Bioinformatics Chapter 1.

Introduction

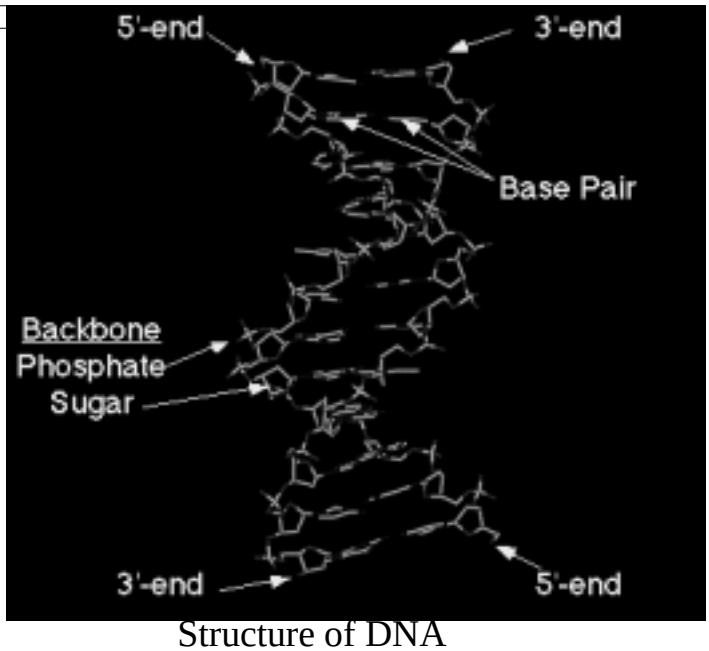
Outline

- Biological Data in Digital Symbol Sequences
- Genomes – Diversity, Size, and Structure
- Proteins and Proteomes
- On the Information Content of Biological Sequences
- Prediction of Molecular Function and Structure

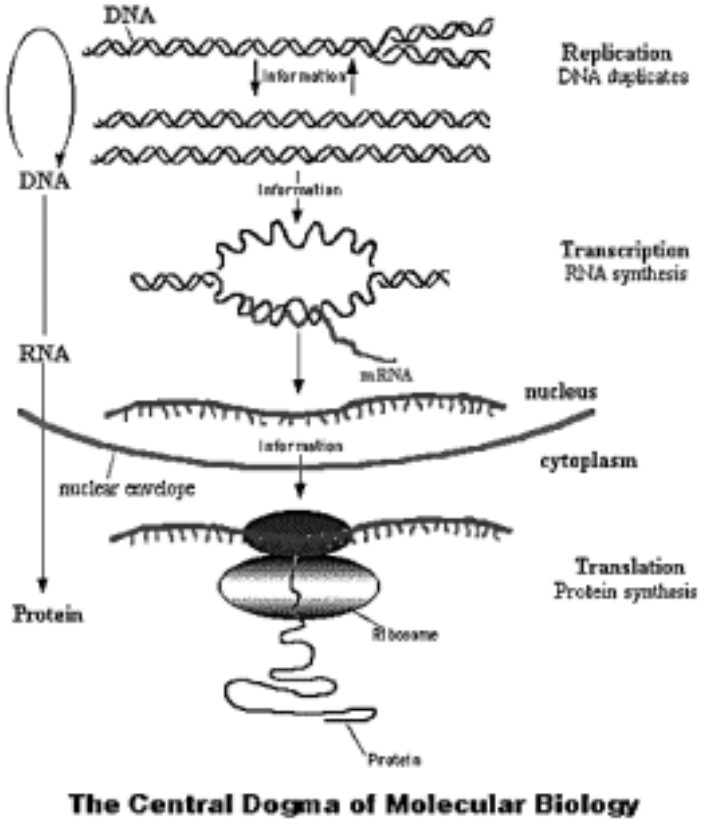
- A liger is the result of breeding a male lion with a female tiger. It has stripes and spots.
- All ligers are presumed to be born sterile like mule. This is not unusual for hybrids.
- Extravagant combination of genomes: this organism is unable to contribute to further to the evolution of the gene pool.

1. Biological Data in Digital Symbol Sequence

- Digital nature of genetic data/biological sequence
- DNA sequence:
  4 nucleotides (A, T, G, C)
- Protein sequence:
  20 amino acids

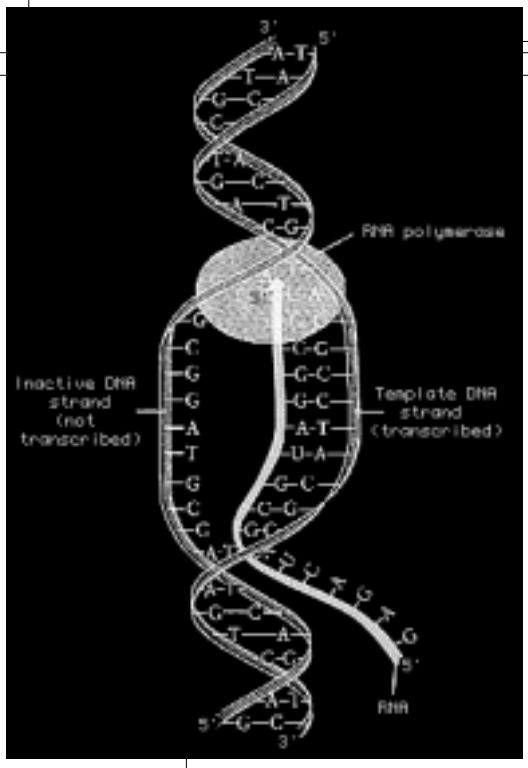


Structure of DNA

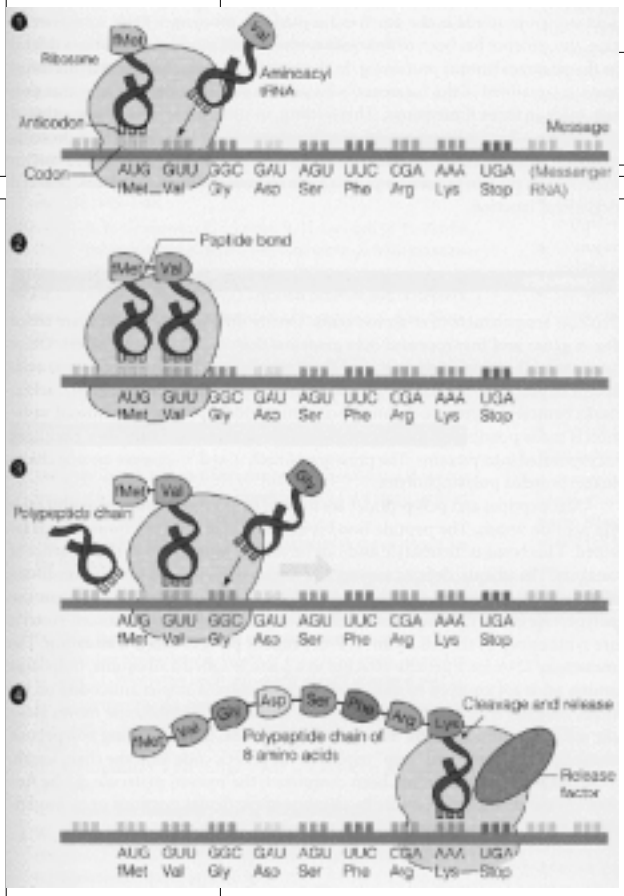


The Central Dogma of Molecular Biology

- Central Dogma
  - Information flow from DNA to protein
  - Proteins are synthesized based on the information of DNA
  - DNA: information storage
  - RNA: information intermediate
  - Protein: various cellular functions

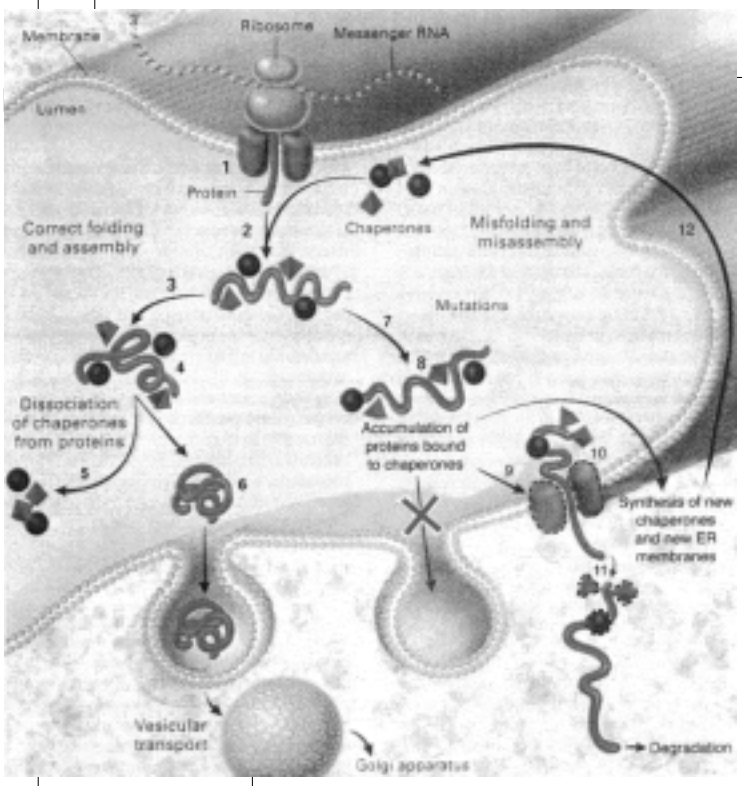


- RNA transcription
  - Newly synthesized RNAs have complementary base sequences to template DNAs. (A – U, C – G)
  - Three major RNA classes:
  - mRNA: protein coding
  - tRNA: at translation process, brings amino acids to ribosomes
  - rRNA: ribosome component
  - This transcription process is performed by RNA polymerase.



- Translation: synthesis of proteins
  - Ribosome, tRNA and other components carry out this process.
  - Codon (mRNA): anticodon (tRNA) recognition
  - Codon triplelets give more diverse compositions

  (4 nucleotide compositions = give different 64 codons)



- Post translational protein folding
  - Proteins exist not as linear form but as complex 3D form.
  - Proteins are folded by many molecular interactions (hydrophobic interactions, disulfied bond, ionic bond…)
  - Various molecules like chaperone help protein folding.
  - This protein conformation construction is determined by information on its sequence.

Why Probabilistic Models?

- This course focuses on the probabilistic models of sequences. Why?
- While the goal is to study a particular sequence and its molecular structure and function, the analysis typically proceeds through the study of an ensemble of sequences consisting of its different versions in different species or different versions in the same species.
- Thus, comparison of sequence patterns across species must take into account that biological sequences are inherently noisy, the variability resulting in part from random events amplified by evolution.

1.1 Database Annotation Quality

- High error rate after handling of information
  - Handling by a diverse group of people
  - Incorrect assignments by experimentalists: simple consistency check
  - Features are indicated by listing the relevant positions in numeric form.
  - Fuzzy statements in annotation: comments like “potential” or “probable”
- Bioinformaticians
  - Should consider potential sources of errors when creating machine-learning approaches for prediction and classification.
  - Prepare the data carefully
  - Discard data from unclear sources

1.1 Database Annotation Quality

- Machine learning techniques
  - can handle noise if large corpora of sequences are available.
- Learning from DNA sequences as language acquisition
  - Infants can detect linguistic regularities and learn simple statistics for the recognition of word boundaries
  - Learning techniques are useful for revealing similar regularities ("grammar" or rules) in genomic data ("sentences" or examples).

1.2 Database Redundancy

- Redundancy: sources of error
  - Data for large families of closely related sequences
    → biased and over-represent features
  - Apparent correlations between different positions in the sequences may be artifact of biased sampling of the data.
  - Predictive performance may be overestimated if training/test data are closed related.
- Over-representation problem
  - It is necessary to avoid too closely related sequences in a data set.
  - Keep all sequences in a data set and assign weights according to their novelty
  - Redundancy reduction
    - arbitrary similarity threshold
    - "representative" data set

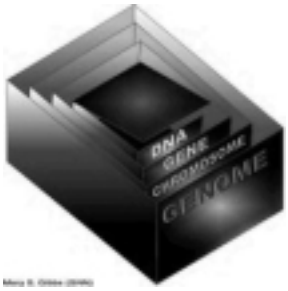
2. Genomes – Diversity, Size, and Structure

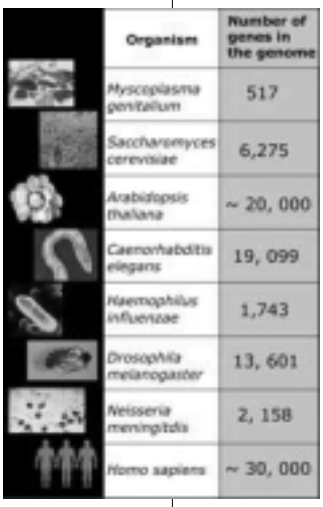
- Genomes: total genetic material within a cell or organism

  DNA (e.g. cellular organism) or RNA (e.g. HIV virus)

- Genome size: 5,386bp (bacteriophage φX174)

  ~ 3Gbp (*homo sapiens*)



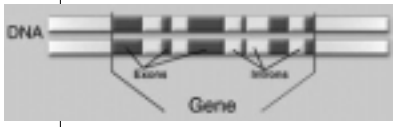


Organism	Number of genes in the genome
Myscoplasma genitalium	517
Saccharomyces cerevisiae	6,275
Arabidopsis thaliana	~ 20, 000
Caenorhabditis elegans	19, 099
Haemophilus influenzae	1,743
Drosophila melanogaster	13, 601
Neisseria meningitidis	2, 158
Homo sapiens	~ 30, 000

• Gene numbers in several organisms

2. Genomes – Diversity, Size, and Structure

- Genome size does not directly relate to cell complexity
  - Many plant and amphibian genomes are larger than the human genome.
  - Cell complexity is related to gene numbers or gene-to-gene interactions.
  - Most regions of genome are not expressive units.



Gene: expressive segment - coding proteins or functional RNA molecules

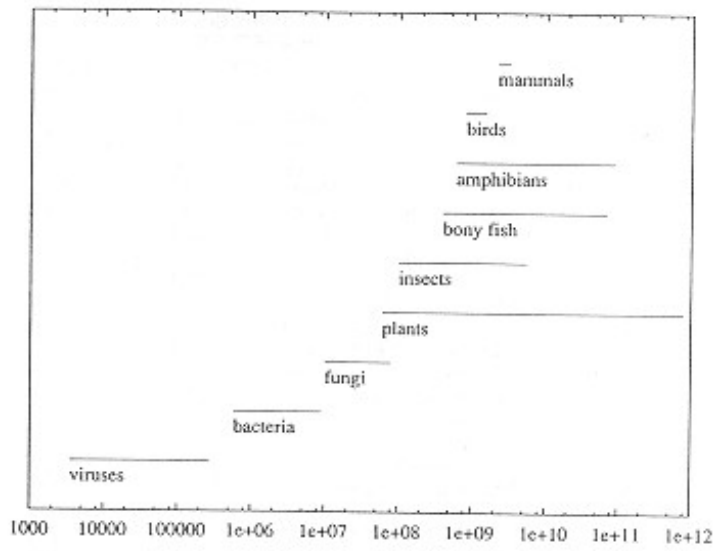


Figure 1.1: Intervals of Genome Sizes for Various Classes of Organisms. Note that the plot is logarithmic in the number of nucleotides on the first axis. Most commonly, the variation within one group is one order of magnitude or more. The narrow interval of genome sizes among mammals is an exception to the general picture. It is tempting to view the second axis as "organism complexity", but it is most certainly not a direct indication of the size of the gene pool. Many organisms in the upper part of the spectrum, e.g., mammals, fish, and plants, have comparable numbers of genes (see Table 1.1).

Group	Species	Genes	Genome size
Phages	Bacteriophage MS2	4	0.003569
	Bacteriophage T4	270	0.168899
Viruses	Cauliflower mosaic virus	8	0.008016
	HIV type 2	9	0.009671
	Vaccinia virus	260	0.191737
Bacteria	*Mycoplasma genitalium*	473	0.58
	Mycoplasma pneumoniae	716	0.82
	Haemophilus influenzae	1,760	1.83
	Bacillus subtilis	3,700	4.2
	Escherichia coli	4,100	4.7
	Myxococcus xanthus	8,000	9.45
Archaea	*Methanococcus jannaschii*	1,735	1.74
Fungi	*Saccharomyces cerevisiae*	5,800	12.1
Protoctista	*Cyanidioschyzon merolae*	5,000	11.7
	Oxytricha similis	12,000	600
Arthropoda	*Drosophila melanogaster*	12,000	165
Nematoda	*Caenorhabditis elegans*	14,000	100
Mollusca	*Loligo pealii*	>35,000	2,700
Plantae	*Nicotiana tabacum*	43,000	4,500
	Arabidopsis thaliana	25,000	70–145
Chordata	*Giona intestinalis*	N	165
	Fugu rubripes	70,000	400
	Danio rerio	N	1,900
	Mus musculus	70,000	3,000
	Homo sapiens	70,000	3,000

Table 1.1: Approximate Gene Number and Genome Sizes in Organisms in Different Evolutionary Lineages. Genome sizes are given in megabases. N = not available. Data were in part taken from [292] and references therein; others were compiled from a number of different Internet resources, papers, and books.

Species	Haploid genome size	Bases	Entries
Homo sapiens	3,000,000,000	502,036,490	960,074
Mus musculus	3,000,000,000	120,306,608	256,900
Caenorhabditis elegans	100,000,000	108,081,367	75,887
Arabidopsis thaliana	100,000,000	40,392,224	48,473
Saccharomyces cerevisiae	12,067,280	28,511,474	10,457
Drosophila melanogaster	165,000,000	26,701,988	25,489
Escherichia coli	4,639,221	17,231,610	4,681
Rattus norvegicus	3,000,000,000	13,516,612	10,307
Oryza sativa	400,000,000	8,868,147	22,218
HIV type 1	9,750	8,591,298	19,005
Fugu rubripes	400,000,000	7,385,060	15,051
Schizosaccharomyces pombe	14,000,000	6,361,177	1,185
Gallus gallus	1,200,000,000	4,993,215	4,338
Bacillus subtilis	4,170,000	4,799,275	1,015
Mycobacterium tuberculosi	4,397,000	4,349,989	559
Toxoplasma gondii	89,000,000	4,347,115	10,817
Brugia malayi	100,000,000	4,118,477	10,826
Bos taurus	3,000,000,000	4,068,817	4,609
Synechocystis sp.	3,573,470	3,826,212	146
Xenopus laevis	3,000,000,000	3,078,555	2,024

Table 1.2: The Number of Bases in GenBank rel. 103, October 1997, for the 20 Most Sequenced Organisms. For some organisms there is far more sequence than the size of the genome, due to strain variation and pure redundancy.

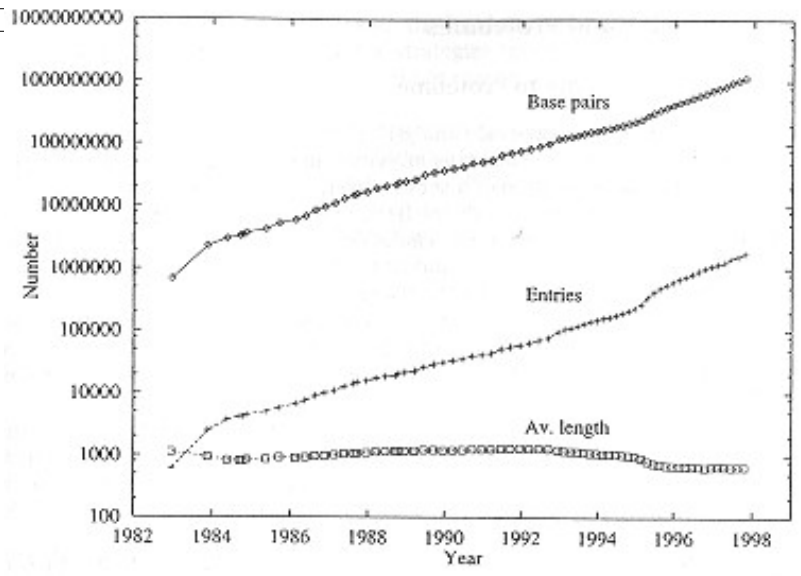


Figure 1.2: The Exponential Growth in the Size of the GenBank Database in the Period 1983-1997. Based on the development in the last year, the doubling time is around 15 months. The complete size of GenBank rel. 103 is 1,160,300,687 nucleotides in 1,765,847 entries (average length 657). Currently the database grows by more than 1,700,000 bases per day.

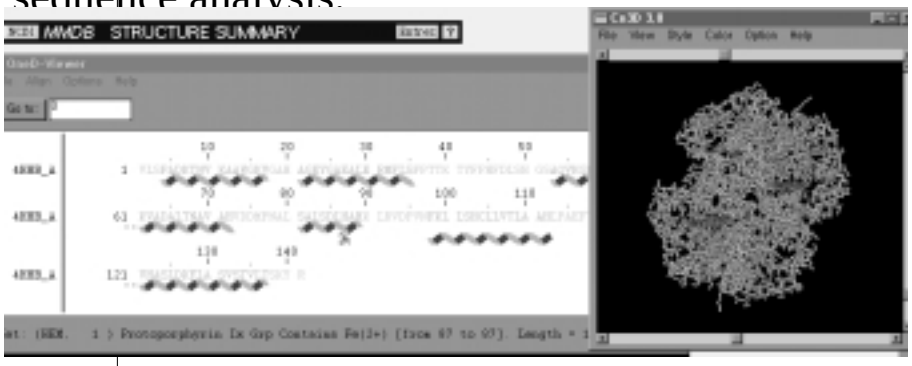
3. Proteins and Proteomes

3.1 From Genome to Proteome

- Gene : genome = protein : proteome
- Proteome: total protein expression of a set of chromosomes
- Protein: dynamic structure with various modification
- Proteome analysis:
  - Not only dealing with sequences and functions but also concerning biochemical states of proteins in its posttranslational form

3.2 Protein Length Distributions

- Amino acid sequences in protein direct its structure which in turn influences its functions.
- To make a suitable structure, regularity of amino acid sequences does exist.
- Statistical analysis has played a major role in protein sequence analysis.



Sequence of hemoglobin and its structure

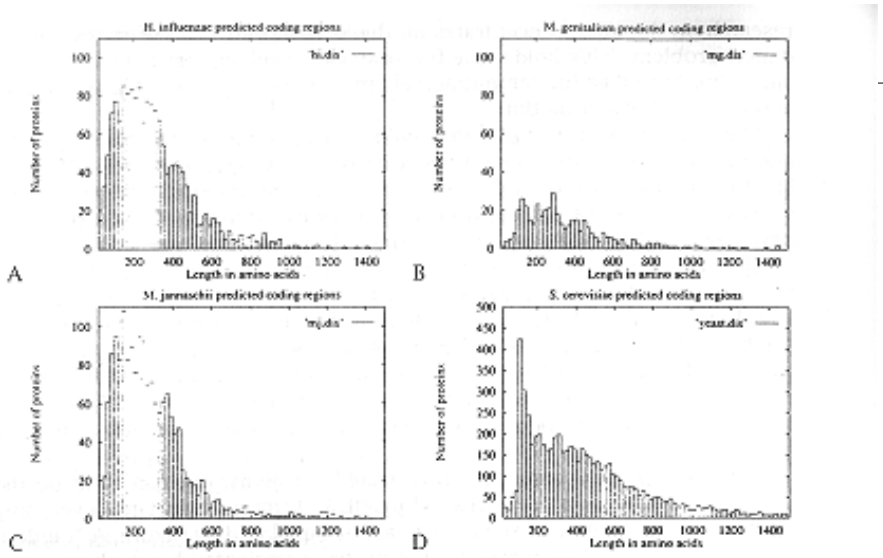
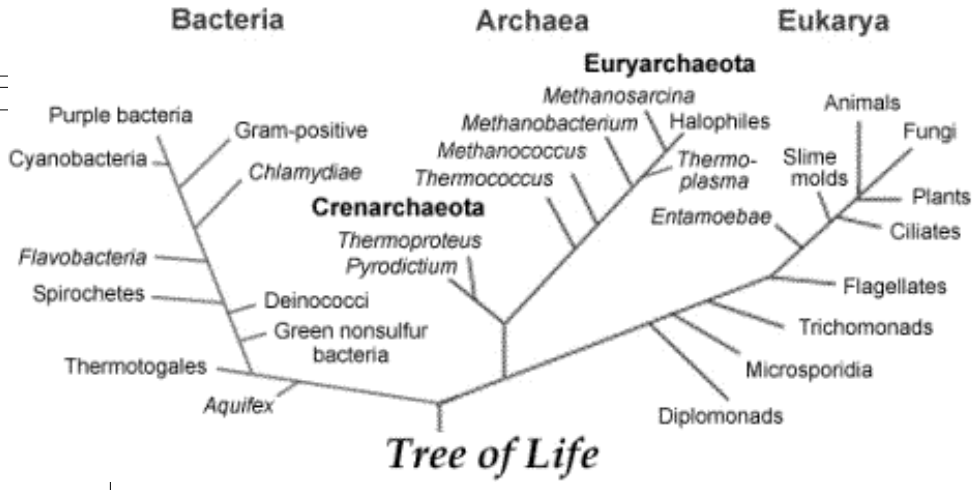


Figure 1.3: Length Distributions for Predicted Protein Coding Regions in Entire Genomes. A. *H. influenzae*, among the 1,743 regions, amino acid chains of lengths between 140 and 160 are the most frequent. B. *M. genitalium* with 468 regions, and preferred amino acid chains of length between 120 and 140 or 280 and 300. C. The archaeon *M. jannaschii* with 1,735 regions; amino acid chains of length between 140 and 160 are the most frequent. D. *S. cerevisiae*, among the 6,200 putative protein coding regions, amino acid chains of length between 100 and 120 are the most frequent; this interval is followed by the interval 120 to 140. As described in a 1997 correspondence in *Nature*, the *S. cerevisiae* set clearly contains an overrepresentation (of artifact sequences) in the 100-120 length interval [107].

Some Observations

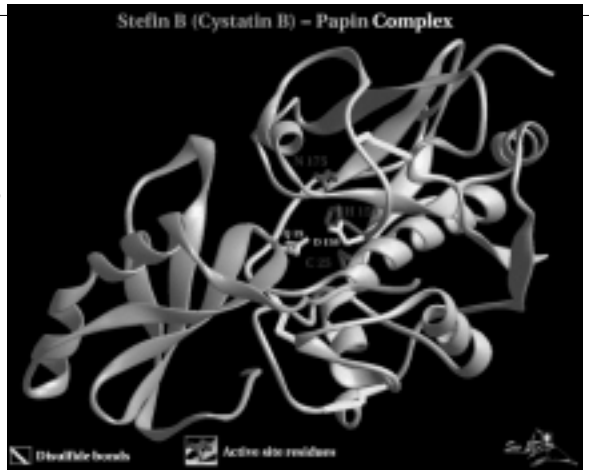
- It is observed that the peaks for the prokaryote *H. influenzae* and the eukaryote S. *cerevisiae* are positioned in different intervals: at 140-160 and 100-120, respectively.
- Further studies show that a eukaryotic distribution from a wide range of species peaks at 125 amino acids and that the distribution displays a periodicity based on this size unit.
- Interestingly, the distribution for the achaeon *M. jannaschii* is sort of inbetween the *H. influenzae* (eukaryote) and the *S. cerevisiae* (prokaryote) distributions.



- Tree of life: Three great kingdoms
  - Drawn by sequence analysis of rRNA
  - Three kingdoms: bacteria, archaea, eukarya

* Former five kingdoms: Animalia, Plantae, Fungi, Protista, Monea

3.3 Protein Function



- Protein function
  - Does not depend on whole structure
  - Determined mainly by local sequences

  (e.g. active site of enzyme)

- End of human genome project: finding genes, measuring its function and activity and many other analysis are required.
  - Prediction by experiment: time costing and hard job & non-real environment (under laboratory environment)
  - Many computational approaches can be helpful for protein analysis.

Species	Sequences
Saccharomyces cerevisiae	4,711
Homo sapiens	4,470
Escherichia coli	3,968
Mus musculus	2,727
Rattus norvegicus	2,200
Bacillus subtilis	1,843
Caenorhabditis elegans	1,666
Haemophilus influenzae	1,632
Schizosaccharomyces pombe	1,015
Drosophila melanogaster	981
Bos bovis	980
Methanococcus jannaschii	884
Gallus gallus	750
Salmonella typhimurium	662
Mycobacterium tuberculosi	644

Table 1.3: The Number of Sequences for the 15 Most Abundant Organisms in SWISS-PROT rel. 35, November 1997.

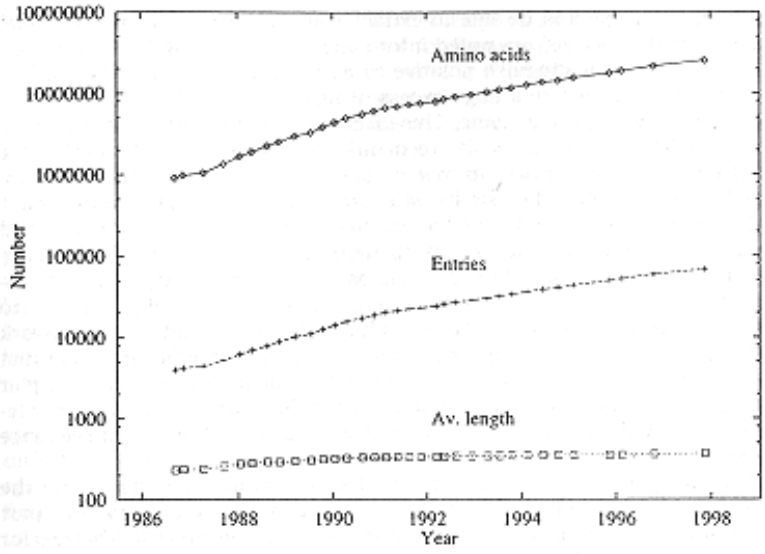


Figure 1.4: The Exponential Growth of the SWISS-PROT Database in the Period 1980-1997. The size of SWISS-PROT rel. 35 is in the order of 25,000,000 amino acids from 68,000 entries.

4. On the Information Content of Biological Sequences

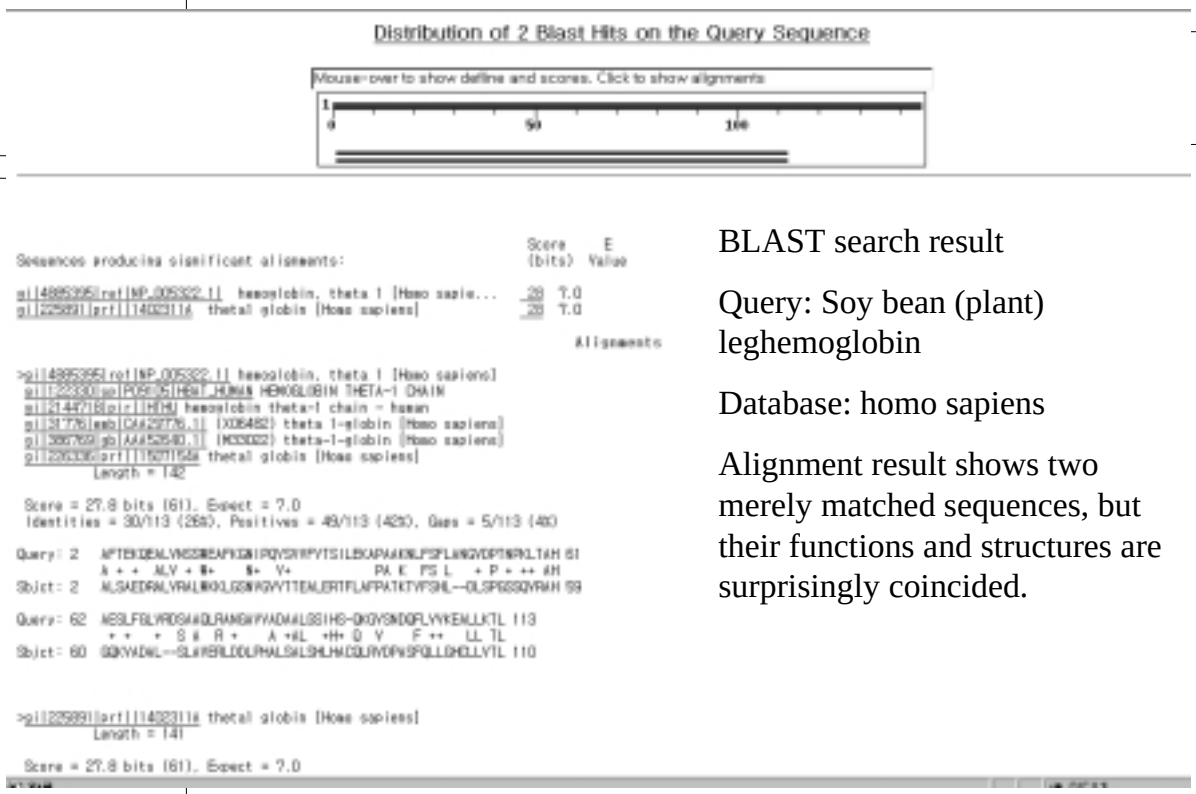
- Concept of information and its quantification is essential for the basic principles of machine-learning approaches in molecular biology
- Data-driven prediction methods can extract essential features from individual examples and discard unwanted information.
- Machine-learning techniques are excellent for the task of discarding of and compacting redundant sequence information

- Information reduction
  - Key feature in the understanding of almost any kind of system
  - Create simple representation of a sequence space → much more powerful and useful than the original data that contain all details
- Compression rate
  - Ratio between size of the an encoded corpus of sequences and the original corpus of sequences
  - Global quantifier of the degree of regularity in the data

$$R_C = \frac{S_E}{S_O}$$

4.2 Alignment vs. Prediction: When are Alignment Reliable?

- New sequences are aligned against all sequences in DB
  - Hints toward structural/functional relationships + functional insights
- Fundamental question
  - When is the sequence similarity high enough that one may infer a structural/functional similarity from the pairwise alignment of two sequences?
  - Given the detected overlap in a sequence segment, can a similarity threshold be defined that sifts out cases where the inference will be reliable?
- Answer
  - It depends on the structural/functional aspect one wants to investigate
  - Different for each task
- Prediction
  - When alignment alone is not enough to lead a reliable inference



BLAST search result

Query: Soy bean (plant) leghemoglobin

Database: homo sapiens

Alignment result shows two merely matched sequences, but their functions and structures are surprisingly coincided.

4.3 Prediction of Functional Features

- Two protein sequences share sequence similarity
  - These proteins share common function?
- New sequence identity threshold is required for prediction of function
  - Threshold used for structural problems can not be used.
- Solution
  - Split each sequence into a number of subsequences
  - The fraction of aligned site per alignment

4.4 Global and Local Alignments and Substitution Matrix Entropies

- Pairwise alignment
  - There is no optimal answer (actually we don't know optimal alignment): changed by various options
  - Dynamic programming: computational demanding job
  - Heuristic approaches: BLAST, FASTA
- Substitution matrix
  - Set of scores s_{ij} of replacing amino acid i by amino acid j
  - Generated from simplified protein evolution model involving amino acid frequencies, p_i

$$s_{ij} = \frac{1}{\lambda}\left(\ln \frac{q_{ij}}{p_i p_j}\right) \quad p_i = \frac{1}{20} \quad q_{ij} = \begin{cases} q \text{ for } i = j \\ \overline{q} \text{ for } i \neq j \end{cases}$$

PAM250 Matrix

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
A	Ala	.18																			
R	Arg	-.15	.61																		
N	Asn	.02	0	.20																	
D	Asp	.03	-.13	.21	.39																
C	Cys	-.20	-.36	-.36	-.51	1.19															
Q	Gln	-.04	.13	.08	.16	-.54	.40														
E	Glu	.03	-.11	.14	.34	-.53	.25	.38													
G	Gly	.13	-.26	.03	.06	-.34	-.53	.25	.38												
H	His	-.14	.16	.16	.07	-.34	.29	.07	-.21	.65											
I	Ile	-.05	-.20	-.18	-.24	-.23	-.20	-.20	-.26	-.24	.45										
L	Leu	-.19	-.30	-.29	-.40	-.60	-.18	-.34	-.41	-.21	.24	.59									
K	Lys	-.12	.34	.10	.01	-.54	.07	-.01	-.17	0	-.19	-.29	.47								
M	Met	-.11	-.04	-.17	-.26	-.52	-.10	-.21	-.28	-.21	.22	.37	.04	.64							
F	Phe	-.35	-.45	-.35	-.56	-.43	-.47	-.54	-.48	-.18	.10	.18	-.53	.02	.91						
P	Pro	.11	-.02	-.05	-.10	-.28	.02	-.06	-.05	-.02	-.20	-.25	-.11	-.21	-.46	.59					
S	Ser	.11	-.03	.07	.03	0	-.05	0	.11	-.08	-.14	-.28	-.02	-.16	-.32	.09	.16				
T	Thr	.12	-.09	.04	-.01	-.22	-.08	-.04	0	-.13	.01	-.17	0	-.06	-.31	.03	.13	.26			
W	Trp	-.58	.22	-.42	-.68	-.78	-.48	-.70	-.70	-.28	-.51	-.18	-.35	-.42	.04	-.56	-.25	-.52	1.73		
Y	Tyr	-.35	-.42	-.21	-.43	.03	-.40	-.43	-.52	-.01	-.09	-.09	-.44	-.24	.70	-.49	-.28	-.27	-.02	1.01	
V	Val	.02	-.25	-.17	-.21	-.19	-.19	-.18	-.14	-.22	.37	.19	-.24	.18	-.12	-.12	-.10	.03	-.62	-.25	.43

Matches involving rare amino acids count more than common amino acids

Mismatches between similar amino acids have higher score than between distinct amino acids

4.5 Consensus Sequences and Sequence Logos

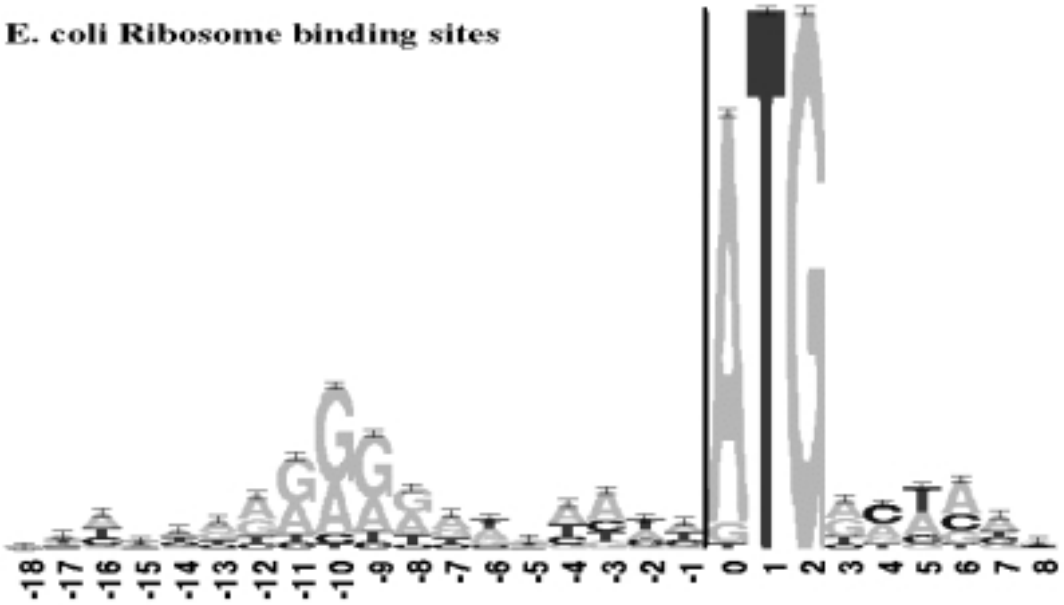
- Creating consensus sequences
  - From alignments
  - Choose most common nucleotides or amino acids
  - Can be highly misleading
- Sequence logo
  - Based on Shannon information content at each position
  - Emphasize the deviation from the uniform or flat distribution

$$D(i) = \log_2 |A| + \sum_{k=1}^{|A|} p_k(i) \log_2 p_k(i)$$

$$|A| = 4 \text{ or } 20$$

  - Displays a significant degree of deviation from flat distribution

E. coli Ribosome binding sites



- Initiation codon (ATG) is very well conserved.
- 5' to the initiation codons (-12~-7) are conserved also.

- Slight variation of the logo formula

$$H(i) = H(P(i), Q(i)) = \sum_{k=1}^{|A|} p_k(i) \log \frac{p_k(i)}{q_k(i)}$$

  - Based on relative entropy
  - Quantification of the contrast between the observed probabilities $P(i)$ and a reference probability distribution $Q(i)$
  - Q: depends on the position i in the alignment

5. Prediction of Molecular Function and Structure

- DNA, RNA and protein sequence analysis based on machine-learning approaches can be helpful for many biological problems.
  - Finding intron splice sites in mRNA processing
  - Gene finding
  - Recognition of promoter and other regulatory regions
  - Prediction of gene expression levels
  - Prediction of DNA bending and bendability
  - Sequence clustering
  - RNA secondary structure prediction
  - Protein function/structure prediction
  - Protein family classification