Bioinformatics

Introduction

Bioinformatics - lectures

- Introduction
- Information networks
- Protein information resources
- Genome information resources
- DNA sequence analysis
- Pairwise sequence alignment
- Multiple sequence alignment
- Secondary database searching
- Analysis packages
- Protein structure modelling

Introduction

- history of sequencing
- what is it Bioinformatics?
- sequence to structure deficit
- genome projects
- why is Bioinformatics important?
- patter recognition and prediction
- folding problem
- sequence analysis
- homo/analogy and ortho/paralogy

History of sequencing

- Protein sequencing
  - separation of peptides, identification and quantification of amino acids
  - Edman degradation
  - mass-spectrometry - advantage in identification of post-translational modifications

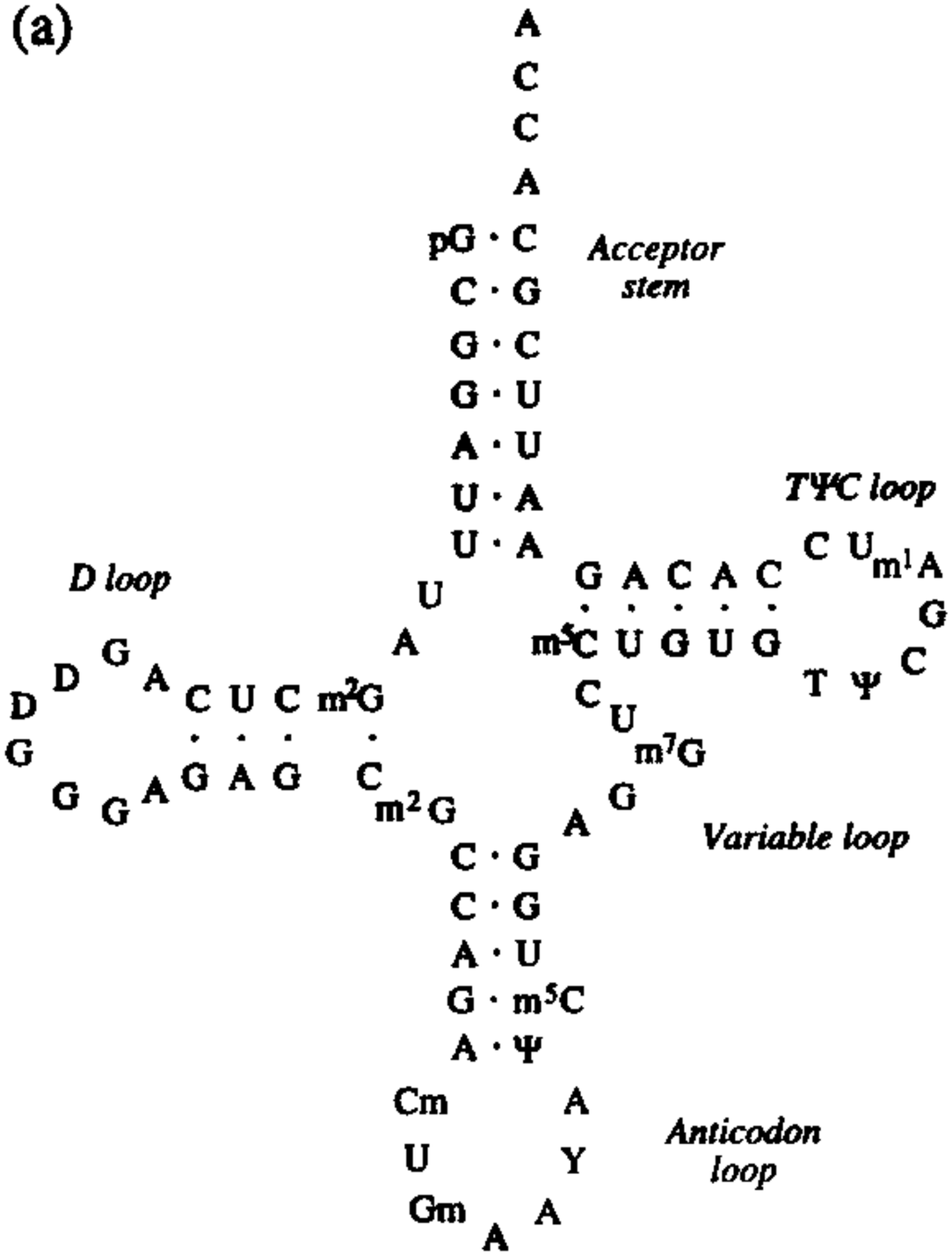
  - 1955 sequencing of peptide insuline
  - 1960 sequencing of enzyme ribonuclease
  - 1980s automated sequencers

History of sequencing

- Nucleic acid sequencing
  - tRNA - short, could be purified
  - DNA - large (human chromosome 55-250 x 10^6 bp); the longest fragment for sequencing is 500 bp; purification is problematic
  - advent of gene cloning and PCR
  - 1972 DNA cloning
  - 1975 DNA sequencing
  - 1980s and 1990s sequence revolution

	Technology development	Structure determination
1950	49 Edman degradation	
		51 α-helix model
	54 Isomorphous replacement	53 DNA double helix model Insulin primary structure
1960		60 Myoglobin tertiary structure
	62 Restriction enzyme	
		65 $tRNA^{Ala}$ primary structure
1970		
	72 DNA cloning	
		73 $tRNA^{Phe}$ tertiary structure
	75 DNA sequencing	
		77 $\phi X174$ complete genome
		79 Z-DNA by single crystal diffraction
1980		
	84 Pulse field gel electrophoresis 85 Polymerase chain reaction	
		86 Protein structure by 2D NMR
	87 YAC vector	
		88 Human Genome Project
1990		
	93 DNA chip	
		95 *H. influenzae* complete genome
2000		

(a)



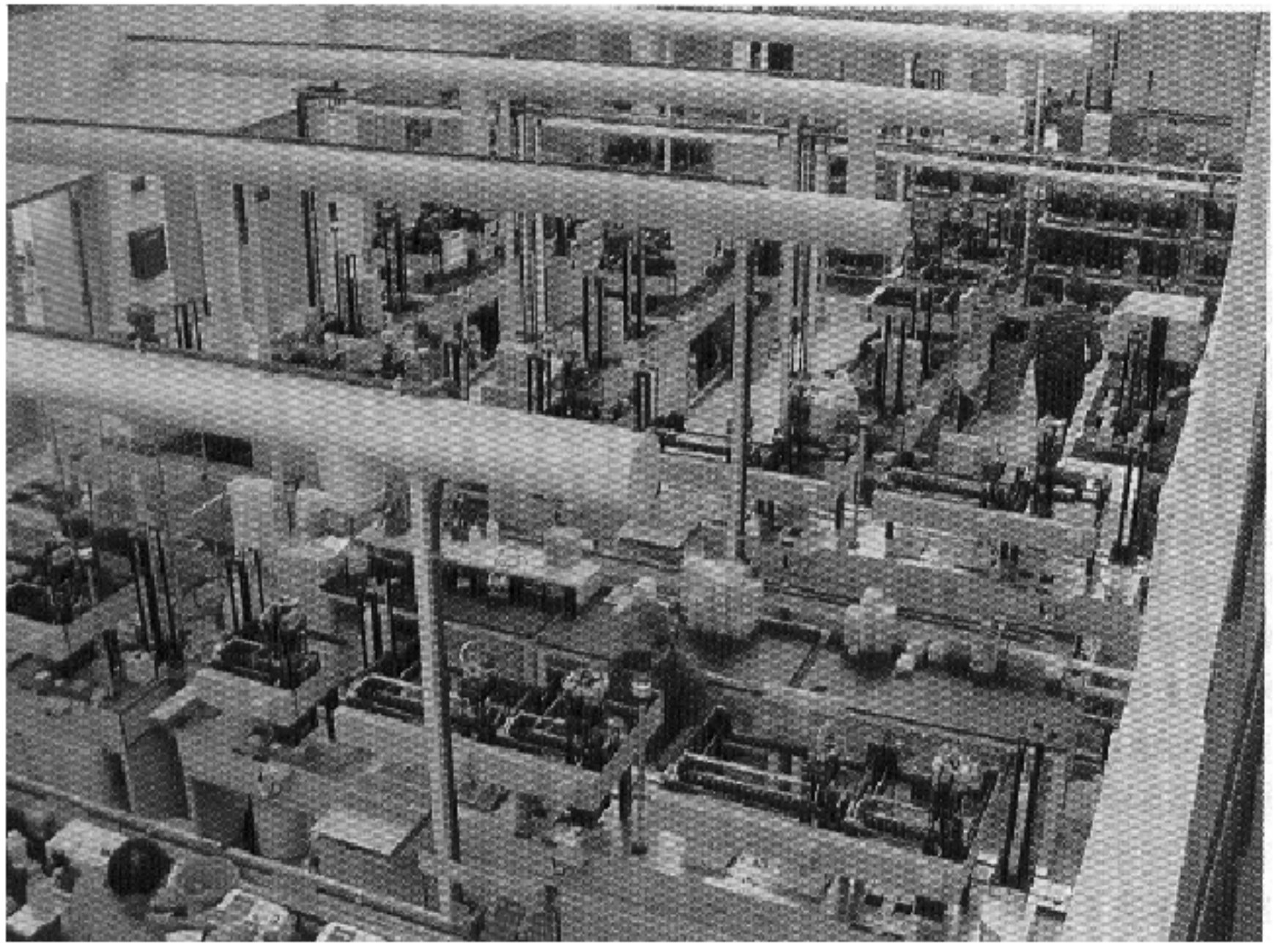
(b)

Fig. 1.8. Transfer RNA. (a) The primary sequence and the secondary structure of yeast alanyl-transfer RNA. (b) The tertiary structure of yeast phenylalanyl-transfer RNA (PDB:1TRA).

Automatic sequencing machine

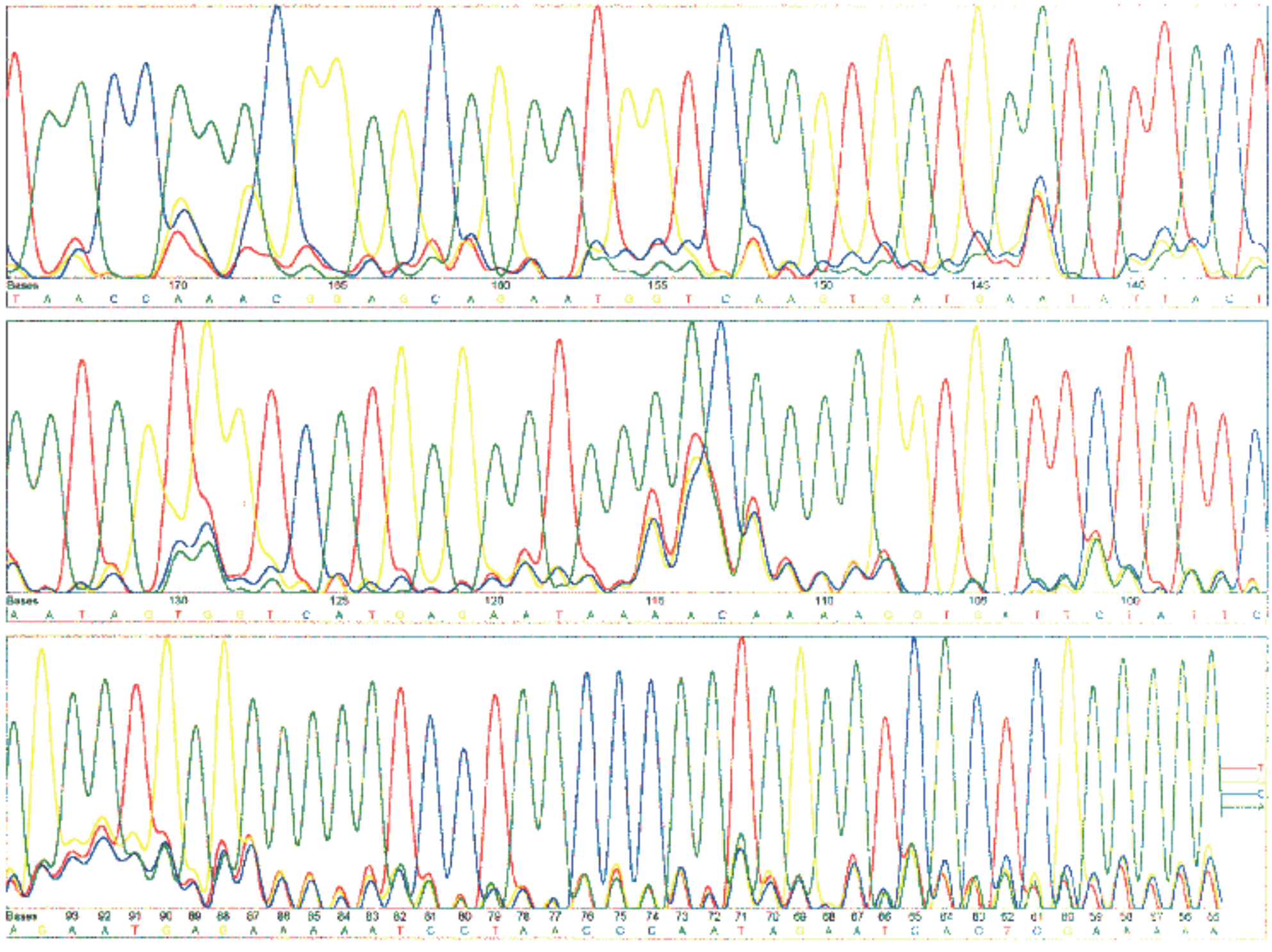
ABI Prism 310, Applied Biosystems

Automated production line in sequencing "factory"



Whitehead Institute, Center for Genome Research, USA

Sequencing chromatogram



What is Bioinformatics?

- improvements in DNA sequencing technologies and computer-based technologies

- originally - analysis of sequence data (1980s)
- presently - also analysis of 3D-structures

- The term bioinformatics is used to encompass almost all computer applications in biological sciences.

- Information technology applied to the management and analysis of biological data.

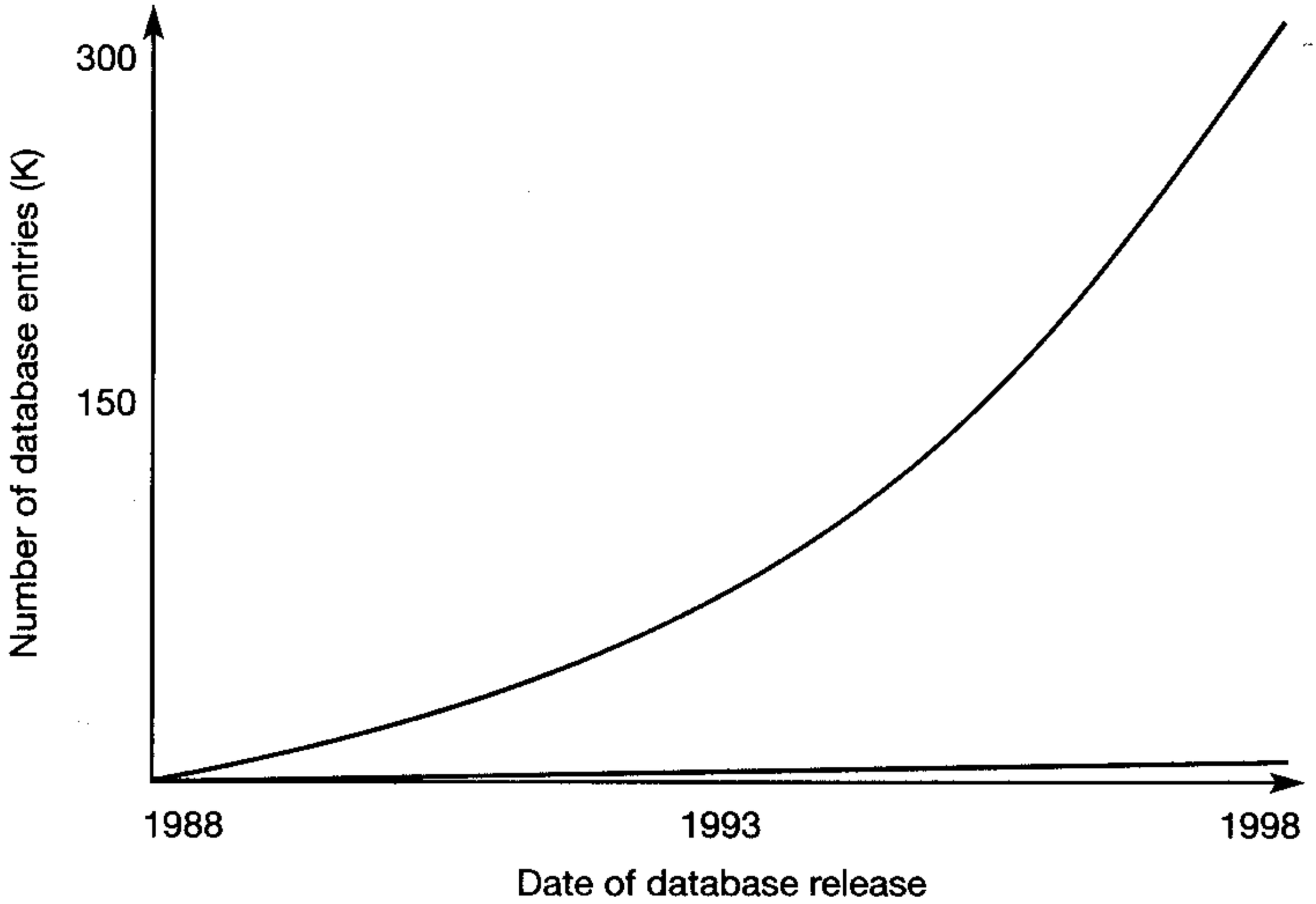


Figure 1.1 The protein sequence/structure deficit in 1998. The graph illustrates the non-redundant growth of sequence data during the last decade (—) and the corresponding growth in the number of unique structures (—).

Genome projects

- 1977 first complete genome - virus ϕX174, 5000 nucleotides; 11 genes
- 1995 first complete genome of living organism *Haemophilus influenzae,* 1.8 million nucleotides and 1700 genes
- sequencing of model systems: *Escherichia coli, Saccharomyces cerevisiae, Cernorhabditis elegans, Drosophila melanogaster, Arabidopsis thaliana, Canis familiaris, Mus Musculus*

(Bits) (Nucleotides)

2^{32}

10^{9}

- Human
- Mouse
- Rice
- Fruit fly
- Nematode (1998)
- *Arabidopsis*
- Budding yeast (1997)
- *Escherichia coli* (1997)
- *Haemophilus influenzae* (1995)
- Cytomegalovirus (1990)
- λ phage (1982)
- ϕX174 phage (1977)

2^{24}

10^{7}

2^{16}

10^{5}

2^{8}

10^{3}

- GenBank 10/97
- GenBank 9/92
- GenBank 9/87
- GenBank 10/82

	Genome size (Mb)	Gene number	Haploid chromosome number
Bacterium (*Escherichia coli*)	~4	4,403	1
Yeast (*Saccharomyces cerevisiae*)	~12	6,190	16
Worm (*Caenorhabditis elegans*)	97	19,730	6
Fruit Fly (*Drosophila melanogaster*)	120	13,601	4
Mouse (*Mus Musculus*)	3,454	~50,000 (estimated)	20
Human (*Homo sapiens*)	2,910	33,609	23

Human Genome Project

- in mid-1980s initiated Human Genome Project
- estimated 100.000 genes and completion in 2005
- need for automated sequencing and improved computational techniques
- shotgun method

- sequencing of rough draft first
- first draft completed in 2000 by publicly funded the International Consortium Human Genome Project and the company Celera Genomics

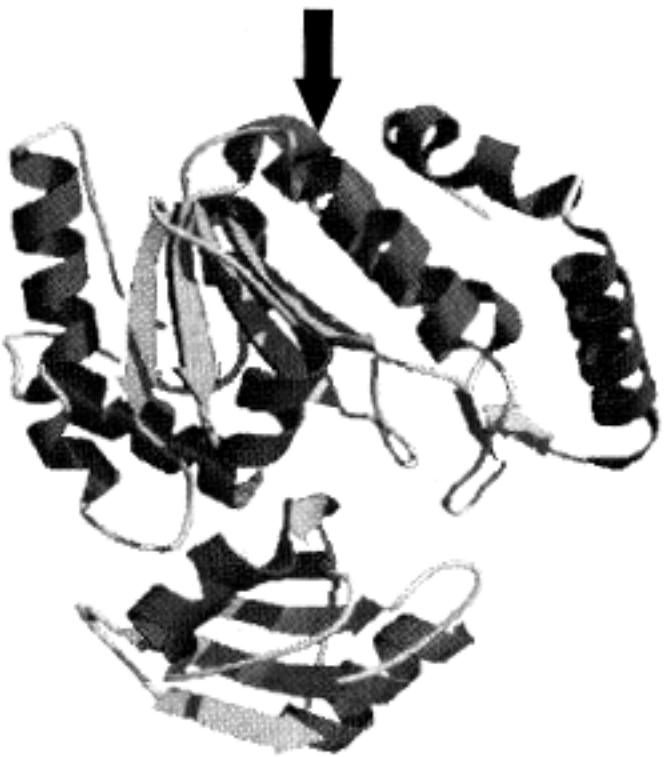
Human Genome Project

- ~33.000 genes
- genes are complex due to alternative splicing
- >1.000.000 proteins (estimated)
- hundreds of genes resulted from horizontal transfer from bacteria (in vertebrate lineage)
- dozen of genes derived from transposable elements (their activity however has declined)
- the mutation rate in male is two-times higher than in female
- >1.400.000 single point polymorphisms (SNPs)

Why is bioinformatics important?

- last 20-30 years - structural biology
- new era - bioinformatics - due to genome projects and sequence/structure deficit
- biological function is not known for about 50% of all genes in every sequenced genome
- role of bioinformatics
  - data management and storage
  - data analysis = conversion of primary sequence to biological knowledge

T M I T D S L A V V L Q R R D W E N P G
V T Q L N R L A A H P P F A S W R N S E
E A R T D R P S Q Q L R S L N G E W R F
A W F P A P E A V P E S W L E C D L P E
A D T V V V P S N W Q M H G Y D A P I Y
T N V T Y P I T V N P P F V P T E N P T
G C Y S L T F N V D E S W L Q E G Q T R
I I F D G V N S A F H L W C N G R W V G
Y G Q D S R L P S E F D L S A F L R A G
E N R L A V M V L R W S D G S Y L E D Q
D M W R M S G I F R D V S L L H K P T T
Q I S D F H V A T R F N D D F S R A V L



Primary structure:	the linear sequence of amino acids in a protein molecule
Secondary structure:	regions of local regularity within a protein fold (e.g., α-helices, β-turns, β-strands)
Super-secondary structure:	the arrangement of α-helices and/or β-strands into discrete folding units (e.g., β-barrels, βαβ-units, Greek keys, etc.)
Tertiary structure:	the overall fold of a protein sequence, formed by the packing of its secondary and/or super-secondary structure elements
Quaternary structure:	the arrangement of separate protein chains in a protein molecule with more than one subunit
Quinternary structure:	the arrangement of separate molecules, such as in protein–protein or protein–nucleic acid interactions

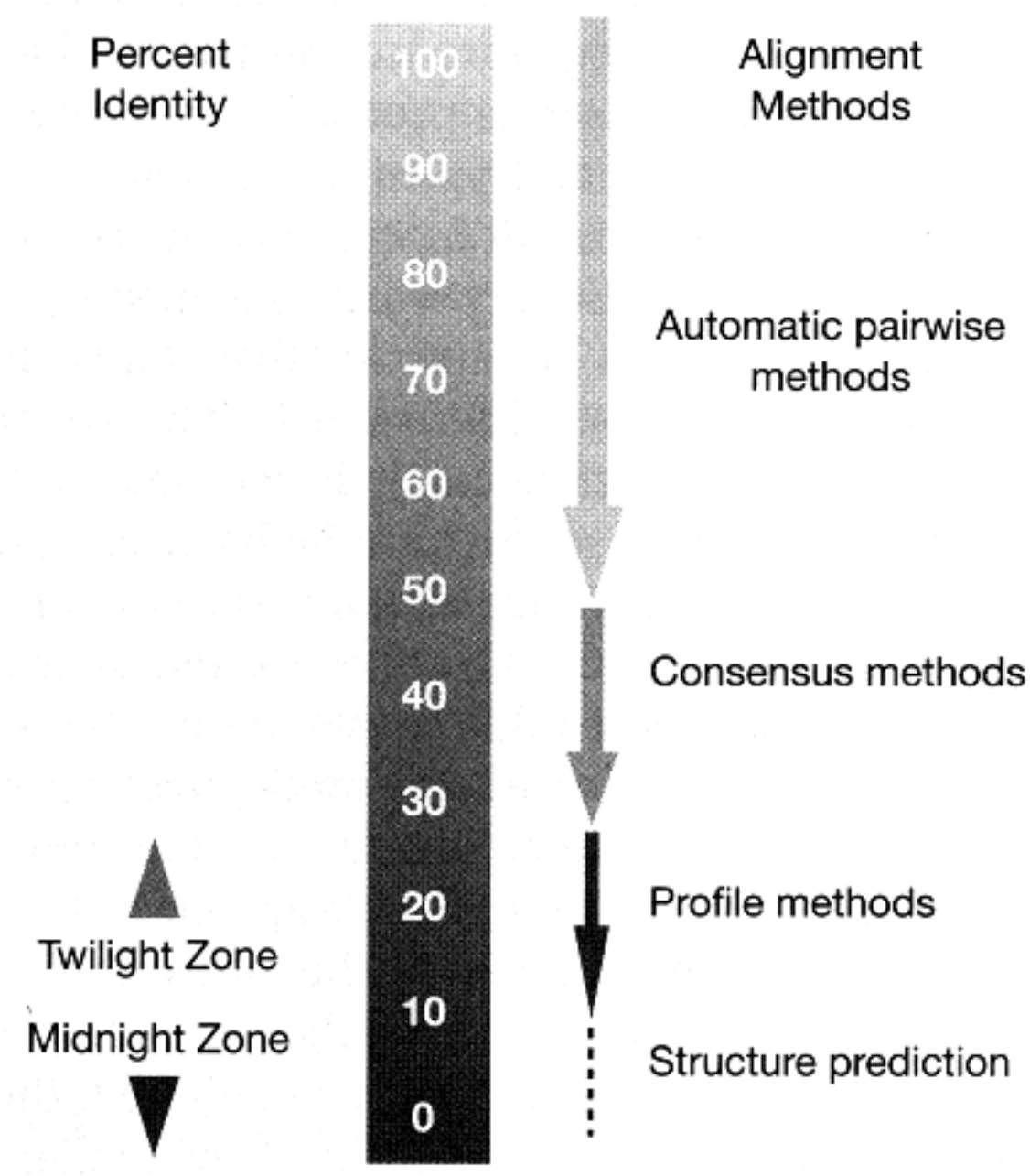
Homology and analogy

- Sequences are said to be homologous if they are related by divergence from a common ancestor.

- Proteins can share similar folds (e.g., β-barrel) or similar catalytic residues (e.g., serine proteases) without any sequential similarity. Convergence to similar biological solutions from different evolutionary starting points results in analogy.

- Sequence analysis assumes homologous proteins.

- Homology is not a measure of similarity.


Percent Identity
100
90
80
70
60
50
40
30
20
10
0
Alignment Methods
Automatic pairwise methods
Consensus methods
Profile methods
Structure prediction
Twilight Zone
Midnight Zone

Orthology and paralogy

- Proteins performing the same function in different species - orthologues.

- Proteins performing different, but related functions within same organism - paralogues.

- Sequence comparison of orthologuos proteins - phylogenetic analysis.

Self-opening umbrella