

A sequencing farm

Sanger Centre data storage



Progress in HGP



Structure of a gene

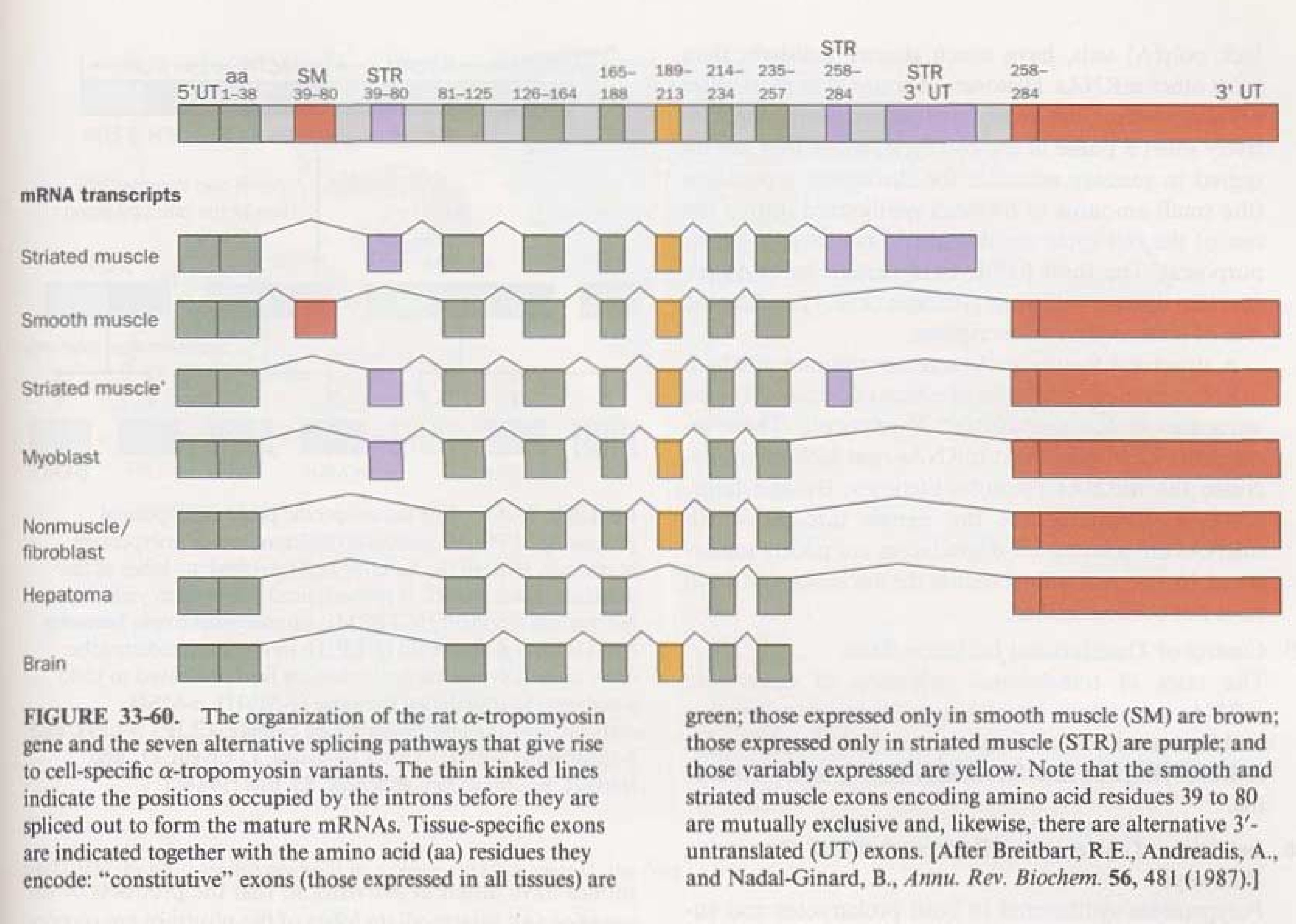
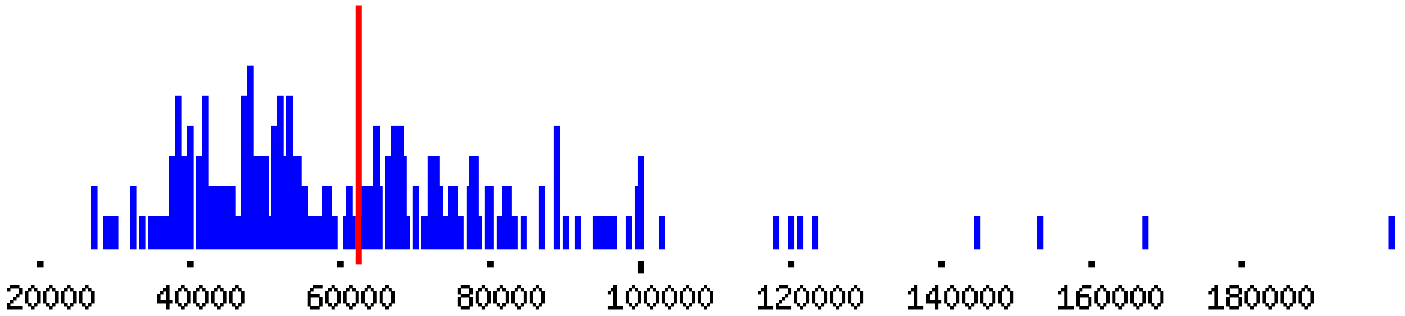
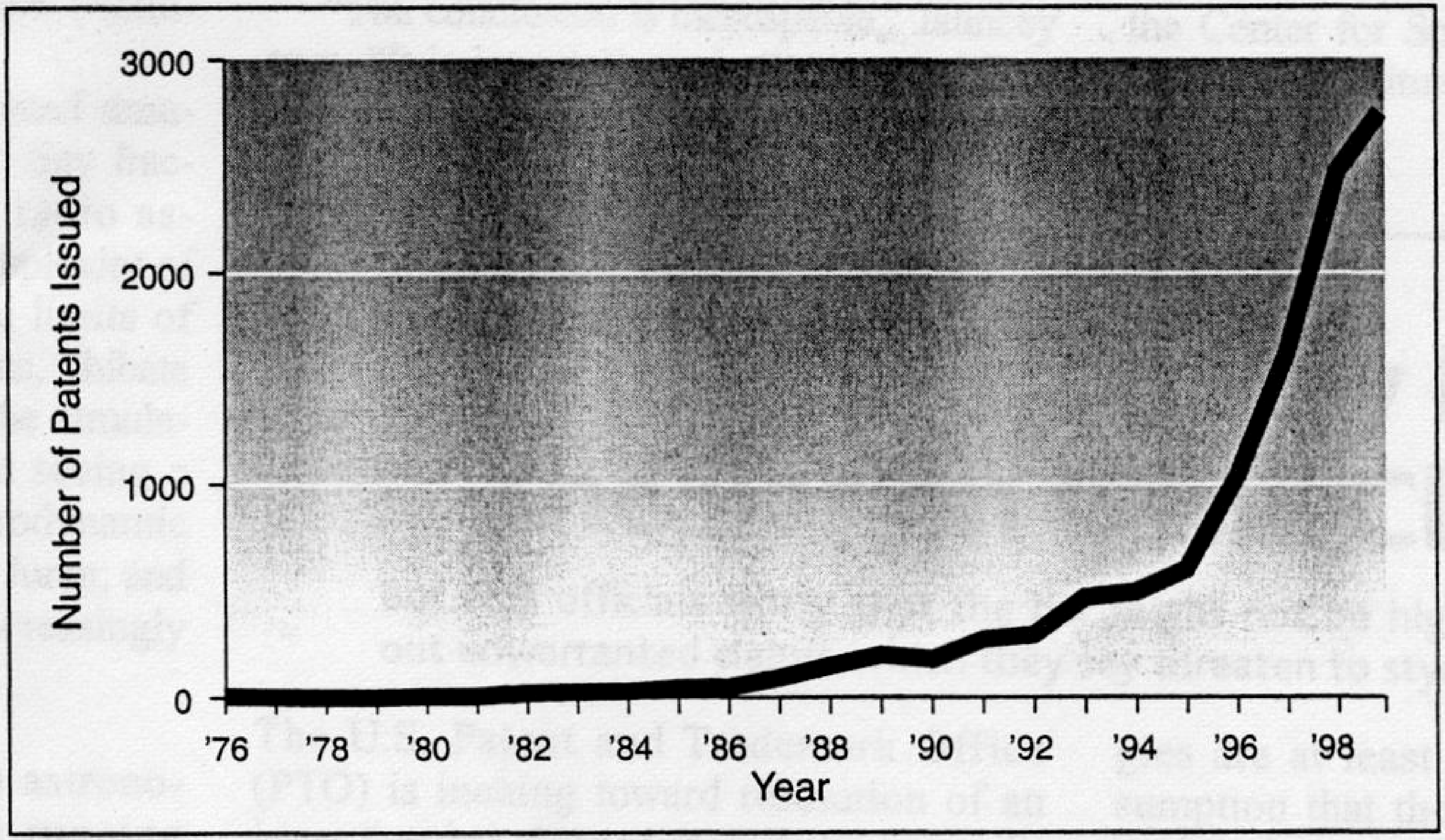


FIGURE 33-60. The organization of the rat α-tropomyosin gene and the seven alternative splicing pathways that give rise to cell-specific α-tropomyosin variants. The thin kinked lines indicate the positions occupied by the introns before they are spliced out to form the mature mRNAs. Tissue-specific exons are indicated together with the amino acid (aa) residues they encode: "constitutive" exons (those expressed in all tissues) are green; those expressed only in smooth muscle (SM) are brown; those expressed only in striated muscle (STR) are purple; and those variably expressed are yellow. Note that the smooth and striated muscle exons encoding amino acid residues 39 to 80 are mutually exclusive and, likewise, there are alternative 3′-untranslated (UT) exons. [After Breitbart, R.E., Andreadis, A., and Nadal-Ginard, B., *Annu. Rev. Biochem.* **56,** 481 (1987).]



GeneSweep entries

“Patenting genes”



Gene boom. The number of patent applications containing a genetic sequence has exploded over the past decade.

- Novelty
- Invention
- Utility
- not Excluded

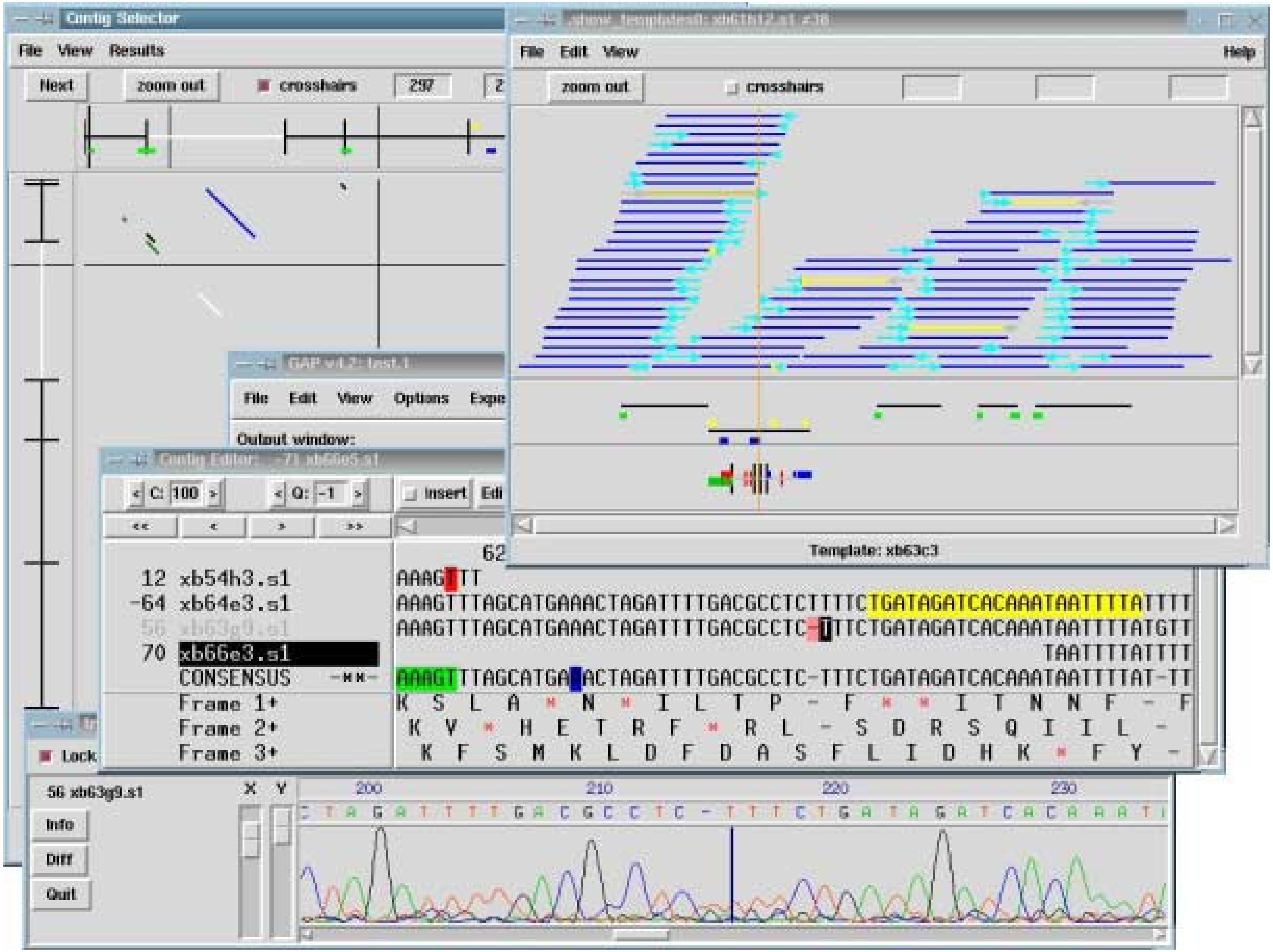
Differences in U.S.

- Year's grace
- "Utility" narrower

  e.g. U.S. requires clinical data
- Filing date

  U.S. "first to invent"

Open Source Bioinformatics

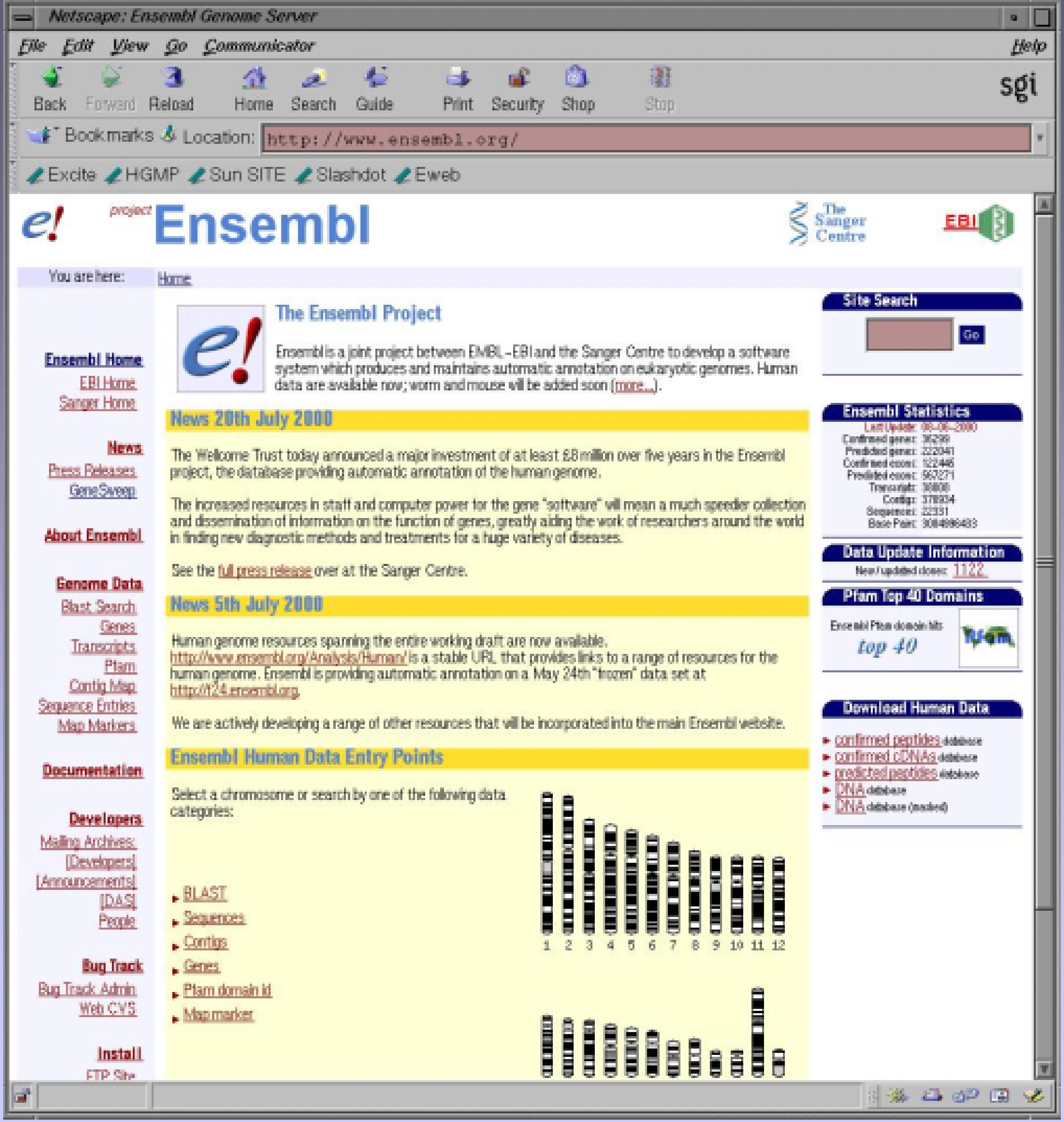


```
                                   62
 12 xb54h3.s1      AAAGTTT
-64 xb64e3.s1      AAAGTTTAGCATGAAACTAGATTTTGACGCCTCTTTTCTGATAGATCACAAATAATTTTATTTT
 56 xb63g9.s1      AAAGTTTAGCATGAAACTAGATTTTGACGCCTC-TTTCTGATAGATCACAAATAATTTTATGTT
 70 xb66e3.s1                                                                  TAATTTTATTTT
    CONSENSUS  -**- AAAGTTTAGCATGAAACTAGATTTTGACGCCTC-TTTCTGATAGATCACAAATAATTTTAT-TT
    Frame 1+       K  S  L  A  *  N  *  I  L  T  P  -  F  *  *  I  T  N  N  F  -  F
    Frame 2+        K  V  *  H  E  T  R  F  *  R  L  -  S  D  R  S  Q  I  I  L  -
    Frame 3+         K  F  S  M  K  L  D  F  D  A  S  F  L  I  D  H  K  *  F  Y  -
```

```
CTAGATTTTGACGCCTC-TTTCTGATAGATCACAAATI
```

Ensembl aims

- Find all protein-coding genes
- (guess at gene identity)
- distribute the data
- find additional features


Netscape: Ensembl Genome Server
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Shop Stop
Bookmarks Location:
Excite HGMP Sun SITE Slashdot Eweb

http://www.ensembl.org/

e! project Ensembl

You are here: Home

Ensembl Home
EBI Home
Sanger Home

News
Press Releases
GeneSweep

About Ensembl

Genome Data
Blast Search
Genes
Transcripts
Pfam
Contig Map
Sequence Entries
Map Markers

Documentation

Developers
Mailing Archives:
[Developers]
[Announcements]
[DAS]
People

Bug Track
Bug Track Admin
Web CVS

Install
FTP Site

The Ensembl Project

Ensembl is a joint project between EMBL-EBI and the Sanger Centre to develop a software system which produces and maintains automatic annotation on eukaryotic genomes. Human data are available now; worm and mouse will be added soon (more...).

News 20th July 2000

The Wellcome Trust today announced a major investment of at least £8 million over five years in the Ensembl project, the database providing automatic annotation of the human genome.

The increased resources in staff and computer power for the gene "software" will mean a much speedier collection and dissemination of information on the function of genes, greatly aiding the work of researchers around the world in finding new diagnostic methods and treatments for a huge variety of diseases.

See the full press release over at the Sanger Centre.

News 5th July 2000

Human genome resources spanning the entire working draft are now available. http://www.ensembl.org/Analysis/Human/ is a stable URL that provides links to a range of resources for the human genome. Ensembl is providing automatic annotation on a May 24th "frozen" data set at http://f24.ensembl.org.

We are actively developing a range of other resources that will be incorporated into the main Ensembl website.

Ensembl Human Data Entry Points

Select a chromosome or search by one of the following data categories:

- BLAST
- Sequences
- Contigs
- Genes
- Pfam domain id
- Map marker

Site Search

Go

Ensembl Statistics

Last Update: 08-08-2000

Confirmed genes:	36299
Predicted genes:	222041
Confirmed exons:	122446
Predicted exons:	967271
Transcripts:	38000
Contigs:	378934
Sequences:	22331
Base Pairs:	3084896433

Data Update Information

New/updated clones: 1122

Pfam Top 40 Domains

Ensembl Pfam domain hits

top 40

Download Human Data

- confirmed peptides database
- confirmed cDNAs database
- predicted peptides database
- DNA database
- DNA database (masked)

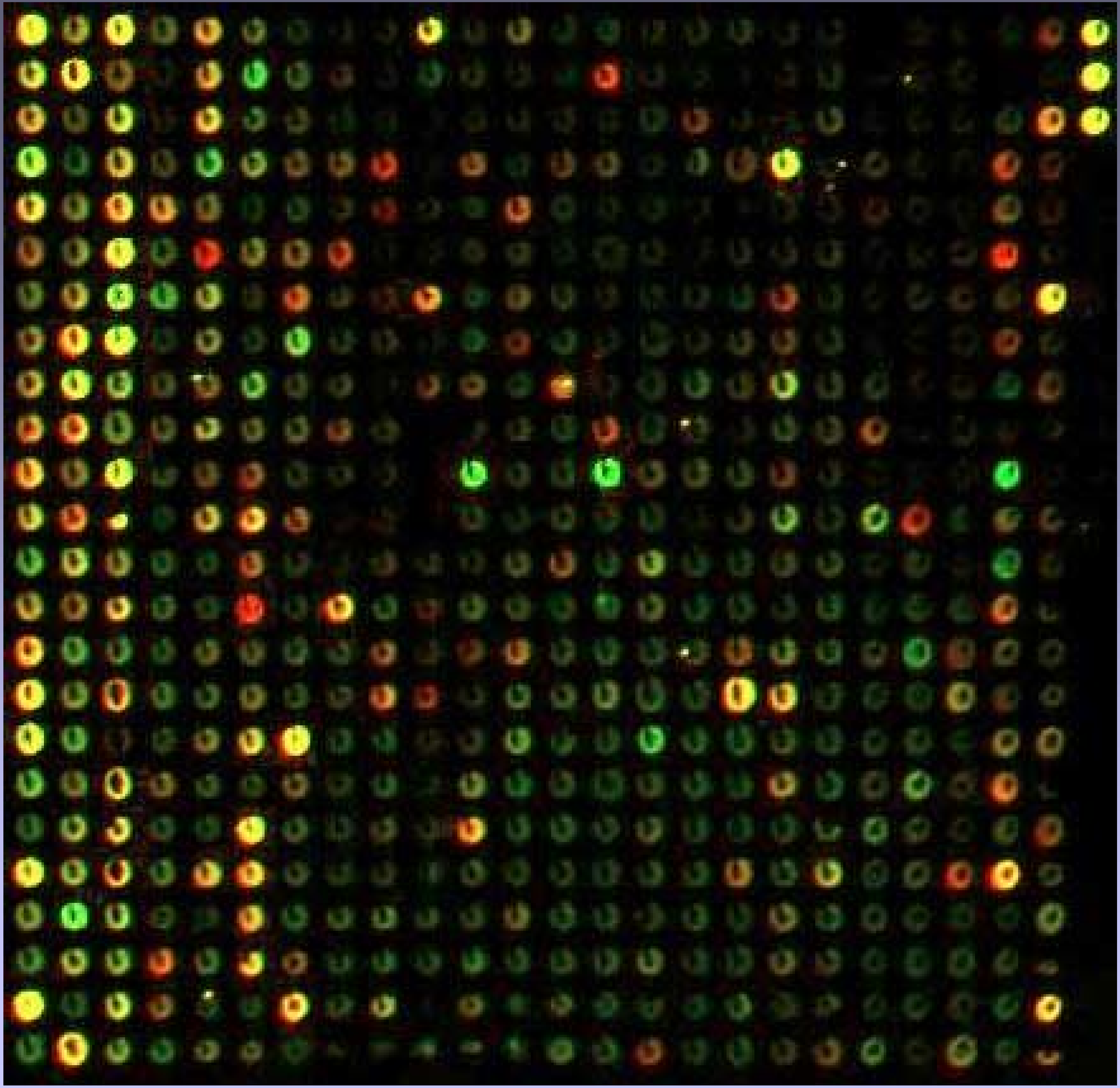
EMBOSS

- EMBOSS
- NUCLEUS

  bioinformatics-specific code
- AJAX

  data structures

The next wave



Acknowledgments

- Steve Roberts
- Keith Willison
- Malcom Herbert
- Gino Bellavia

The Institute of Cancer Research

the cancer research campaign