

- Bioinformatics and Computational Biology: History and Biological Background (JH) 10.10
- The Parsimony criterion GKN 13.10

- Stochastic Models of Sequence Evolution GKN 17.10
- The Likelihood criterion GKN 20.10
- Tut: 9-10 11=12 (Friday)

- Trees in phylogenetics and population genetics GKN 24.10
- Estimating phylogenies and genealogies I GKN 27.10
- Tut: 9-10 11-12 (Friday)

- Estimating phylogenies and genealogies II GKN 31.10
- Estimating phylogenies and genealogies III 3.11
- Tut: 9-10 11-12 (Friday)

- Alignment Algorithms I (Optimisation) (JH) 7.11
- Alignment Algorithms II (Statistical Inference) (JH) 10.11
- Tut: 9-10 11-12 (Friday)

- Finding Signals in Sequences (JH) 14.11
- Stochastic Grammars and their Biological Applications: Hidden Markov Models (JH) 17.10
- Tut: 9-10 11-12 (Friday)

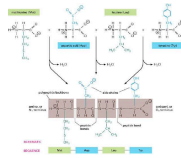
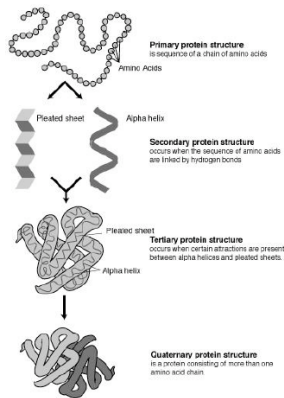
- Stochastic Grammars and their Biological Applications: Context Free Grammars (JH) 21.11
- RNA molecules and their analysis (JH) 24.11
- Tut: 9-10 11-12 (Friday)

- Open Problems in Bioinformatics and Computational Biology I (JH) 28.11
- Possibly: Evolving Grammars, Pedigrees from Genomes
- Open Problems in Bioinformatics and Computational Biology II (GKN) 1.12
- Possibly: The phylogeny of language: traits and dates, What can FIV sequences tell us about their host cat population?
- Tut: 9-10 11-12 (Friday)

Schedule

Proteins

Proteins: a string of amino acids. Often folds up in a well defined 3 dimensional structure. Has enzymatic, structural and regulatory functions.



$S \in \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}^*$

Name	Three and one-letter code	R group (side chain)	Polarity	Molecular weight
Alanine	Ala, A	-CH ₃	Nonpolar	89.09
Leucine	Leu, L	-CH ₂ CH(CH ₃) ₂	Nonpolar	133.17
Isoleucine	Ile, I	-CH(CH ₃)CH ₂ CH ₃	Nonpolar	133.17
Valine	Val, V	-CH(CH ₃) ₂	Nonpolar	117.15
Proline	Pro, P		Nonpolar	115.09
Threonine	Thr, T	-CH(CH ₃)OH	Nonpolar	119.09
Tryptophan	Try, W		Nonpolar	204.23
Methionine	Met, M	-CH ₂ CH ₂ CH ₂ SCH ₃	Nonpolar	149.21
Cysteine	Cys, C	-SH	Polar(S)	75.07
Serine	Ser, S	-CH ₂ OH	Polar	105.09
Threonine	Thr, T	-CH(OH)CH ₃	Polar	119.12
Tyrosine	Tyr, Y	-CH ₂ -C ₆ H ₄ -OH	Polar	181.19
Cysteine	Cys, C	-CH ₂ -SH	Polar	119.09
Asparagine	Asn, N	-CH ₂ CONH ₂	Polar	132.12
Glutamine	Gln, Q	-CH ₂ CH ₂ CONH ₂	Polar	146.15
Aspartic acid	Asp, D	-CH ₂ COOH	- Charged	133.10
Glutamic acid	Glu, E	-CH ₂ CH ₂ COOH	- Charged	147.13
Lysine	Lys, K	-CH ₂ (CH ₂) ₄ NH ₃ ⁺	+ Charged	146.19
Arginine	Arg, R	-CH ₂ (CH ₂) ₃ NH ₂ -C(NH ₂) ₂	+ Charged	174.20
Histidine	His, H		+ Charged(S)	155.15

Bioinformatics and Computational Biology: History & Biological Background

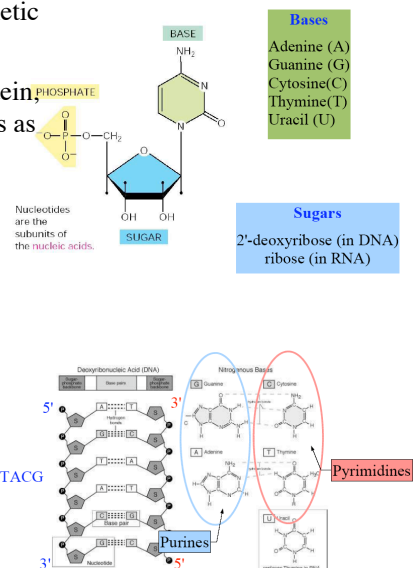
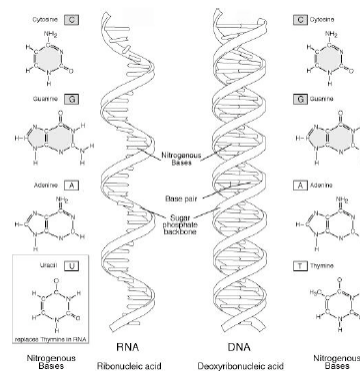
Early History up to 1953

- 1838 Schwann and Schleiden Cell Theory
- 1859 Charles Darwin publishes Origin of Species
- 1865 Mendel discovers basic laws of inheritance (largely ignored)
- 1869 Miescher Discovers DNA
- 1900 Mendels laws rediscovered.
- 1944 Avery shows DNA contains genetic information
- 1951 Corey & Pauling Secondary structure elements of a protein.
- 1953 Watson & Crick proposes DNA structure and states

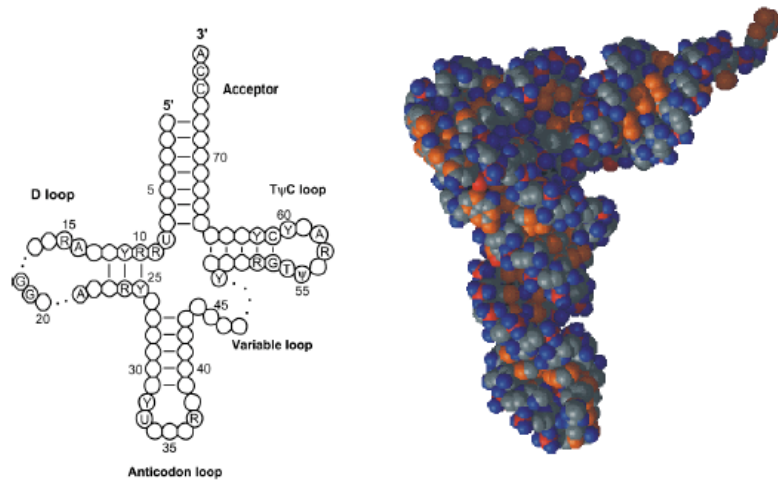
DNA & RNA

DNA: The Information carrier in the genetic material. Usually double helix.

RNA: messenger tape from DNA to protein, regulatory, enzymatic and structural roles as well. More labile than DNA

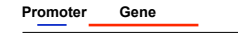


An Example: t-RNA



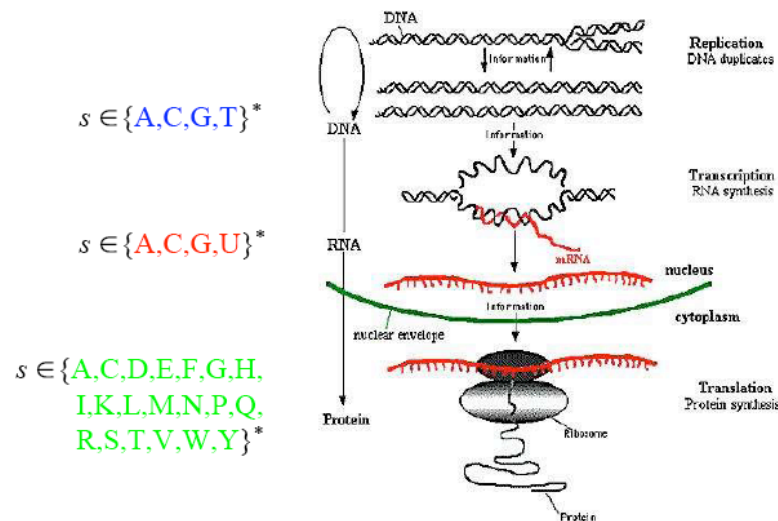
From Paul Higgs

History up to 1953-66



- 1955 Sanger first protein sequence – Bovine Insulin
- 1957 Kendrew structure of Whale Myoglobin
- 1958 Crick, Goldschmidt,.... Central Dogma
- 1958 First quantitative method for phylogeny reconstruction (UGPMA - Sokal and Michener)
- 1959 Operon Models proposed (Jakob and Monod)
- 1966 Genetic Code Determined
- 1967 First RNA sequencing

The Central Dogma



The Genetic Code

Genetic Code: Mapping from 3-nucleotides (codons) to amino acids (20) + stop codon. This 64-->21 mapping creates the distinction silent/replacement substitution.

TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys
TTA	Leu	TCA	Ser	TAA	!!!	TGA	!!!
TTG	Leu	TCG	Ser	TAG	!!!	TGG	Trp
CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
CTC	Leu	CCT	Pro	CAC	His	CGC	Arg
CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
ATG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Substitutions	Number	Percent
Total in all codons	549	100
Synonymous	134	25
Nonsynonymous	415	75
Missense	392	71
Nonsense	23	4

Ser Thr Glu Met Cys Leu Met Gly Gly
TCA ACT GAG ATG TGT TTA ATG GGG GGA
*** * * * * * * ***
TCG ACA GGG ATA TAT CTA ATG GGT ATA
Ser Thr Gly Ile Tyr Leu Met Gly Ile

History 1966-80

1969-70 Temin + Baltimore Reverse Transcriptase

1970 Needleman-Wunch algorithm for pairwise alignment

1971-73 Hartigan-Fitch-Sankoff algorithm for assigning nucleotides to inner nodes on a tree.

1976/79 First viral genome – MS2/φX174

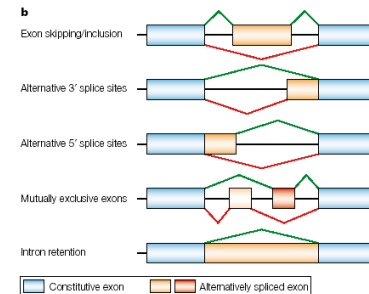
1977/8 Sharp/Roberts Introns

1979 Alternative Splicing

1980 Mitochondrial Genome (16.569bp) and the discovery of alternative codes

Genes, Gene Structure & Alternative Splicing

- Presently estimated Gene Number: 24.000, Average Gene Size: 27 kb
- The largest gene: Dystrophin 2.4 Mb - 0.6% coding – 16 hours to transcribe.
- The shortest gene: tRNA^{TYR} 100% coding
- Largest exon: ApoB exon 26 is 7.6 kb Smallest: <10bp
- Average exon number: 9 Largest exon number: Titin 363 Smallest: 1
- Largest intron: WWOX intron 8 is 800 kb Smallest: 10s of bp
- Largest polypeptide: Titin 38.138 smallest: tens – small hormones.
- Intronless Genes: mitochondrial genes, many RNA genes, Interferons, Histones,...



1. A challenge to automated annotation.
2. How widespread is it?
3. Is it always functional?
4. How does it evolve?

Cartegni, L. et al. (2002) "Listening to Silence and understanding nonsense: Exonic mutations that affect splicing" Nature Reviews Genetics 3.4:285-300, HMG p291-294

Strings and Comparing Strings

1970 Needleman-Wunch algorithm for pairwise alignment for maximizing similarity

1972 Sellers-Sankoff algorithm for pairwise alignment for minimizing distance (Parsimony)

Initial condition: $D_{0,0}=0$, $D_{i,j} := D(s1[1:i], s2[1:j])$

$D_{i,j} = \min\{D_{i,i,j-1} + d(s1[i], s2[j]), D_{i,j-1} + g, D_{i-1,j} + g\}$

Alignment: CTAGG
I=2 v=5) g=10 i v
Cost 17 TT-GT

	40	32	22	14	9	17
T						
	30	22	12	4	12	22
G						
	20	12	2	12	22	32
T						
	10	2	10	20	30	40
T						
	0	10	20	30	40	50
	C	T	A	G	G	

1973-5 Sankoff algorithm for multiple alignment for minimizing distance (Parsimony) and finding phylogeny simultaneously

History 1980-95

1981 Felsenstein Proposes algorithm to calculate probability of observed nucleotides on leaves on a tree.

1981-83 Griffiths, Hudson The Ancestral Recombination Graph.

1987/89 First biological use of Hidden Markov Model (HMM) (Lander and Green, Churchill)

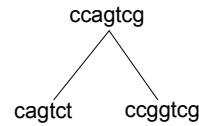
1991 Thorne, Kishino and Felsenstein proposes statistical model for pairwise alignment.

1994 First biological use of stochastic context free grammar (Haussler)

Genealogical Structures

Homology:

The existence of a common ancestor (for instance for 2 sequences)

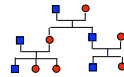


Phylogeny

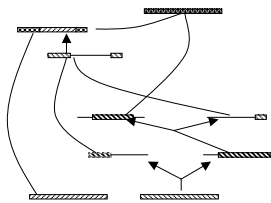


Only finding common ancestors. Only one ancestor.

Pedigree:

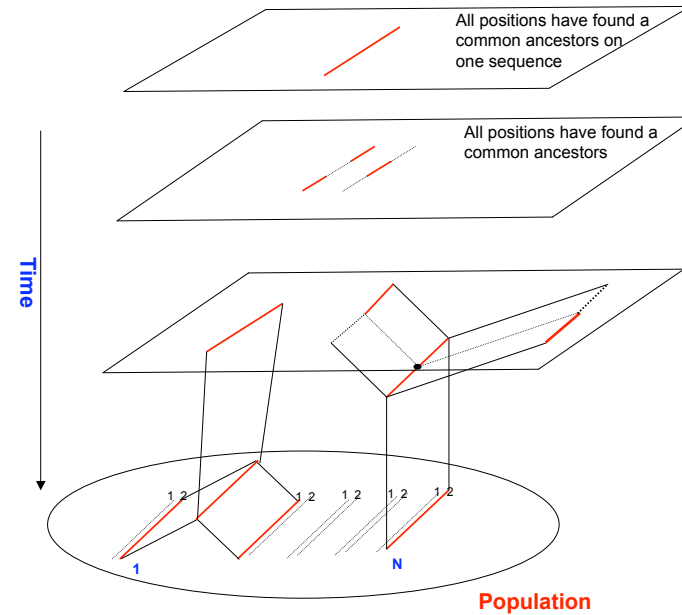


Ancestral Recombination Graph – the ARG

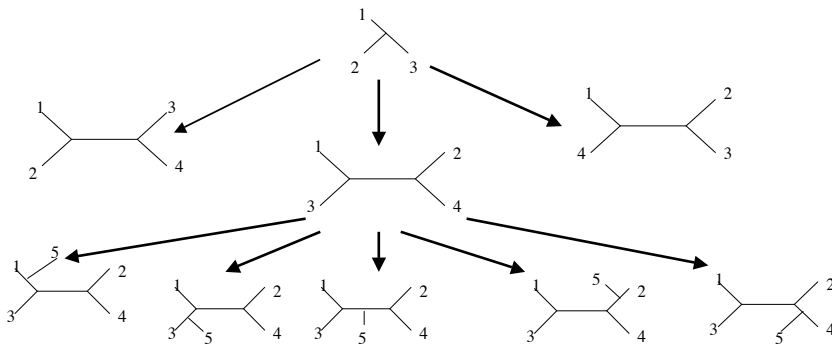


- i. Finding common ancestors.
- ii. A sequence encounters Recombinations
- iii. A “point” ARG is a phylogeny

Time slices



Enumerating Trees: Unrooted & valency 3



Recursion: $T_n = (2n-5) T_{n-1}$ Initialisation: $T_1 = T_2 = T_3 = 1$

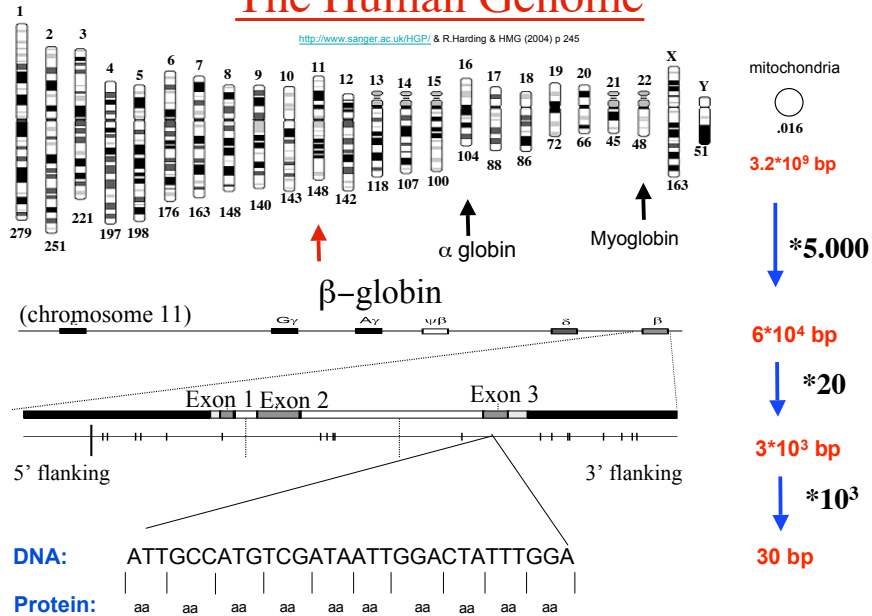
$$\prod_{j=3}^{n-1} (2j-3) = \frac{(2n-5)!}{(n-2)! 2^{n-2}}$$

4	5	6	7	8	9	10	15	20	
3	15	105	945	10345	$1.4 \cdot 10^5$	$2.0 \cdot 10^6$	$7.9 \cdot 10^{12}$	$2.2 \cdot 10^{20}$	

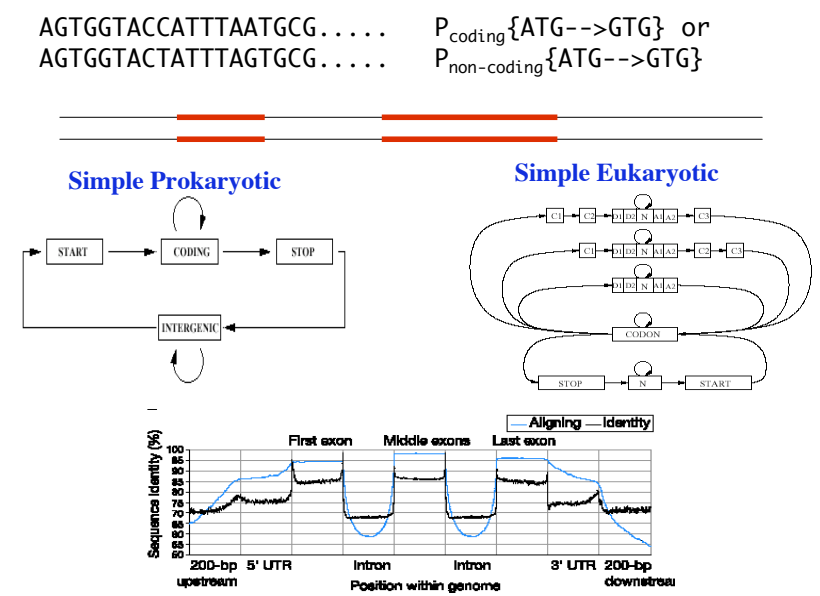
History 1995-2005

- 1995 First prokaryotic genome – *H. influenzae*
- 1996 First unicellular eukaryotic genome – Yeast
- 1998 The first multi-cellular eukaryotic genome – *C.elegans*
- 2000 *Drosophila melanogaster*, *Arabidopsis thaliana*
- 2001 Human Genome
- 2002 Mouse Genome
- 2005 Chimp Genome

The Human Genome



Molecular Evolution and Gene Finding: Two HMMs

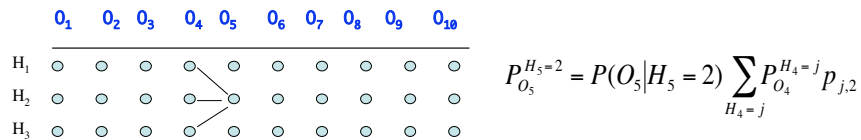


Three Questions for Hidden Structures.

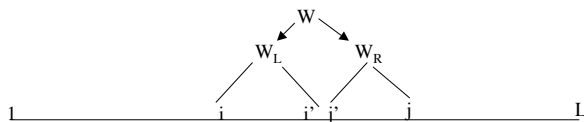
- What is the probability of the data?
- What is the most probable "hidden" configuration?
- What is the probability of specific "hidden" state?

Training: Given a set of instances, find parameters making them probable if they were independent.

HMM/Stochastic Regular Grammar



SCFG - Stochastic Context Free Grammars



- Bioinformatics and Computational Biology: History and Biological Background (JH)
 - The Parsimony criterion GKN
 - Stochastic Models of Sequence Evolution GKN
 - The Likelihood criterion GKN
 - Trees in phylogenetics and population genetics GKN
 - Estimating phylogenies and genealogies I GKN
 - Estimating phylogenies and genealogies II GKN
 - Estimating phylogenies and genealogies III GKN
 - Alignment Algorithms I (Optimisation) (JH)
 - Alignment Algorithms II (Statistical Inference) (JH)
 - Finding Signals in Sequences (JH)
 - Stochastic Grammars and their Biological Applications: Hidden Markov Models (JH)
 - Stochastic Grammars and their Biological Applications: Context Free Grammars (JH)
 - RNA molecules and their analysis (JH)
 - Open Problems in Bioinformatics and Computational Biology I (JH)
 - Possibly: Evolving Grammars, Pedigrees from Genomes
 - Open Problems in Bioinformatics and Computational Biology II (GKN)
- Possibly: The phylogeny of language: traits and dates, What can FIV sequences tell us about their host cat population?