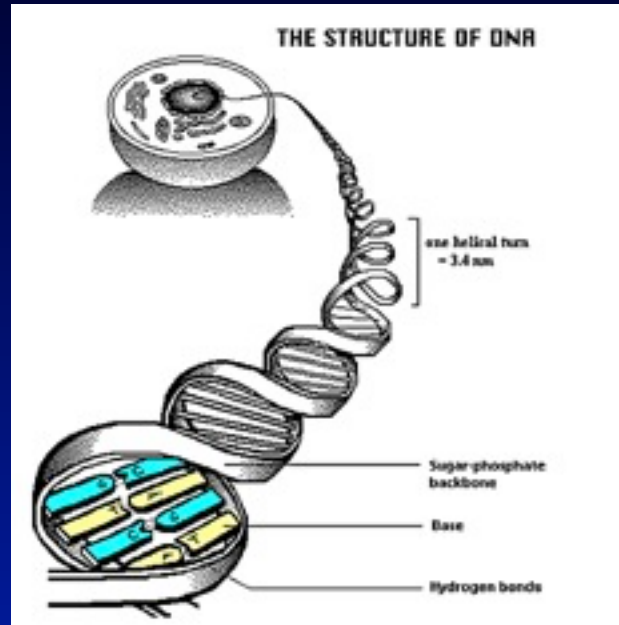


Introduktion til Bioinformatik

Anders Gorm Pedersen

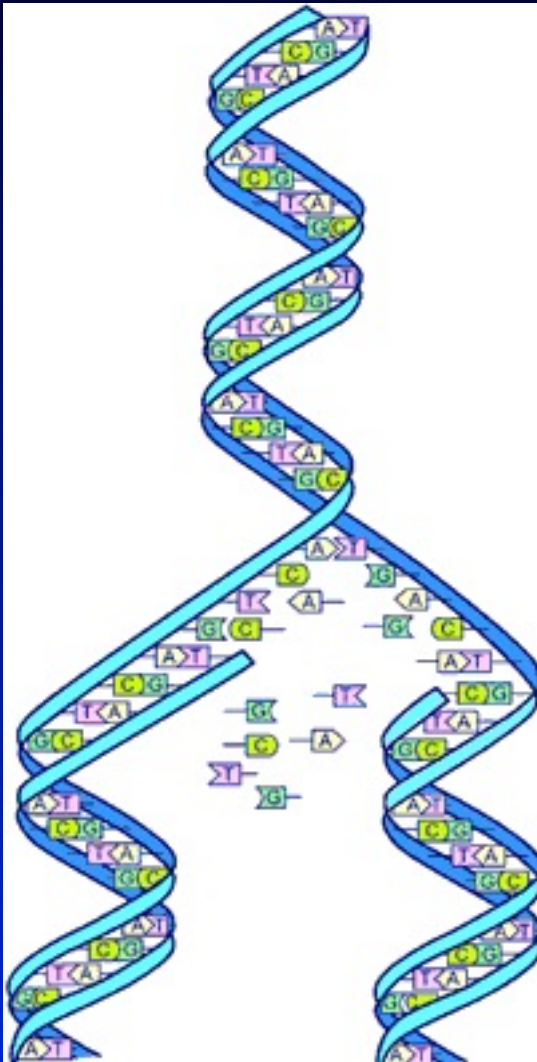
Molekylær Evolutions Gruppen
Center for Biologisk Sekvensanalyse
DTU Systembiologi
gorm@cbs.dtu.dk

What is bioinformatics?



Bioinformatics: computational analysis of biological data

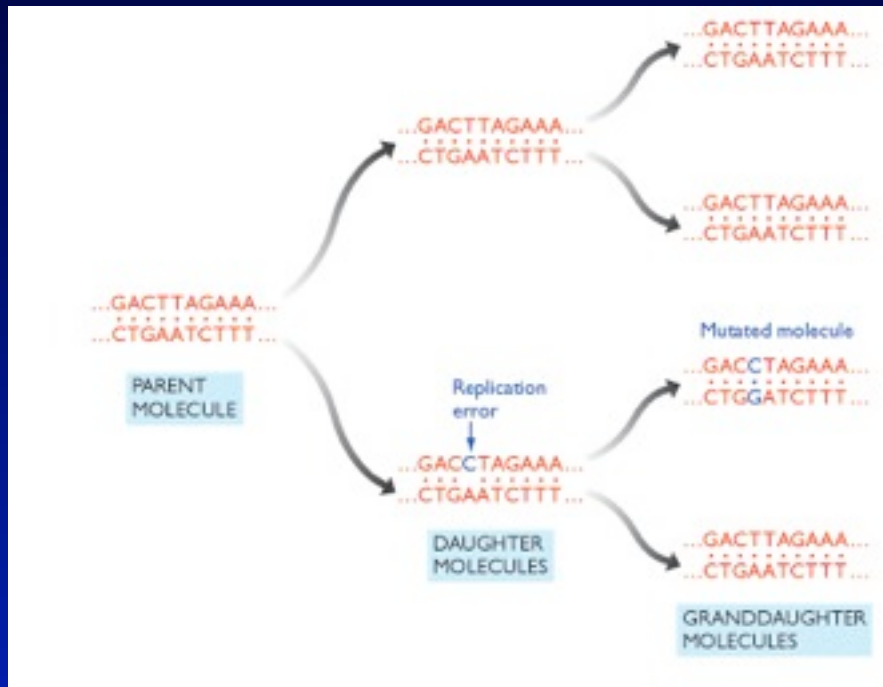
Molecular Basis for Heredity: DNA



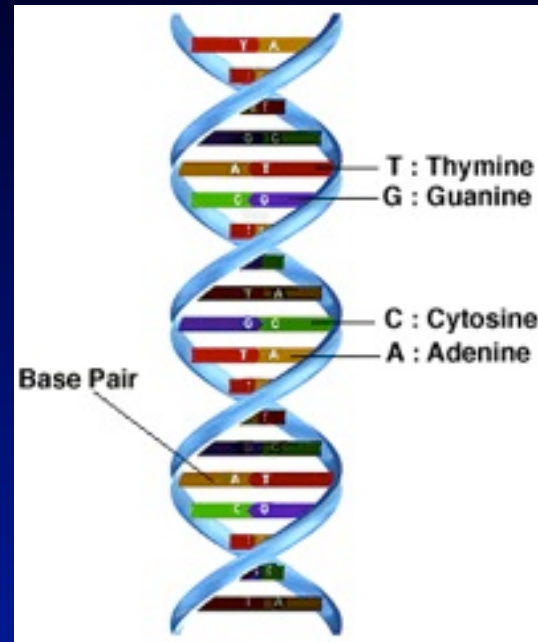
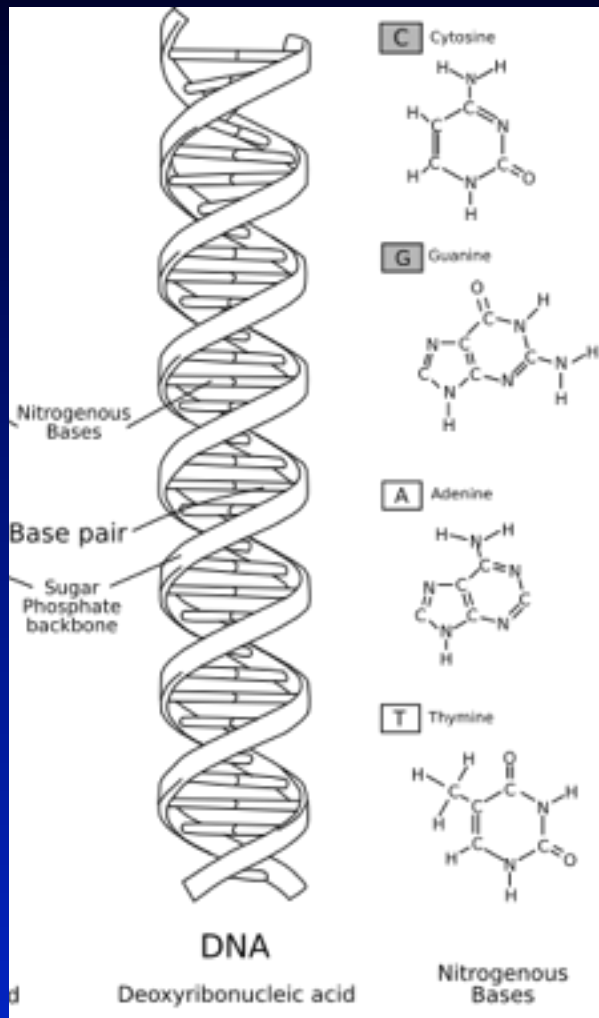
A always paired with T

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)

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS

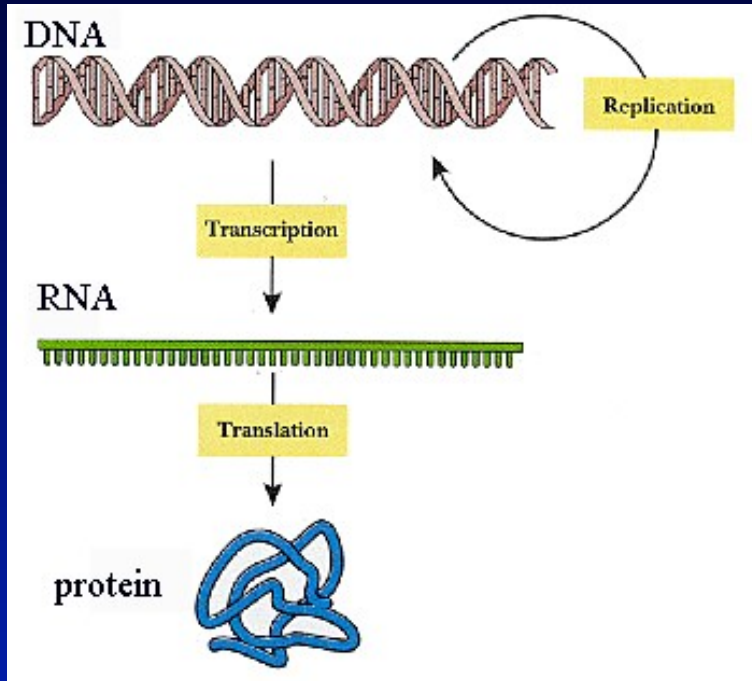
```
GCAGAAAGCGAAAGTAGAGCCAGAGGAGATCTCTCGACGCGAGGACTCGGCTTGCTGAAGTGCATCGGCAAGAGGCGAGAGCGCGACTGGTGTAGTACGCCATTTATATTTGACTAGCGGAGGCTAGAAGGAGAGAGATGGGTGCGAGAGCGTCAATATTA
AGAGGGGAAAAATTAGATAAATGGGAAAAAATTAGTGTAAAGCCAGGAGAAAACACTATATGCTAAAAACCTAGTATGGGCAAGCAGGAGCTGGAAAGATTTGCACTTAACCCCTGGCCCTTTAGAAAACAGCAGATGGCTGTAAACAAATAATA
AACAGCTCAACACAGCTTAAAGCAGGAAACAGAGGAACCTTAGATCATTATCAACACAGTAGCACTCTCTATTGTGTACATAAAAAGGATAGTGTACGAGACACCAAAAGAGTGTAGCAGAGTATAGGAGAAACCAAAAGCTTCAGCAAAAAAC
ACAGCAGGCAAAAGGAGGCTGACGGGAAAGCTCAGTCAAAAATTTCCCTATAGTGCAGAACTCCAAAGGCCAAATGGTACACCAGGCCATATCACCTAGAAGCTTTAAATGCATGGTAAAAGTGGTAGAAGAGAAAGGCTTTTAGTCCGAAAGTAAATACCCATG
TTTTACAGATTTACAGAAAGGAGCCACCCCAAGATTTAAACACCATGTGTAACACAGTGGGGGAGCCATCAAGCAGCCATGCAAAATATTAAGAAATACCATCAATGAAGAGGCTGCAGAAATGGATATACATCCAGTACATGCAGGCGCTTAATCCAG
TAGCCCAATCAGACAACCAAGGCAAGTGCATAGCAGGAACCTACTAGCTTCCAGGACAAATAGCATGGATGATGACAAGTAAACCCACTGTCCAGTAGGACATATAAAAGATTTGCGGATTAATAAATAGTAAAGAAATGTATAG
CCCCACAGCATTTCTGGACATAAAAACAGGGCCAAAAGAACCCCTTAGAGACTATGTAGACCGGTTCTTTAAAACCTTAAAGAGCCGAAACAGCTACACAAGATGTAAAAAATTTGATGACAGACACTTGTGGTCCAAAATGCCAACCCAGATTTAG
ACCAATCTTAAGAGATTTAGCCAGAGGCTCTCAATAGAGAAATATGTAGACAGATCTCAGGAGTGGGAGGACCTAGTCAATAAGCAAGATGTTGGCTGAGGCAATGACCAAAAACAAAATGACATAATGTGCAGAGAACAATTTAAAGGCCCTA
AARAAATTTGTAAATTTTCAACTGTGGCAAAAGGCAACATGATGCGAGAAATGCGAGGCCCTAGGAAAAGGGCTTTGAAATGTGGA AAAAGAAAGCAACCACTGAAAAGATTTGACTGAGAGACAGCTTAATTTTTAGGAAAATCTGGCCCTC
CCACAAGGAAAGGCCAGGGAATTTCTTCAGAGCAGACCAGGCCAACAGCCCCACAGAGGAGCCTCAGGCTTGGGGGAGAGACAACAACCCAGCTCAGAAGCAGGAGTCAACAGACAAGGAACTATATCCCTTAAACCTCCCTCAGATCACTCTTT
GGCAACGACCCCTCGTCAAAATAAAGATAGGGGGGCAATTAAGGAAAGCTCTATTAGATACAGGAGCAGATGATACAGTATTAGAAGACATGAATTTGCCAGGGAATGGAAACAAAATGATAGGGGGAATTTGAGGCTTTTATCAAAGTAAAGCAGTA
TGAAACAAGTACCATAAGAAATCTGTGGACACAAGCTATGGGTACAGTATTAGTGGGACCTACACCTGTCAACATAATTTGGGAGAAATCTGTTGACTCAGCTTGGTGCATTTAAATTTTCCAATTAGTCCCAATGAACTTACCAGTAAATTAAG
CCAGGAATGGATGGCCAAAAGTTTAAACAATGGCCATGTACAGAAGAAAAAATAAAGCATTAACAGAAATTTGTAATGAAATGGAGAAGGAAGAAAAATTAACAAAAATTTGGCCCTGAAAATCCATATAAACAATCCAATATTTGCCATAAAAAAGAA
ACAGTACTAAGTGGAGGAAATTAGTAGATTTTCAGGGAACCTCAATAAAAGAACTCAAGACTTTTGGGAAAGTCAATTAGGAATACCACACCCAGCGGTTAAAAAAGAAAAAATCAGTGACAGTACTAGTGTGGGGGATGCATATTTTCAGTTCCTTT
ATATGAAGAATTCAGGAAATATACTGCATTCACCCATACCTAGTATAAAATTAAGAACCCAGGGATCAGGTATCAATCAATGTTCCACAGGGGTGAAAAGGATCACCCGCAATATTTCCAAAGTAGCATGACAAAAATCTTAGAGCCTTTCAGAAA
CAAAAATCCAGACATAGTTATCTATCAATACATGGATAACTGTGATAGCCACTGGATGAGGAAAGGAAAGGAAAGGAAAGGAAAGGAAAGGAAAGGAAAGGAAAGGAAAGGAAAGGAAAGGAAAGGAAAGGAAAGGAAAGGAAAGGAAAGGAAAG
TTCTTTGGATGGGGTATGAACTCCATCCTGCACAAATGGACAGT
TAAACTCCCTTAGGGGACCAAAAGCACTAACAGACATAGTACCAG
CAGGACCAATGGACATATCAAAATTTACCAAGAACCATTCAAAA
CTAAATTTAGATTACCCATTCAAAAAGAAACATGGGAAACATGG
CTATGTAGATGGAGCAGCTAATAGGAAACATAAGGAAAG
AACATAGTAACAGACTCACAGTATGCATTAGGGATCATTCAAG
AAGTAGATAAATAGTAAAGTAGTGGAAATCAGAAAAAGTGCATTT
TGATCAATGTCACTAAAAGGAGAAGCCATGCATGACAAGTAG
GGACAAGAAACAGCATACTTTTACTAAAATAGCAGGAAGATG
GTCAGGAGTAGTAGAATCCATGAATAAAGAAATTAAGAAAAAT
AATAGATAAATAGCAACAGACATAACAAGCTAAAGAAATTAACA
AACAGTGCATAAAGGTAGTACCAGAGGAAAGCAAAAATCAT
AAATGGGTTATAGACATATTAAGAAAGCAGACATAAAGAGT
GAATAGAAAAATATACCACAAAATAGAACCCTGGCCTGGCAGT
AGGATCTCTACAATCTTGGCACTGCAGCATTGATAAAACCA
TAGAAGAGCTCAAGCAAGAGCTGTGACAGACTTTCTTAGACC
TCAGCATAGCAGAATAGGCATTTTGAGACAGAGAAGAAACAAG
AAAAAAGGCTTAGGCATCTTCTATGGCAGGAAGAAAGCAAG
GCAATAATAGCAATAATTTGTGTGGACCATAGTATTTATAGAA
AGATGGGGACCATGCTCCTTGGAAATTTGATGATCTGTAGGG
GGCTACACATGCTGTACCTACAGCCCAAGCCCAAGAA
GACCCACTCTGTGCTACTTTAGAATGTACAGATGCTAAGAAATG
GGATGTCACTTTTTATAGACTGTATGTAGTACCCTGGTGAGAG
CTCCAGCTGGTATGCACTTCTAAAGTGTAAATAAAGACATTTA
TGACTGTGAAATCTGACAAAACATGCCAAAATAATAGTCAAT
AGACAAGCATATTTGATCATTAATGAAAGTCAATGAAATGATA
AATTTTTCTATGTGTAATACAAAATGTTTAAATAGTACATAC
GTAATGGCCCTCCATTCAGGAAAACATAAACACTAATCAATCA
AAATATAAAGTGGTAGAAATTAAGCCATTAGGAGTAGCTCCC
CGTACAGGCGACAGCAATTTGCTGTCGGTATAGTGCACAAAG
GCTCTTAGGCTTTGGGGCTGCTCAGGAAAACCTCATCTGCAC
TTGATTTGAAGATCGCAAAACCAGCAGGAAAGGAATGAAAGG
TTTTTTGCTGCTTTCTATAGTAAATAGAGTTAGGAGGATAC
ATTTTAGTCTTTGCTGGGACGATTCGGCAACCTGTGCCTCT
AGCCTTGTGCACTACTGGGGTCAGGAGCTAAAAAGAGTGTAT
TGAAAGCAGCTTGCATAAATTTGGGGGCAAGTGGTCAAAAG
TGAAAGACATGGAGCAATCACAAGTAGCAACTACATAATG
TTTTAAAGAAAAGGGGGACTGGAAGGTTAATTTACTCTAAG
AAGCTAGTACCAGTGCACCAAGGAAAGTGAAGAGACCAAG
AACACATCCGGAATTTACAAGAACTGTGCACAGAGGAACT
TCTAGTTGGACCAGATCTGAGCTGGGAGCTCTCTGGCTATC

```

```
AGCCACTGGATTCCTGATGGATGGATTTGTTAA
AGAGGAAGGAAAAGGTTGTTTCCCTAACTGAA
AATCAGAGTGTAGTTAACCATAATAGAACAAT
AGCTCAAGAAGAACATGAAAGATATCACAGCAA
TGGCAATTAGATTGTACCCATTTAGAAGGAAA
ATACAGACATAATGGTAGTAACTTCCAGGTG
TCAAGCTGAGCACCTTAAGACAGCAGTACAAT
CACAAATTTCCGGTTTATTACAGAGCAGCAGA
AGATGGCAGGTGTGATTGTGTGGCAGTAGC
TCCAGTAGGAGAGGCTAAATTAGTAATAAAA
TTATTTTGATTGTTTTGCAGACTCTGCCATAAG
```

```
TTGGGTGGGGATTACACACCAGACAAGAAACATCAGAAAAGAACCTCCAT
AGTTAACTGGGCAAGTCAAGTTTATCCTGGAATTAAGTAAGGCAACTTT
GGTATACATGACCCATAAAGACTGTATAGCTGAAATACAGAAACAGGGG
GAGGCTGTGCAGAAAATAGCCATGGAAGCATAGTAATGGGGTAGAGACC
TAAAATTTAGTACAGCTGGAAGAAAGATCCCATAGTAGGAGCAGAACTT
GACGAAATGCAAGCAATTTGATATAGCTTTGCAAGATCAGGTACAGGAT
AGAGTCTACCTGTATGGGTACCAGCATAAAGGAAATGGAGGAAATGAAC
GTAGTACTTAACTGCCCACCATAGTAGCAAAAAGAAATAGTGGCTAGCTG
AGTCCATGTAGCCAGTGGCCATGGAAGCAGAAAGTATCCAGAGGAAAAC
TGTGGTGGCAGGATTTCAACAGAAATTTGGAATTCCTACAATCCCAAA
ACAAATTTAAAGAAAAGGGGGGATTTGGGGGATCAGTGCAGGGGAAAAGAT
AGCACCCAGCACAACCTACCTGGAAGGAGGAGGAGGAGGAGGAGGAGGAG
ACATGGAATAGTTTATGAAAACACCATATGTAATATTTCAAAGAGAGGTAAT
CAAAACAGGAAAAGAGATTTGGCATTTGGGCTACAGGATCCATAGAATGGA
GACACATAGTATTTCTAGGTGTACTATCAAGCAGGACATAAATAAAGTGT
BAAGATCAGGGGCCGACAGGGAAACCATCAATGAATGGCACTAGAGCTTC
CCATAATAAGAAATCTGCAACAACCTGCTGTTTATTCATTTCAAGATTTGG
TTAAATTTGTGCATCAGATGCCAAAGCATATGAGACAGAGGATACATAAGT
ATGAAGATGTAATCAGTTTATGGGACCAAGGCTTAAAGCCATGTGTAAGT
TAAATGTAATGCAATCTTAACTATTTCTAGCAATAGTAGCCTTAGTAGAGCA
GACAGTGGCAATGAGAGTGTAGGGGATCAGGAAGAAATGGCGGCATTTATGG
TTATTTTGTGCATCAGATGCCAAAGCATATGAGACAGAGGATACATAAGTGT
ATGAAGATGTAATCAGTTTATGGGACCAAGGCTTAAAGCCATGTGTAAGT
```

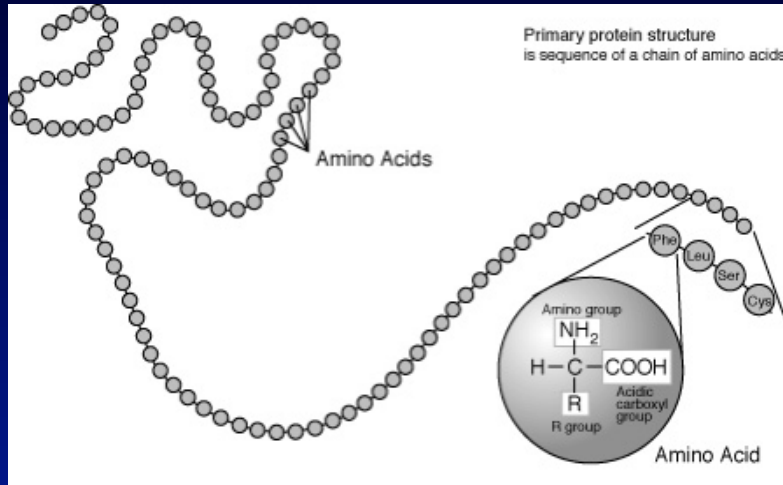
DNA --> RNA --> protein



Standard Genetic Code

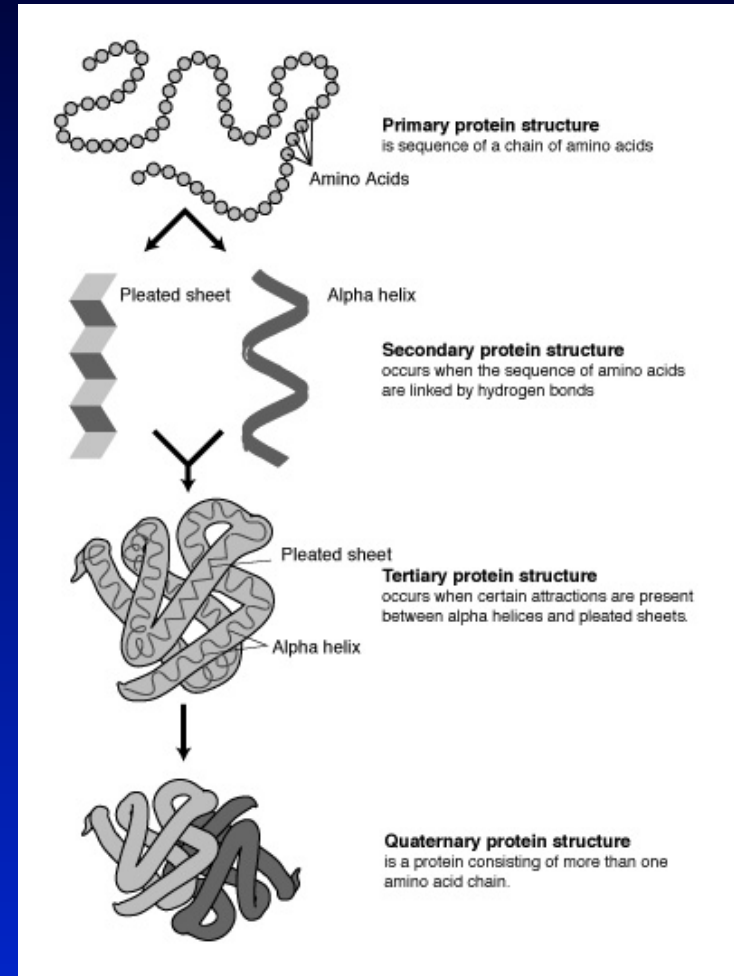
	T	C	A	G
T	TTT Phe F	TCT Ser S	TAT Tyr Y	TGT Cys C
T	TTC Phe F	TCC Ser S	TAC Tyr Y	TGC Cys C
	TTA Leu L	TCA Ser S	TAA Och *	TGA Opa *
	TTG Leu L	TCG Ser S	TAG Amb *	TGG Trp W
C	CTT Leu L	CCT Pro P	CAT His H	CGT Arg R
C	CTC Leu L	CCC Pro P	CAC His H	CGC Arg R
	CTA Leu L	CCA Pro P	CAA Gln Q	CGA Arg R
	CTG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

The screenshot shows the NCBI website interface. At the top, the browser address bar displays 'http://www.ncbi.nlm.nih.gov/'. The search bar contains the text 'human globin' and a 'Search' button. Below the search bar, the page is divided into several sections:

- Resources:** A vertical sidebar on the left lists various categories such as 'NCBI Home', 'All Resources (A-Z)', 'Literature', 'DNA & RNA', 'Proteins', 'Sequence Analysis', 'Genes & Expression', 'Genomes', 'Maps & Markers', 'Domains & Structures', 'Genetics & Medicine', 'Taxonomy', 'Data & Software', 'Training & Tutorials', 'Homology', 'Small Molecules', and 'Variation'.
- Welcome to NCBI:** A central section with the text: 'The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.' Below this is a link: 'More about the NCBI | Mission | Organization | Research | RSS'.
- Genome Reference Consortium:** A section with a blue background and white text: 'Formed to improve human and mouse reference assemblies, GRC will fix loci misrepresented in reference assembly, fill remaining gaps, and make alternate representations of complex loci.' Below the text are four numbered tabs (1, 2, 3, 4), with '2' being the active tab.
- How To...:** A section with a list of links: 'Obtain the full text of an article', 'Retrieve all sequences for an organism or taxon', 'Find a homolog for a gene in another organism', 'Find genes associated with a phenotype or disease', 'Design PCR primers and check them for specificity', 'Find the function of a gene or gene product', and 'Determine conserved synteny between the genomes of two organisms'. Below the list is a link: 'See all ...'.
- NLM/NCBI H1N1 Flu Resources:** A section with a blue background and white text, partially visible at the bottom.
- Popular Resources:** A section on the right with a list of links: 'PubMed', 'PubMed Central', 'Bookshelf', 'BLAST', 'Gene', 'Nucleotide', 'Protein', 'GEO', 'Conserved Domains', 'Structure', and 'PubChem'.
- NCBI News:** A section on the right with a list of news items: 'November and October News' (02 Dec 2009), 'NCBI News - September 2009' (05 Oct 2009), and 'NCBI News - August 2009' (19 Aug 2009). Each item has a brief description and a 'More...' link.

NCBI databases: Genbank feature table

Nucleotide - Homo sapiens hemoglobin, gamma A (HBG1), mRNA
http://www.ncbi.nlm.nih.gov/nuccore/28302130?report=genbank&log\$=seqview&from=54&to=497

Entrez Gene record to access additional publications.
COMPLETENESS: full length.

FEATURES	Location/Qualifiers
source	1..444 /organism="Homo sapiens" /mol_type="mRNA" /db_xref="taxon:9606" /chromosome="11" /map="11p15.5"
gene	<1..>444 /gene="HBG1" /gene_synonym="HBGA; HBGR; HSGGL1; PRO2979" /note="hemoglobin, gamma A" /db_xref="GeneID:3047" /db_xref="HGNC:4831" /db_xref="HPRD:00789" /db_xref="MIM:142200"
exon	<1..92 /gene="HBG1" /gene_synonym="HBGA; HBGR; HSGGL1; PRO2979" /inference="alignment:Splign"
CDS	/number=1 1..444 /gene="HBG1" /gene_synonym="HBGA; HBGR; HSGGL1; PRO2979" /note="hemoglobin gamma-a chain; hemoglobin, gamma, regulator of; gamma globin; gamma A hemoglobin" /codon_start=1 /product="A-gamma globin" /protein_id="NP_000550.2" /db_xref="GI:28302131" /db_xref="CCDS:CCDS7254.1" /db_xref="GeneID:3047" /db_xref="HGNC:4831" /db_xref="HPRD:00789" /db_xref="MIM:142200" /translation="MGHFTEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFPD SFGNLSASAIMGNPKVKAHGKKVLTSLGDATKELDDLKGTFAQLSELHCDKLLIVDPE NFKLLGNVLTIVLAIHFCKEFTPEVQASWQKMTAVASALSSRYH"
exon	93..315 /gene="HBG1" /gene_synonym="HBGA; HBGR; HSGGL1; PRO2979" /inference="alignment:Splign"
STS	/number=2 194..369 /gene="HBG1"

NCBI databases: fasta format

The screenshot shows the NCBI Nucleotide database interface. The browser address bar displays the URL: [http://www.ncbi.nlm.nih.gov/nucleotide/28302130?report=fasta&log\\$=seqview&from=54&to=497](http://www.ncbi.nlm.nih.gov/nucleotide/28302130?report=fasta&log$=seqview&from=54&to=497). The page title is "Nucleotide - Homo sapiens hemoglobin, gamma A (HBG1), mRNA". The search bar contains "Nucleotide" and "for". The format is set to "FASTA". The sequence is displayed in FASTA format, starting with `>gi|28302130:54-497 Homo sapiens hemoglobin, gamma A (HBG1), mRNA` followed by the nucleotide sequence. The "Change Region Shown" section is set to "Selected Region" from base 54 to 497. The "Analyze This Sequence" section includes options for "Run BLAST" and "Pick Primers". The "Articles about the HBG1 gene" section lists several articles related to the gene.

NCBI Reference Sequence: NM_000559.2

Homo sapiens hemoglobin, gamma A (HBG1), mRNA

```
>gi|28302130:54-497 Homo sapiens hemoglobin, gamma A (HBG1), mRNA
ATGGGTCATTTACAGAGGAGGACAAAGGCTACTATCACAAAGCCTGTGGGGCAAGGTGAATGTGGGAAGATG
CTGGAGGAGAAACCTGGGAAGGCTCCTGCTTGTCTACCCATGGACCCAGAGGTTCTTTGCACAGCTTTGG
CAACCTGTCTCTGCTTCTGCCATCATGGGCAACCCAAAGTCAAGGCACATGGCAAGAAGGTGCTGACT
TCCTTGGGAGATGCCACAAGCACCTGGATGATCTCAAGGGCACCTTTGCCAGCTGAGTGAACAGCACT
GTGACAAGCTGCATGTGGATCCTGAGAACTTCAAGCTCCTGGGAAATGTGCTGGTGACCCGTTTTGGCAAT
CCATTTGGGCAAGAATTCAACCCCTGAGGTGCAGGCTTCTTGGCAGAAGATGGTGAAGTGCAGTGGCCAGT
GCCCTGTCTCCAGATACCACTGA
```

Change Region Shown

Whole sequence
 Selected Region
from: 54 to: 497

Customize View

Analyze This Sequence

- Run BLAST
- Pick Primers

Articles about the HBG1 gene

- Molecular analysis of gamma-globin promoters, HS-111 and HS-112 [Hemoglobin. 2009]
- A genome-wide association identified the common genetic variant [Hum Genet. 2009]
- Expression of miR-210 during erythroid differentiation and induction [BMB Rep. 2009]

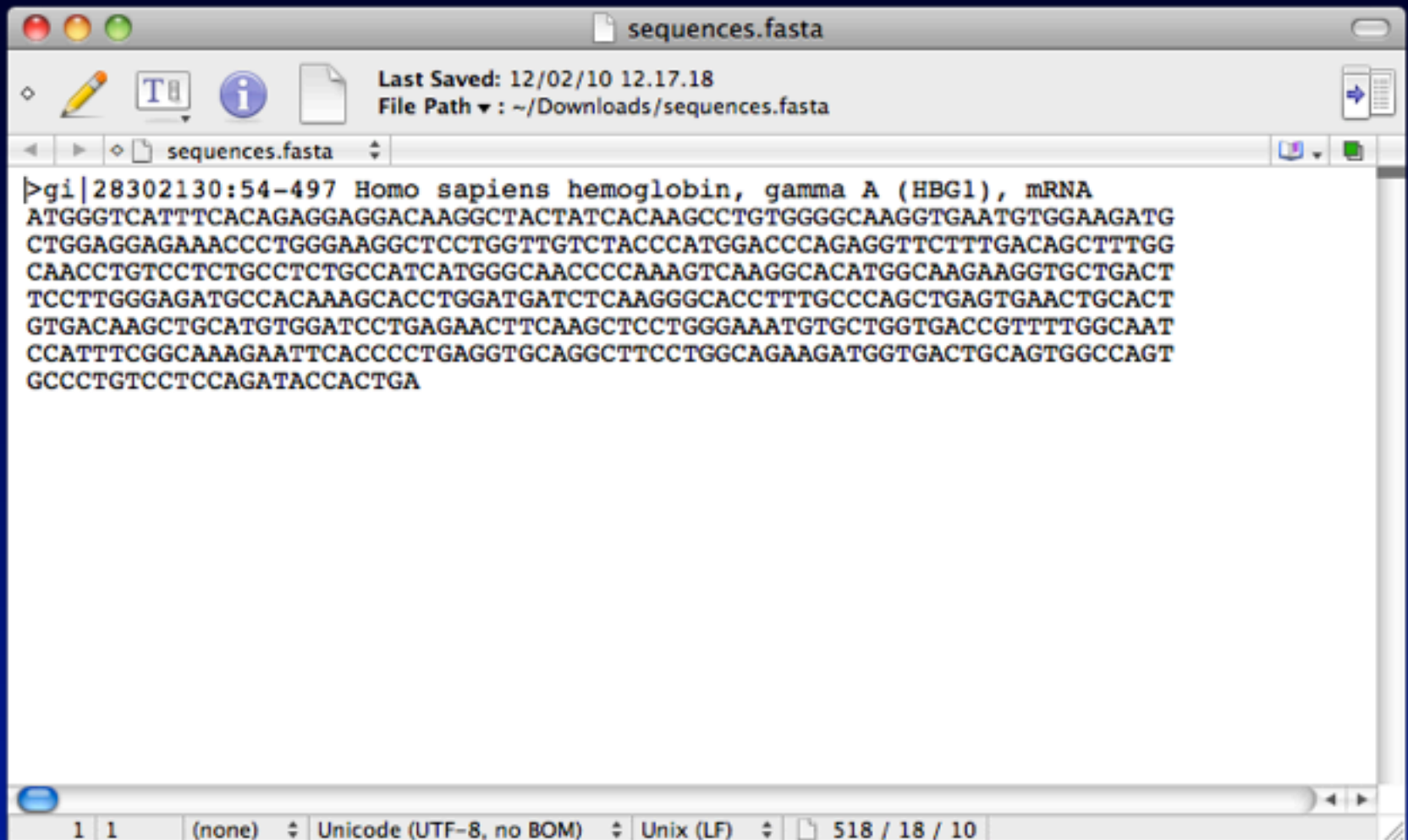
[See all...](#)

[RefSeq Protein Product](#)
See the reference protein sequence for A-gamma globin (NP_000550.2).

[More about the HBG1 gene](#)
The gamma-globin gene (HBG1 and HBG2)

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS

FASTA file



```
>gi|28302130:54-497 Homo sapiens hemoglobin, gamma A (HBG1), mRNA
ATGGGTCATTTACAGAGGAGGACAAGGCTACTATCACAAGCCTGTGGGGCAAGGTGAATGTGGAAGATG
CTGGAGGAGAAACCCTGGGAAGGCTCCTGGTTGTCTACCCATGGACCCAGAGGTTCTTTGACAGCTTTGG
CAACCTGTCCCTCTGCCTCTGCCATCATGGGCAACCCCAAAGTCAAGGCACATGGCAAGAAGGTGCTGACT
TCCTTGGGAGATGCCACAAAGCACCTGGATGATCTCAAGGGCACCTTTGCCAGCTGAGTGAAGTGCAGT
GTGACAAGCTGCATGTGGATCCTGAGAACTTCAAGCTCCTGGGAAATGTGCTGGTGACCGTTTTGGCAAT
CCATTTTCGGCAAAGAATTCACCCCTGAGGTGCAGGCTTCTGGCAGAAGATGGTGACTGCAGTGGCCAGT
GCCCTGTCCCTCCAGATACTACTGA
```