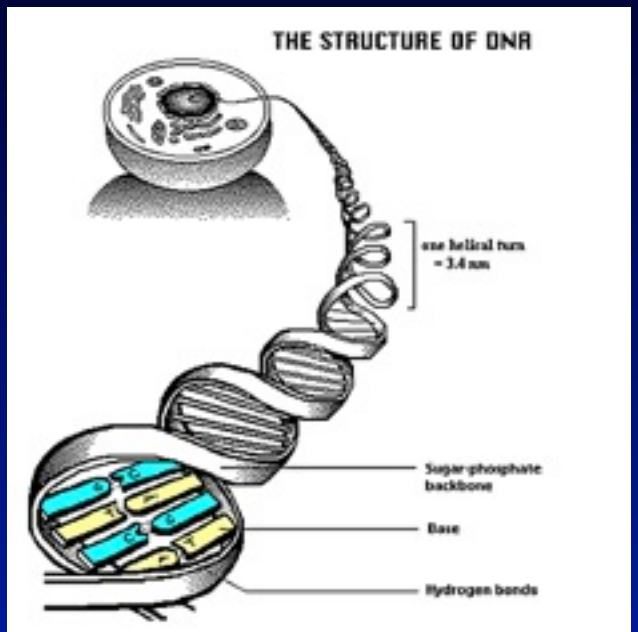
CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS

Introduktion til Bioinformatik

Anders Gorm Pedersen

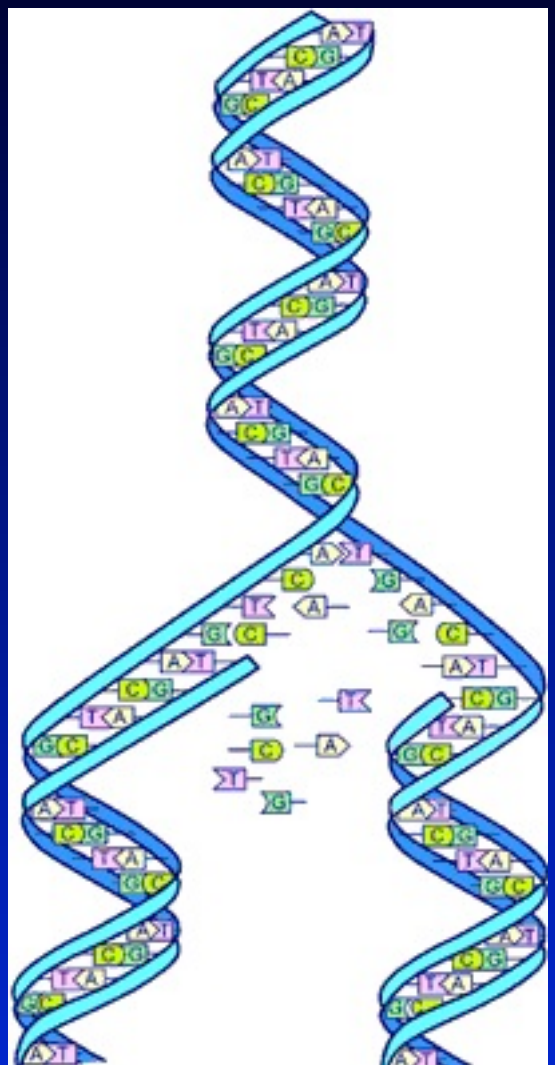
Molekylær Evolutions Gruppen
Center for Biologisk Sekvensanalyse
DTU Systembiologi
gorm@cbs.dtu.dk

What is bioinformatics?



Bioinformatics: computational analysis of biological data

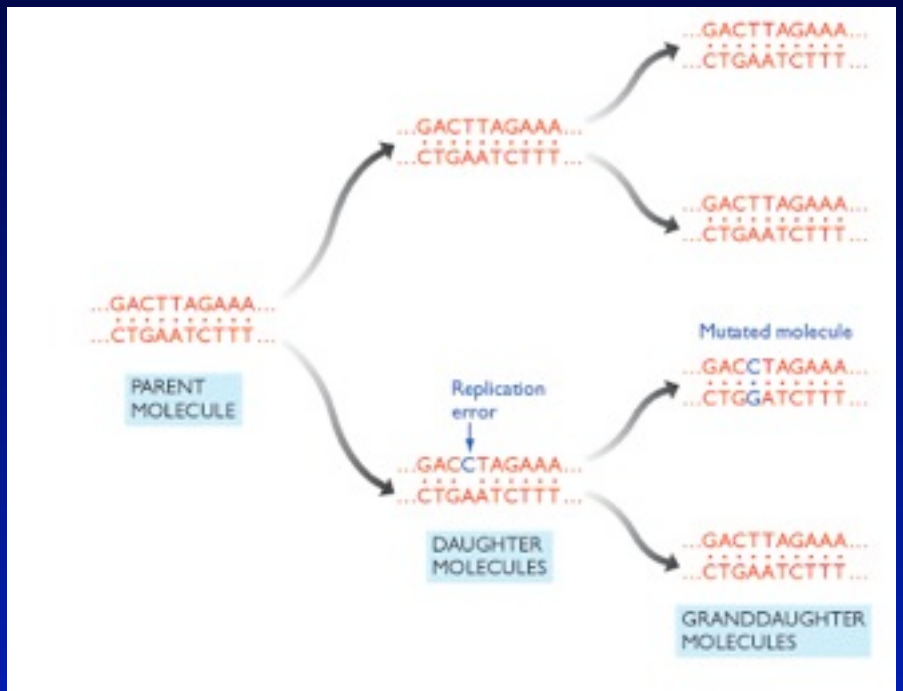
Molecular Basis for Heredity: DNA



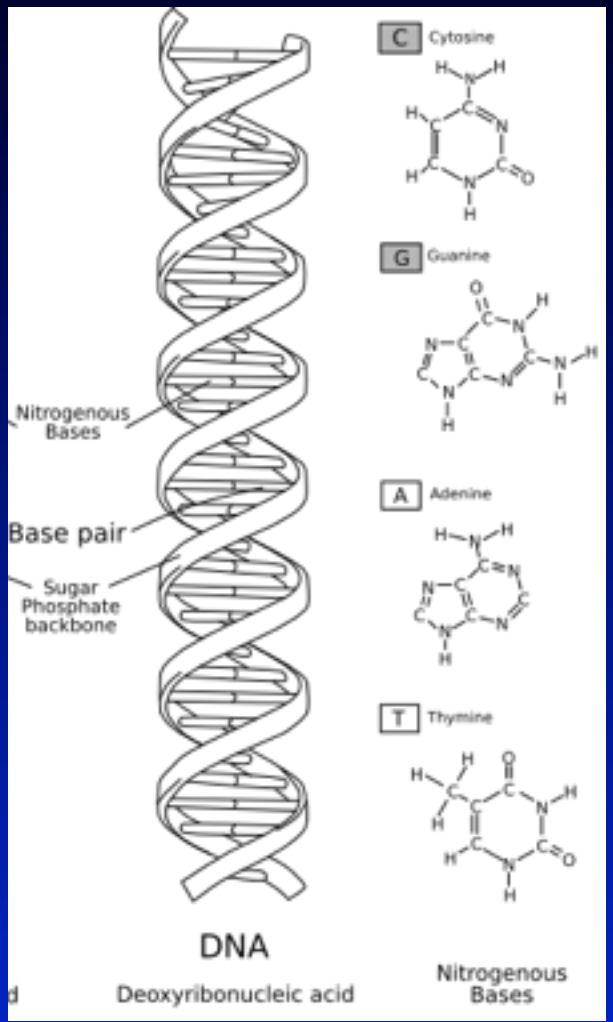
A always paired with T

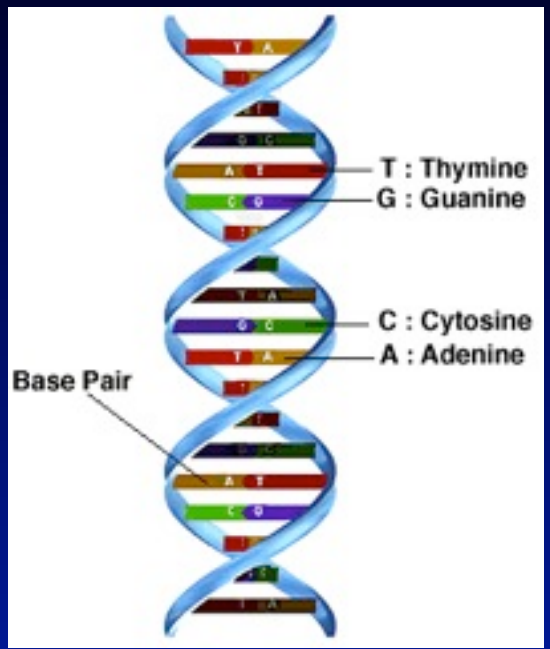
C always paired with G

Molecular Basis for Evolution: DNA Mutation



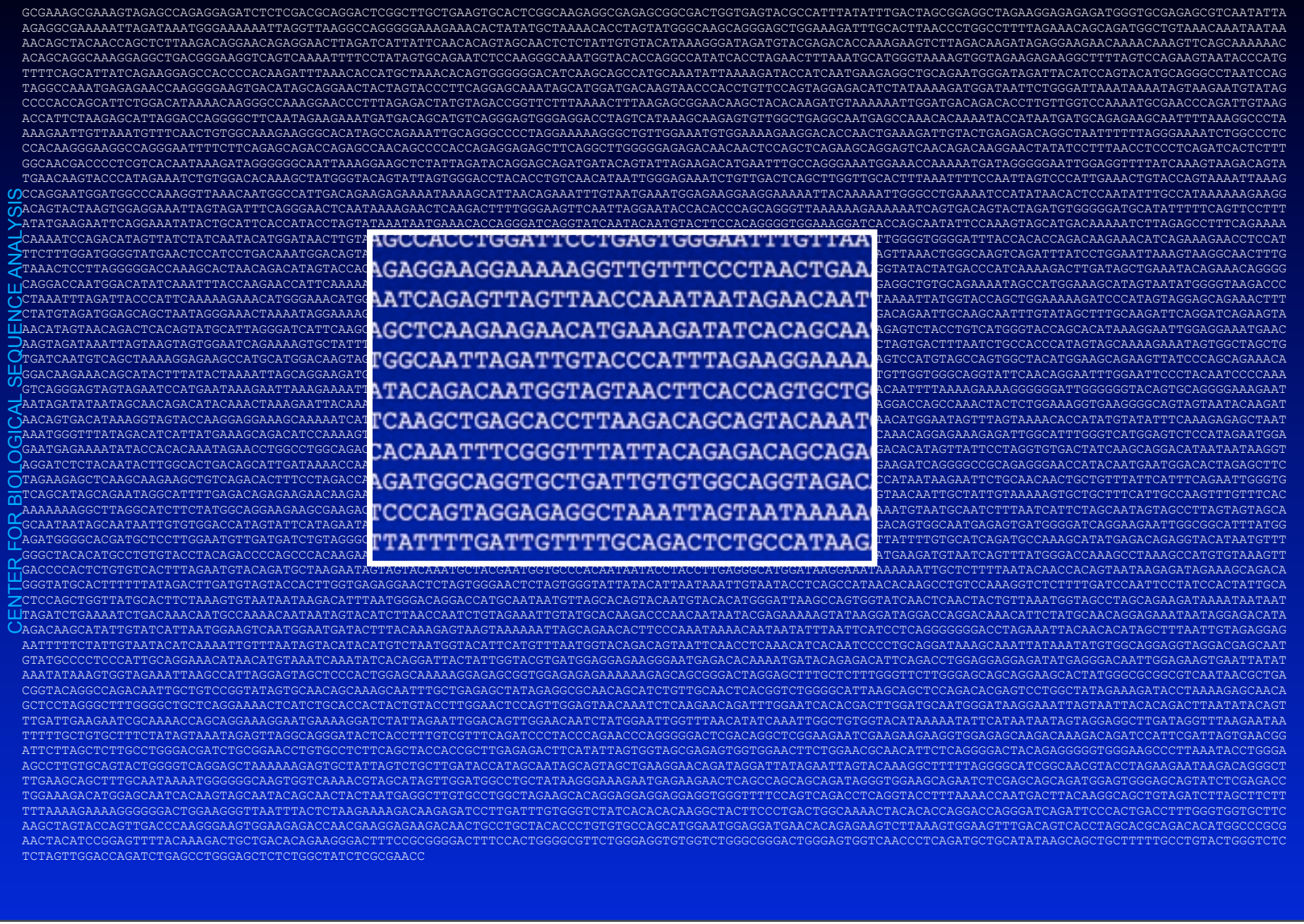
Symbolic representation of DNA structure



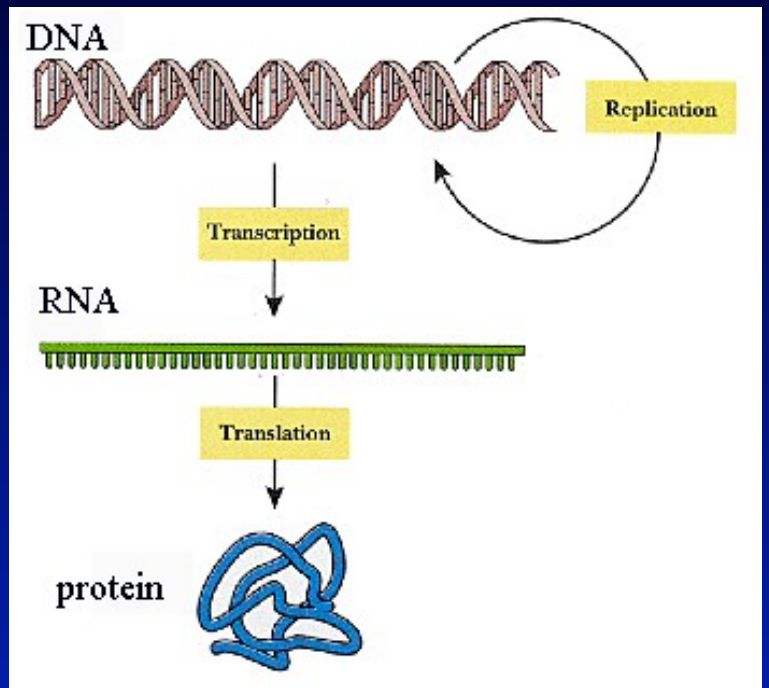


- DNA molecule is a linear polymer
- Structure can be represented as string of 4 symbols: ACTG
- These “sequences” can be analyzed mathematically/ linguistically

HIV genome (approximately 10.000 nucleotides)



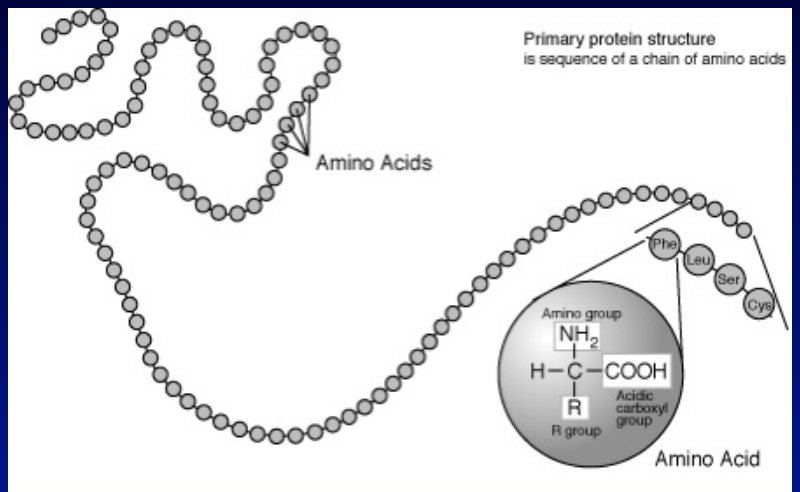
DNA --> RNA --> protein



Standard Genetic Code

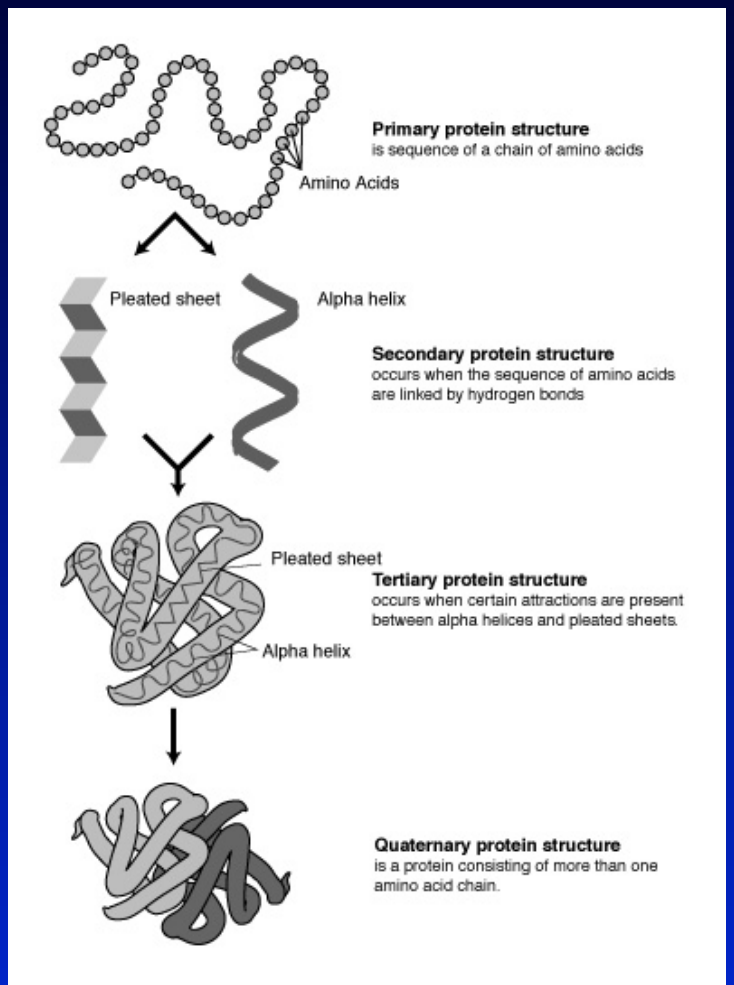
	T			C			A			G			
T	TTT	Phe	F	TCT	Ser	S	TAT	Tyr	Y	TGT	Cys	C	T
	TTC	Phe	F	TCC	Ser	S	TAC	Tyr	Y	TGC	Cys	C	C
	TTA	Leu	L	TCA	Ser	S	TAA	Och	*	TGA	Opa	*	A
	TTG	*Leu*	*L*	TCG	Ser	S	TAG	Amb	*	TGG	Trp	W	G
C	CTT	Leu	L	CCT	Pro	P	CAT	His	H	CGT	Arg	R	T
	CTC	Leu	L	CCC	Pro	P	CAC	His	H	CGC	Arg	R	C
	CTA	Leu	L	CCA	Pro	P	CAA	Gln	Q	CGA	Arg	R	A
	CTG	*Leu*	*L*	CCG	Pro	P	CAG	Gln	Q	CGG	Arg	R	G
A	ATT	Ile	I	ACT	Thr	T	AAT	Asn	N	AGT	Ser	S	T
	ATC	Ile	I	ACC	Thr	T	AAC	Asn	N	AGC	Ser	S	C
	ATA	Ile	I	ACA	Thr	T	AAA	Lys	K	AGA	Arg	R	A
	ATG	*Met*	*M*	ACG	Thr	T	AAG	Lys	K	AGG	Arg	R	G
G	GTT	Val	V	GCT	Ala	A	GAT	Asp	D	GGT	Gly	G	T
	GTC	Val	V	GCC	Ala	A	GAC	Asp	D	GGC	Gly	G	C
	GTA	Val	V	GCA	Ala	A	GAA	Glu	E	GGA	Gly	G	A
	GTG	Val	V	GCG	Ala	A	GAG	Glu	E	GGG	Gly	G	G

Symbolic representation of protein structure

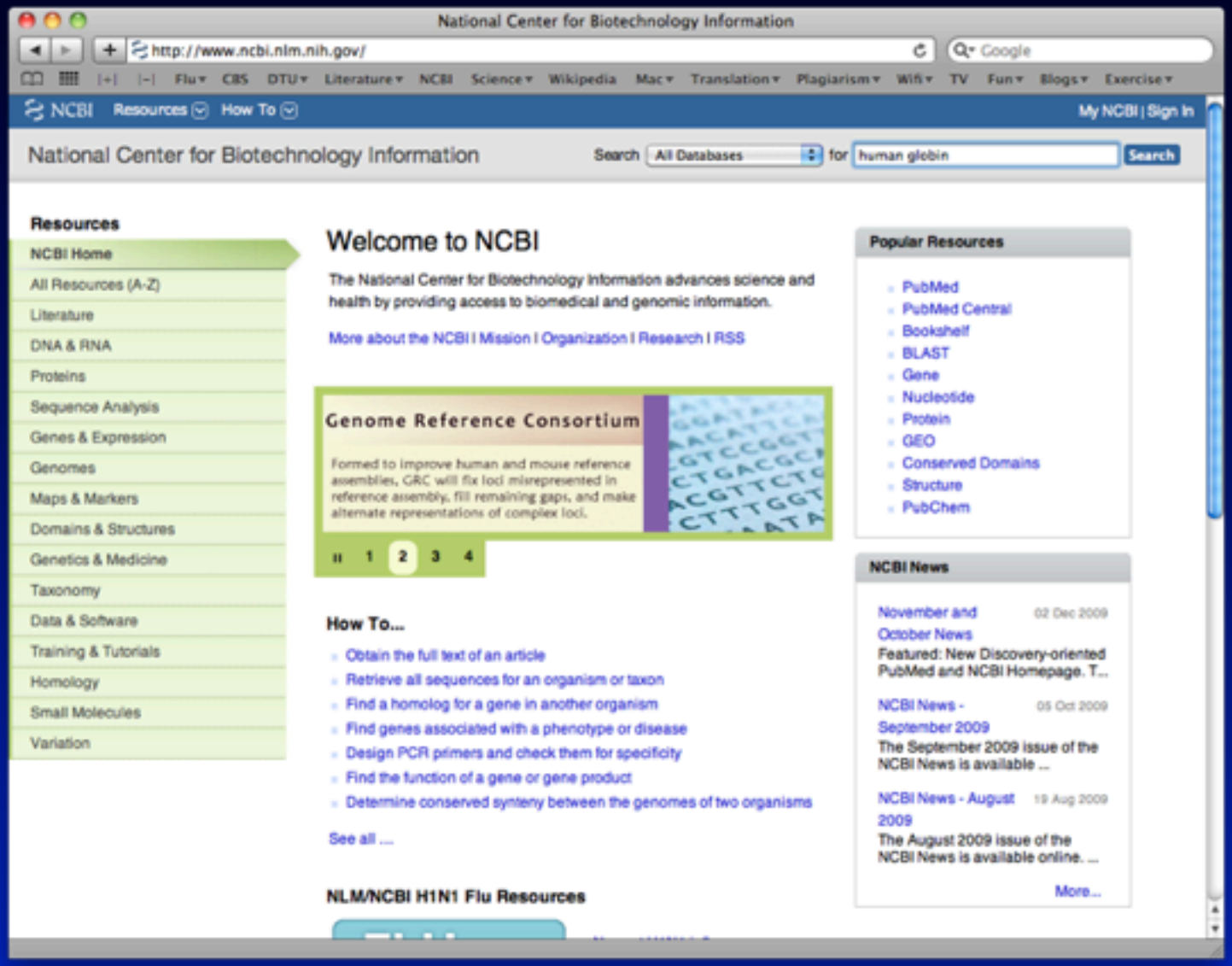


- Proteins are linear polymers
- Built from 20 amino acids
- Can be represented as string of 20 symbols

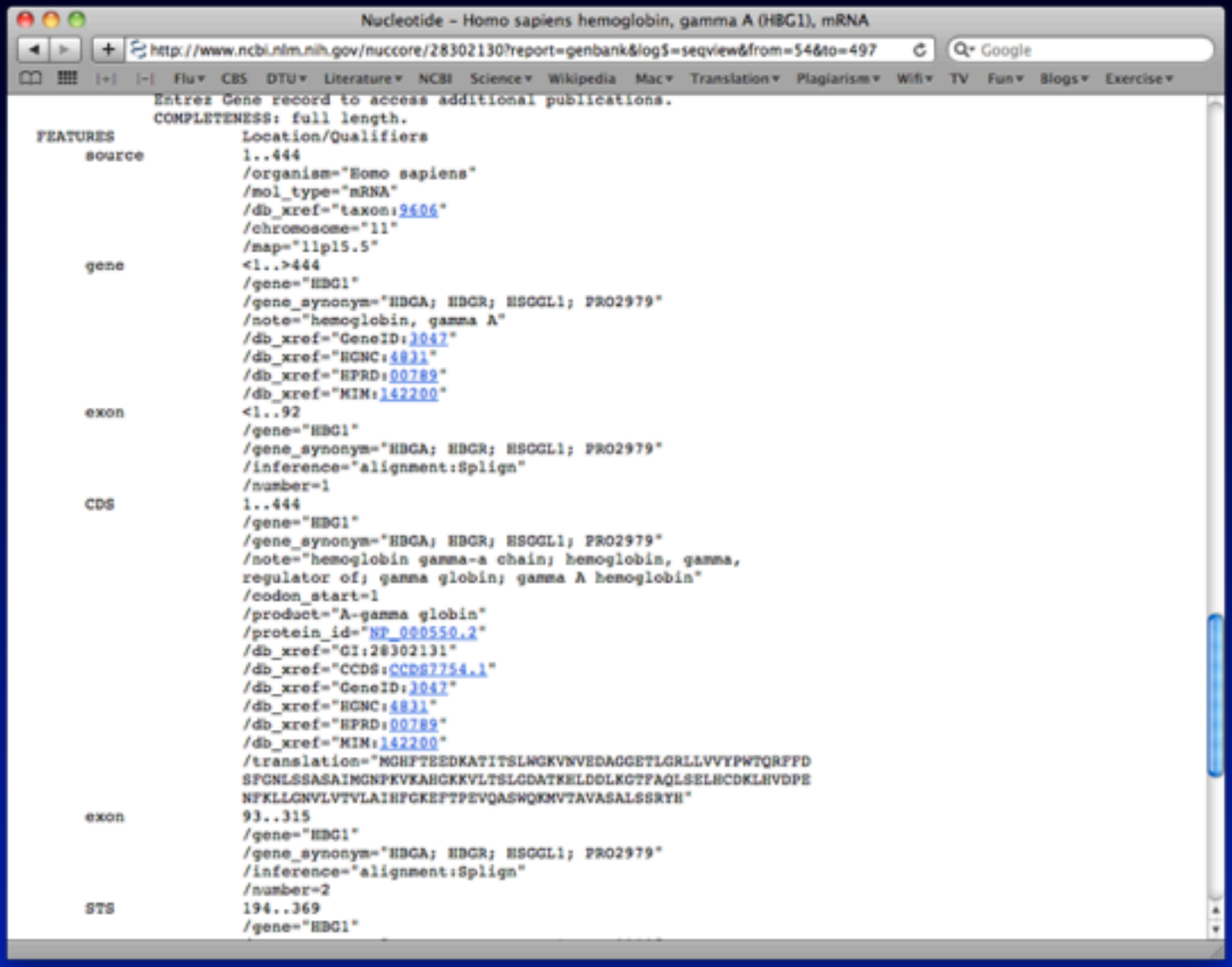
  ACDEFGHIKLMNPQRSTVWY



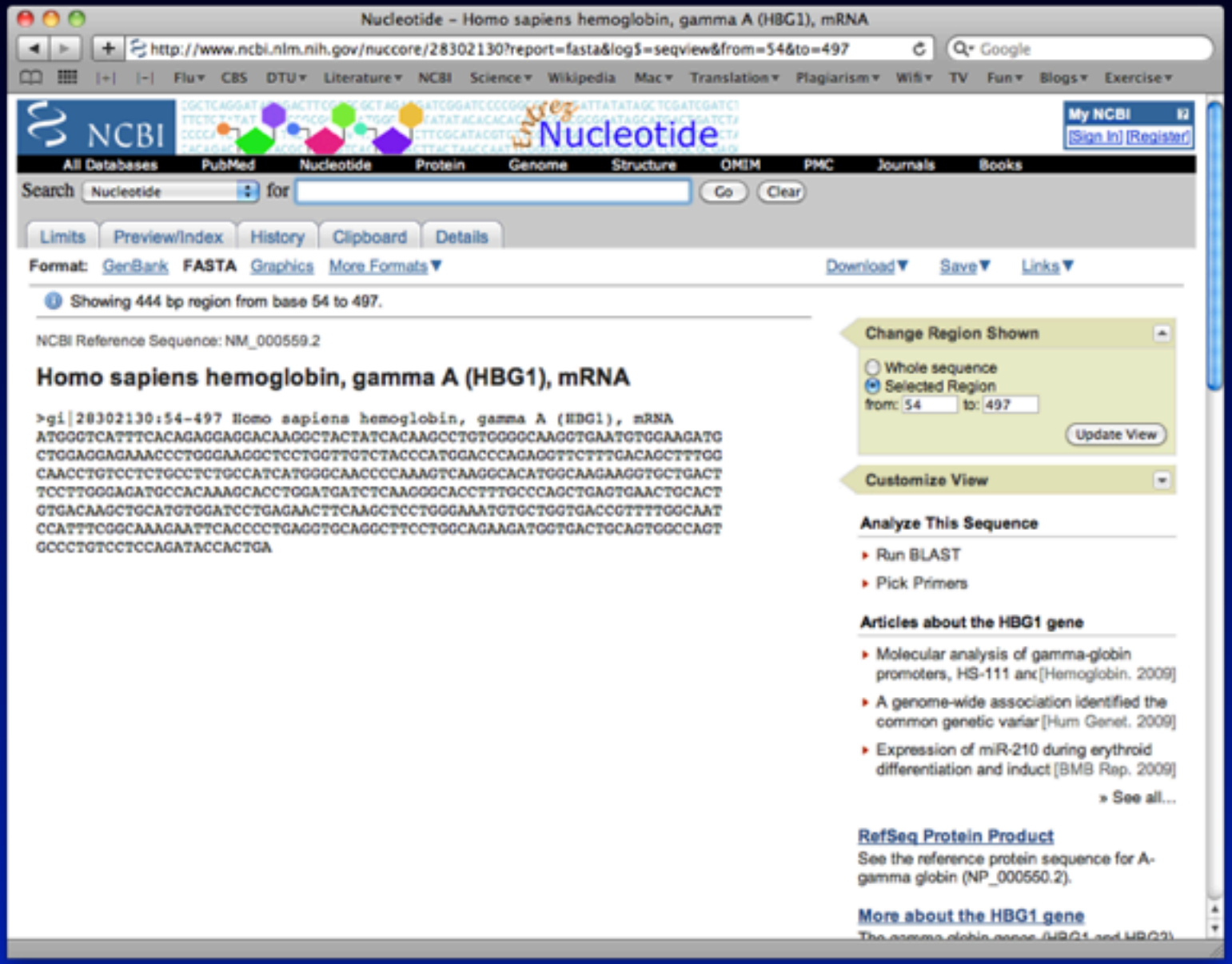
NCBI databases



NCBI databases: Genbank feature table



NCBI databases: fasta format



FASTA file