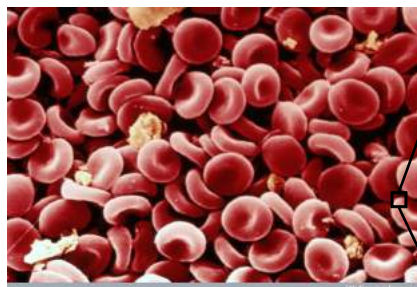


Protein Evolution

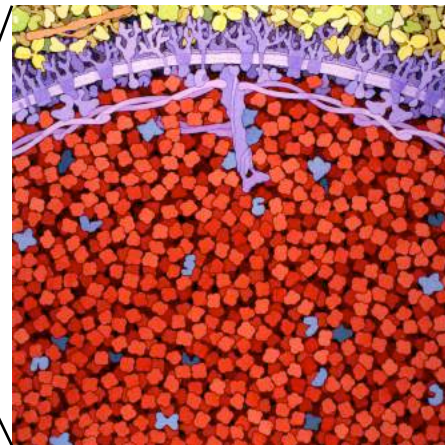
Lecture 1

Lucy Colwell, ICTS Winter School on
Quantitative Systems Biology, Dec 12 2017

Red blood cells

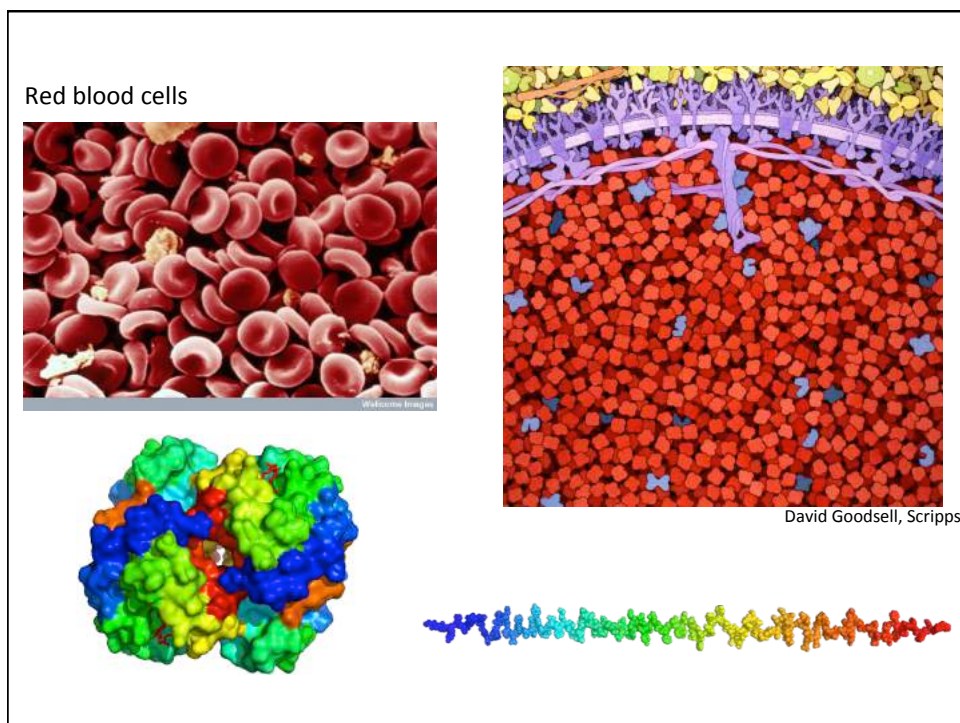
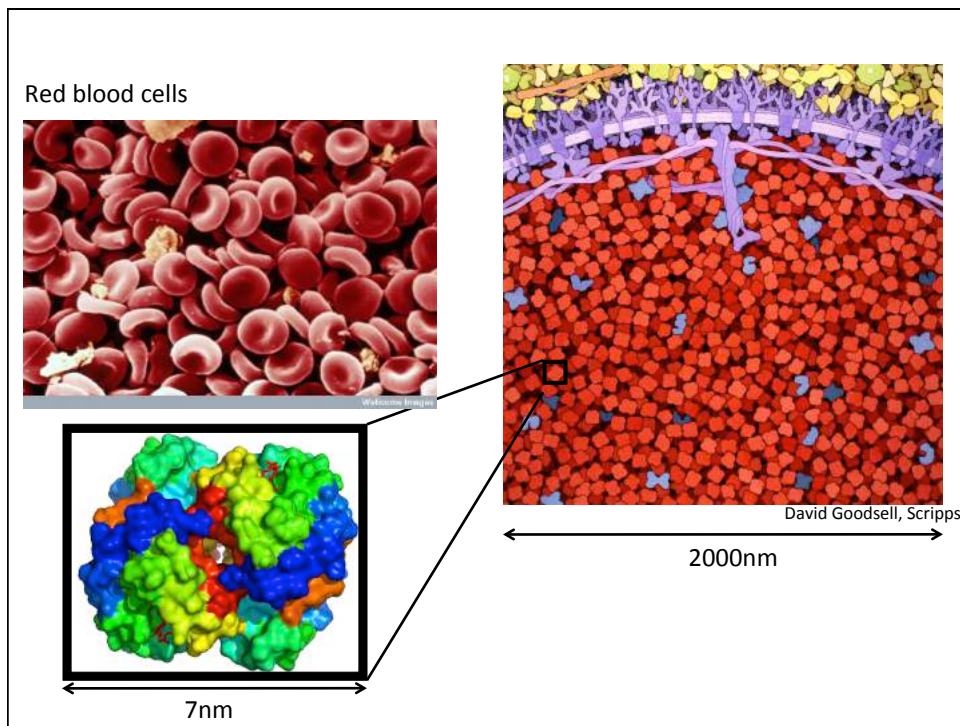


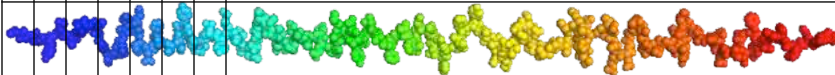
60 μ m



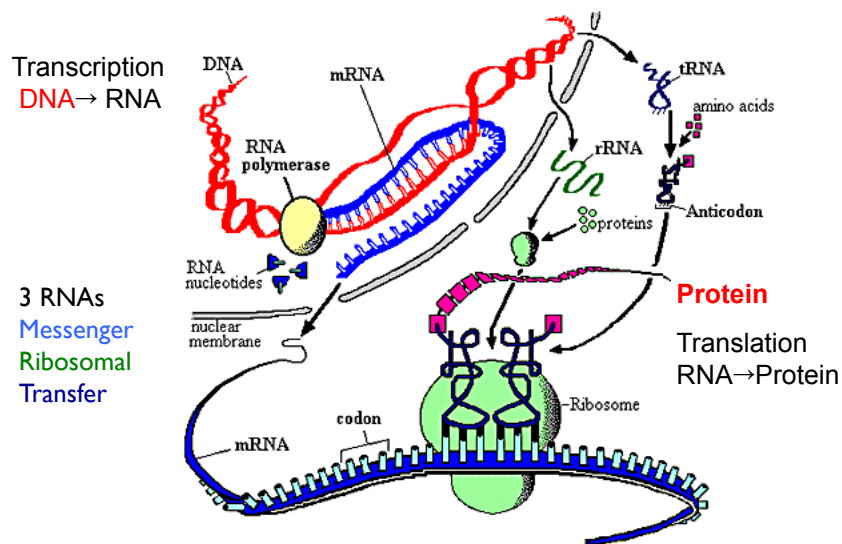
David Goodsell, Scripps

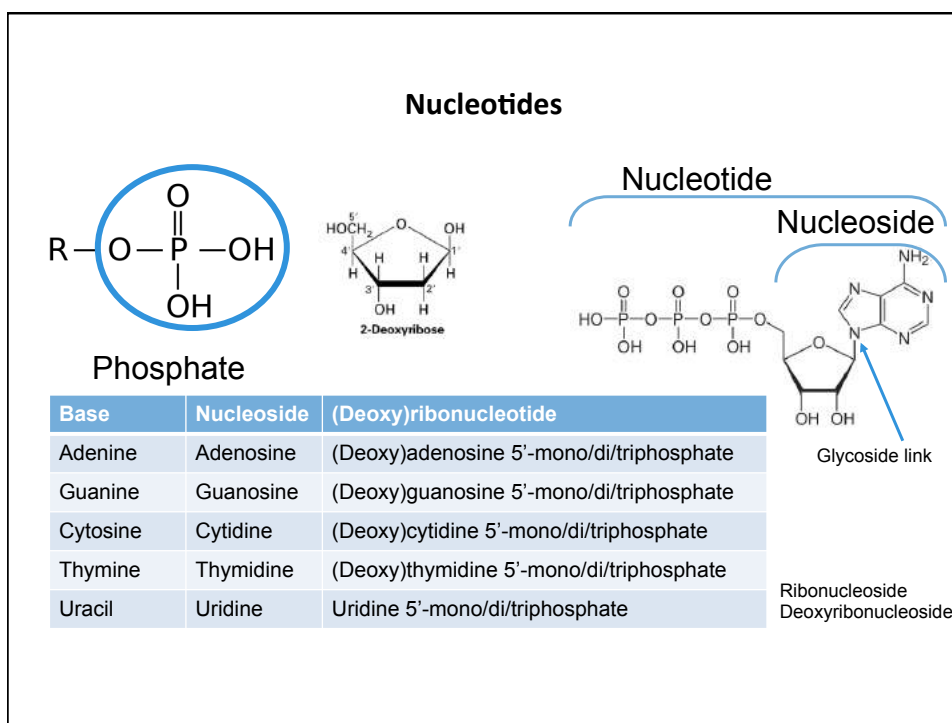
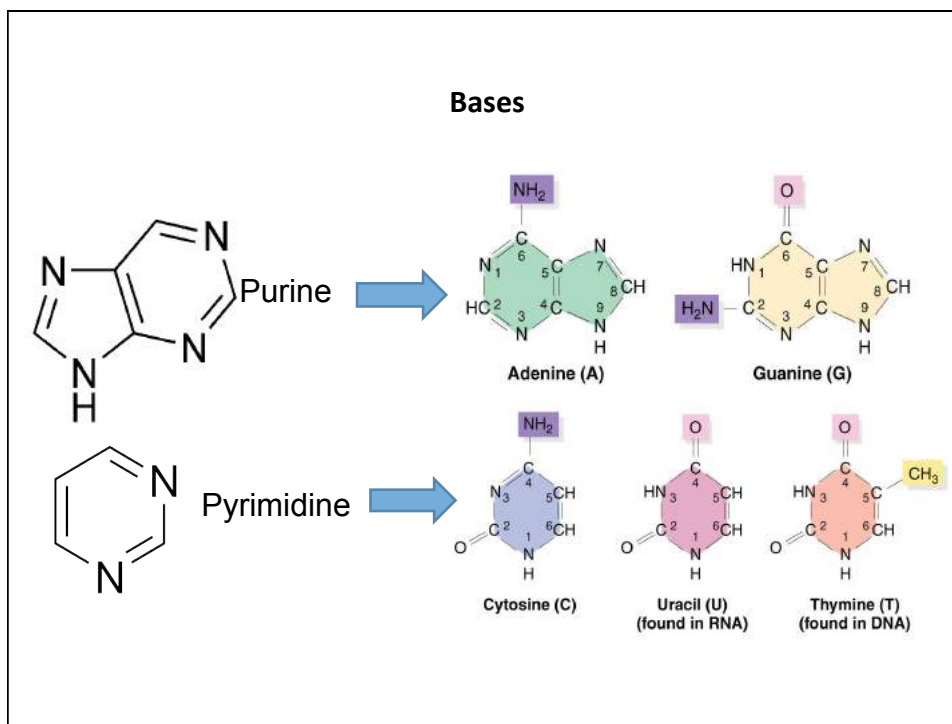
2 μ m

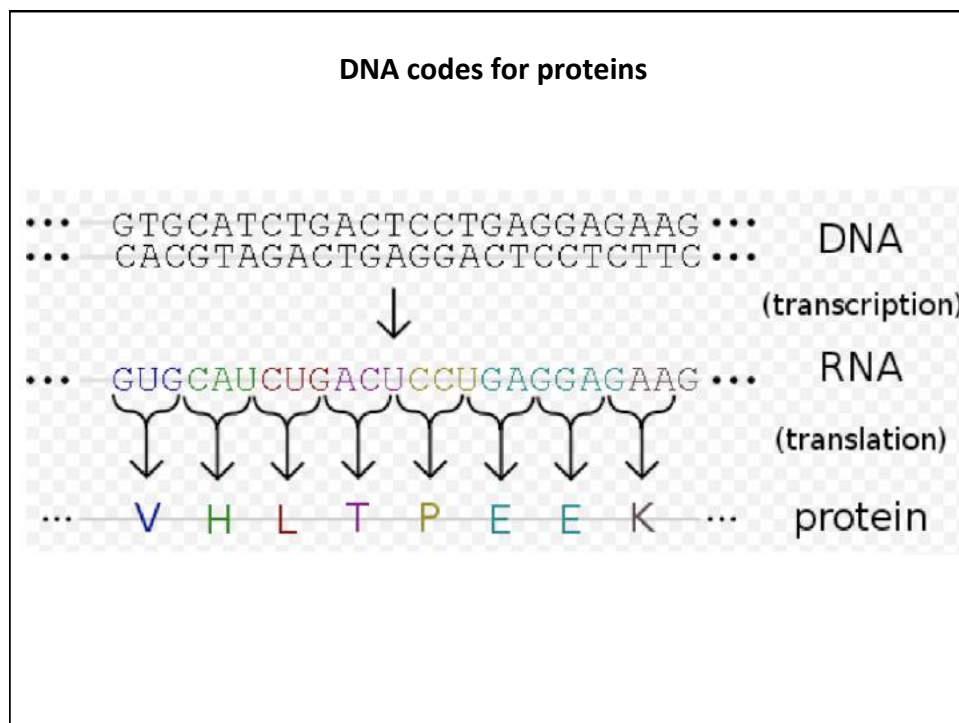
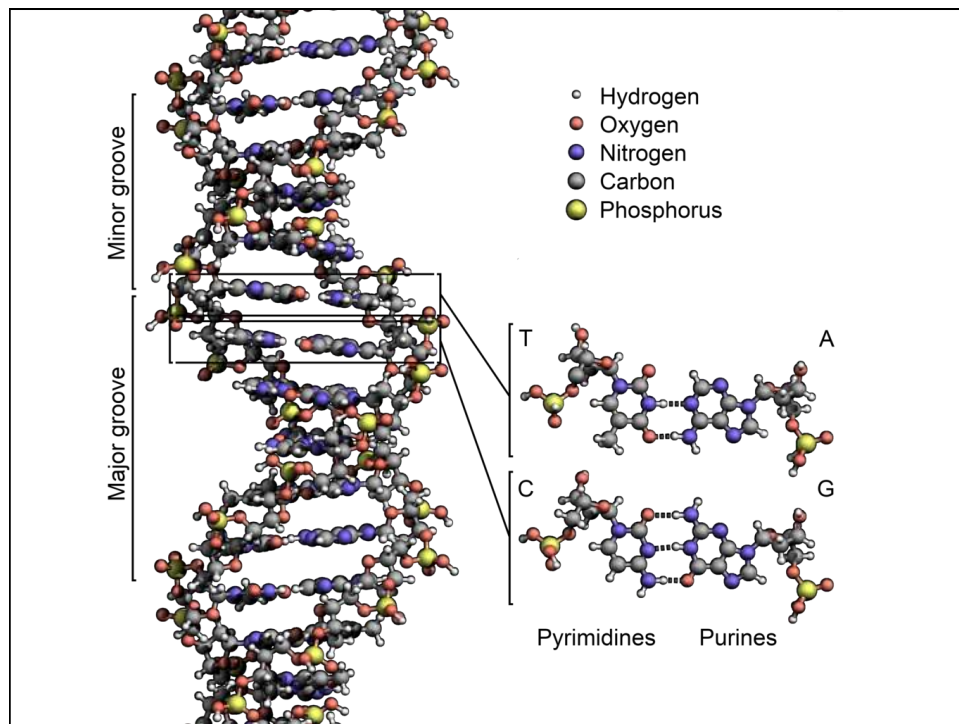


1	2	3	4	5	6	7	8
M	V	L	S	P	A	D	K
							

Central Dogma: DNA to RNA to Protein



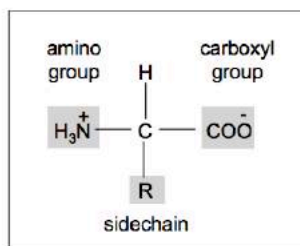
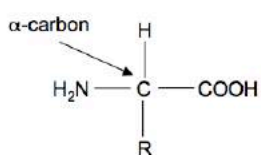




DNA codes for proteins

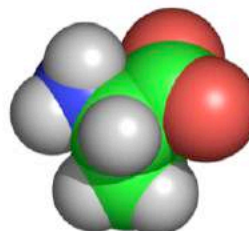
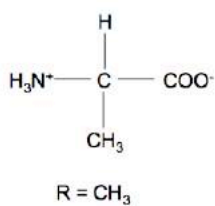
		Second base				
		U	C	A	G	
First base	U	UUU } Phenyl-alanine F UUC } UUA } Leucine L UUG }	UCU } Serine S UCC } UCA } UCG }	UAU } Tyrosine Y UAC } UAA } Stop codon UAG } Stop codon	UGU } Cysteine C UGC } UGA } Stop codon UGG } Tryptophan W	Third base U C A G U C A G U C A G U C A G
	C	CUU } Leucine L CUC } CUA } CUG }	CCU } Proline P CCC } CCA } CCG }	CAU } Histidine H CAC } CAA } Glutamine Q CAG }	CGU } Arginine R CGC } CGA } CGG }	
	A	AUU } Isoleucine I AUC } AUA } AUG } Methionine start codon M	ACU } Threonine T ACC } ACA } ACG }	AAU } Asparagine N AAC } AAA } Lysine K AAG }	AGU } Serine S AGC } AGA } Arginine R AGG }	
	G	GUU } Valine V GUC } GUA } GUG }	GCU } Alanine A GCC } GCA } GCG }	GAU } Aspartic acid D GAC } GAA } Glutamic acid E GAG }	GGU } Glycine G GGC } GGA } GGG }	

Amino acid structure

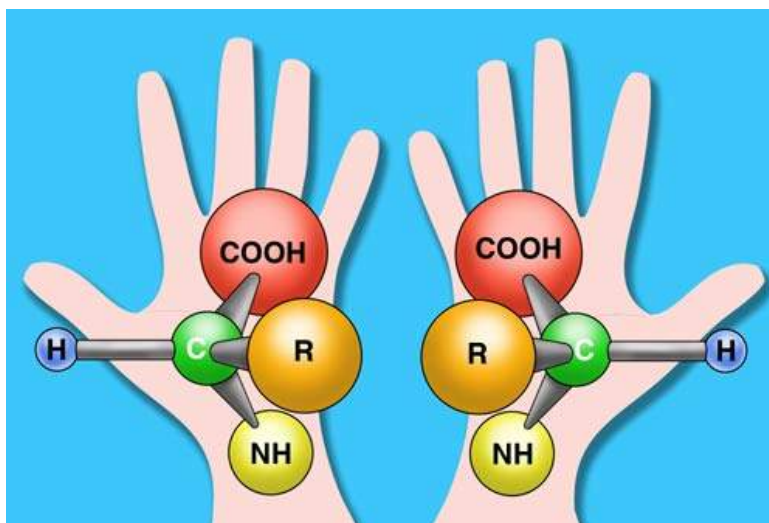


ionized form
at pH 7

e.g Ala

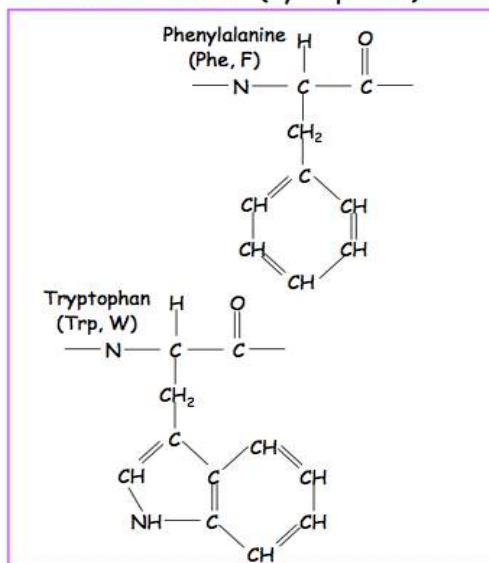


Chirality

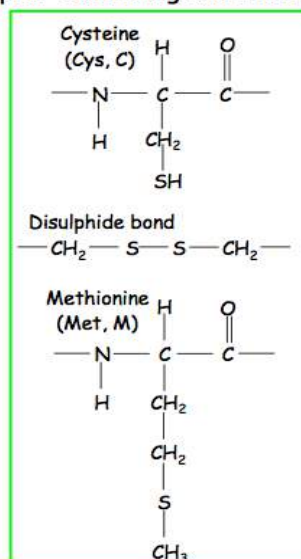


L (or R) form not D (or S)

Aromatic side chains (hydrophobic)

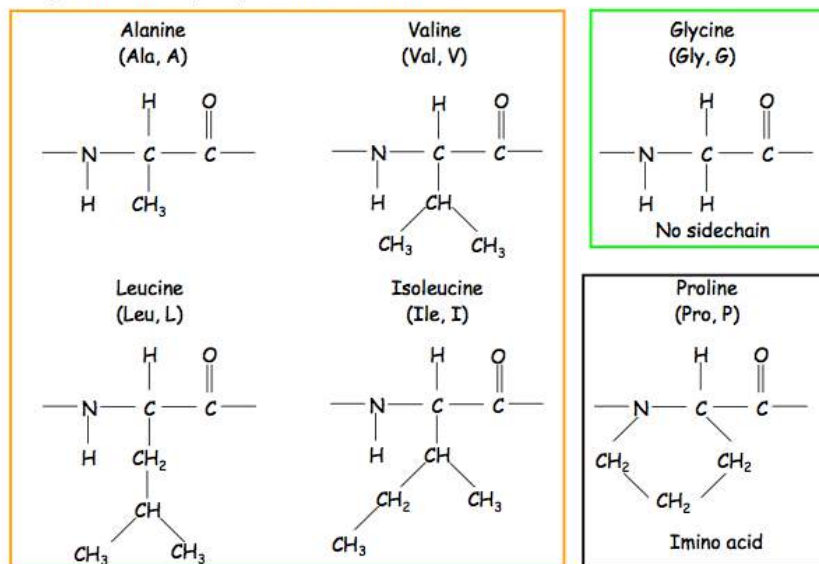


Sulphur-containing side chains



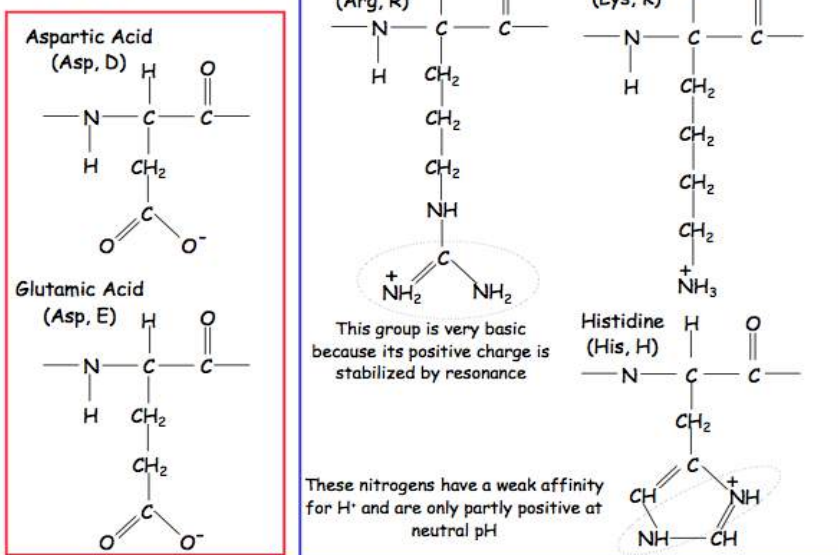
These boxes are coloured according to conventional colouring (eg purple aromatic/hydrophobic)

Aliphatic and hydrophobic side chains



These boxes are coloured according to conventional colouring (eg orange for aliphatic)

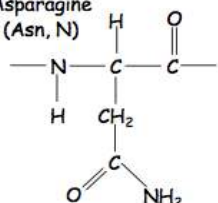
Charged side chains



These boxes are coloured according to conventional colouring (eg red acidic, blue basic)

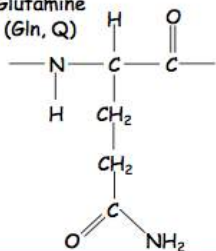
Polar side chains

Asparagine
(Asn, N)

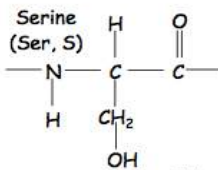


NH₂ is not charged, but is polar

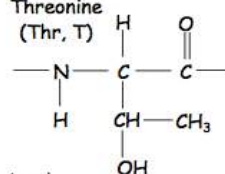
Glutamine
(Gln, Q)



Serine
(Ser, S)

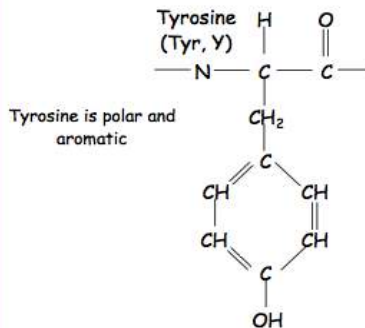


Threonine
(Thr, T)



OH group is polar

Tyrosine
(Tyr, Y)

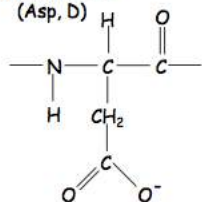


Tyrosine is polar and aromatic

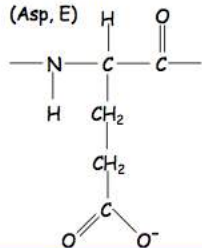
These boxes are coloured according to conventional colouring (eg purple polar, cyan those with hydroxyls -OH)

Charged side chains

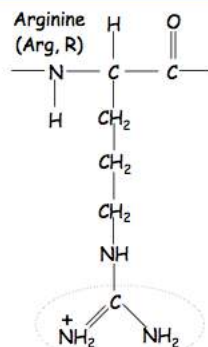
Aspartic Acid
(Asp, D)



Glutamic Acid
(Asp, E)

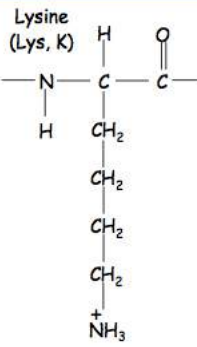


Arginine
(Arg, R)

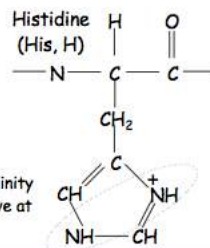


This group is very basic because its positive charge is stabilized by resonance

Lysine
(Lys, K)



Histidine
(His, H)



These nitrogens have a weak affinity for H⁺ and are only partly positive at neutral pH

These boxes are coloured according to conventional colouring (eg red acidic, blue basic)

Amino acid usage in proteins

Based on sequences contained in the UniprotKB/Trembl database:

Highly hydrophobic

Amino Acid	Frequency (%)
Isoleucine	6.00
Valine	6.79
Leucine	9.96
Phenylalanine	4.03
Methionine	2.47

Charged

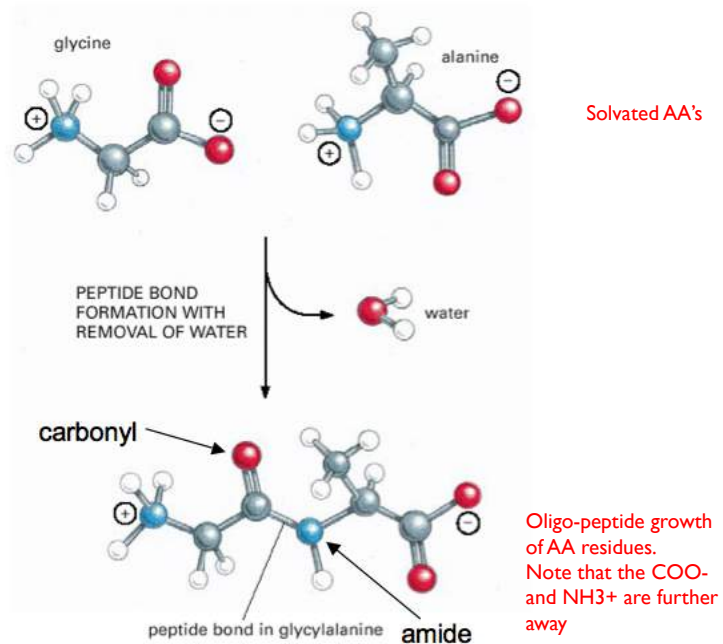
Amino Acid	Frequency (%)
Arginine	5.43
Histidine	2.20
Lysine	5.26
Glutamic Acid	6.19
Aspartic Acid	5.33

Less hydrophobic

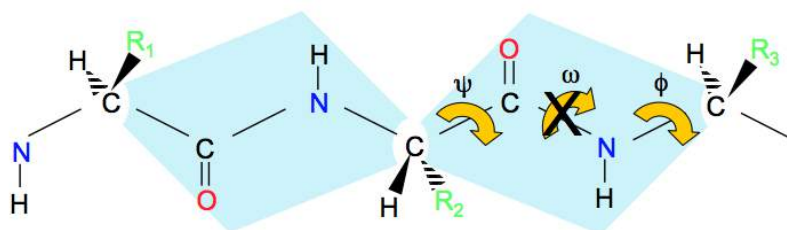
Amino Acid	Frequency (%)
Alanine	8.66
Glycine	7.09
Cysteine	1.23
Tryptophan	1.30
Tyrosine	3.03
Proline	4.65
Threonine	5.55
Serine	6.63

Highly hydrophillic

Amino Acid	Frequency (%)
Asparagine	4.09
Glutamine	3.98



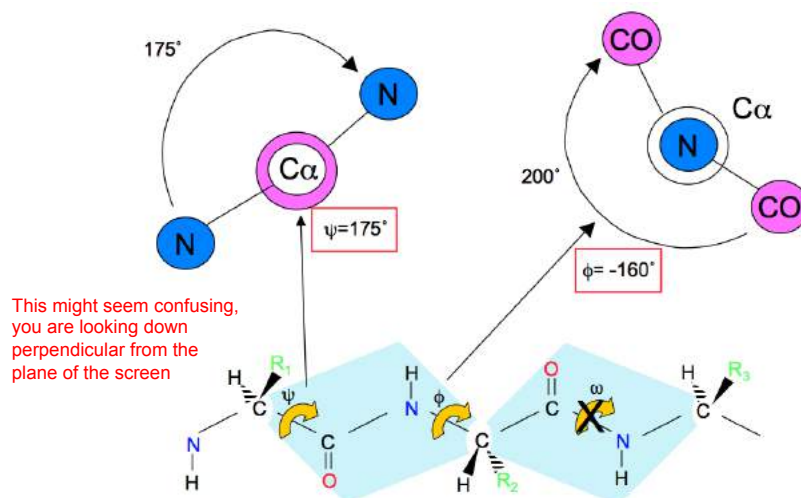
Peptide angles



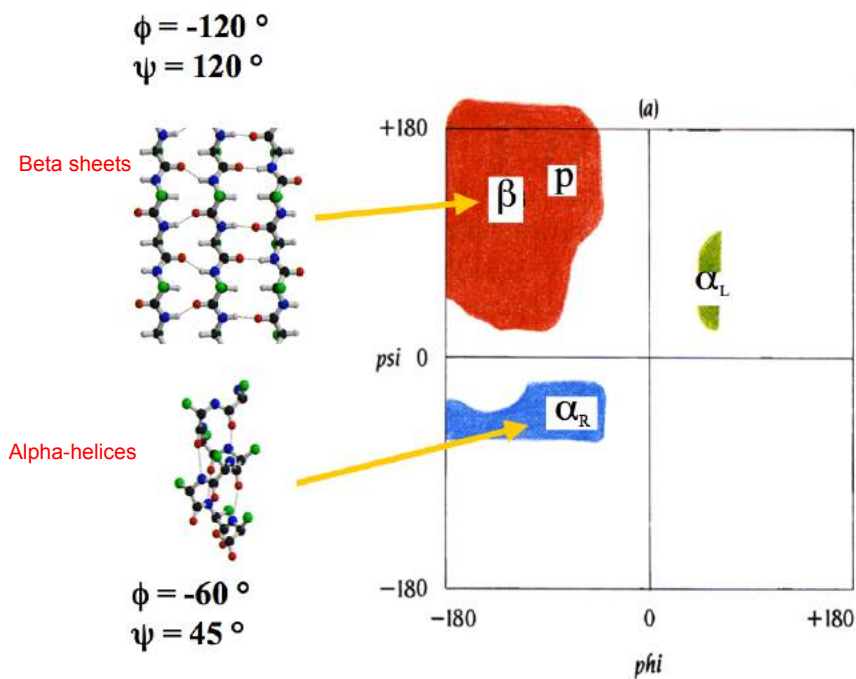
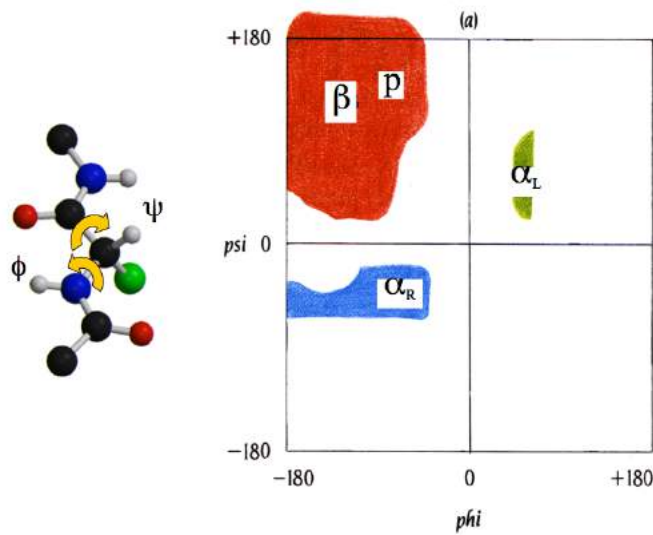
The peptide bond is linear - ω is fixed at 180°
 ϕ and ψ can vary - but not freely!

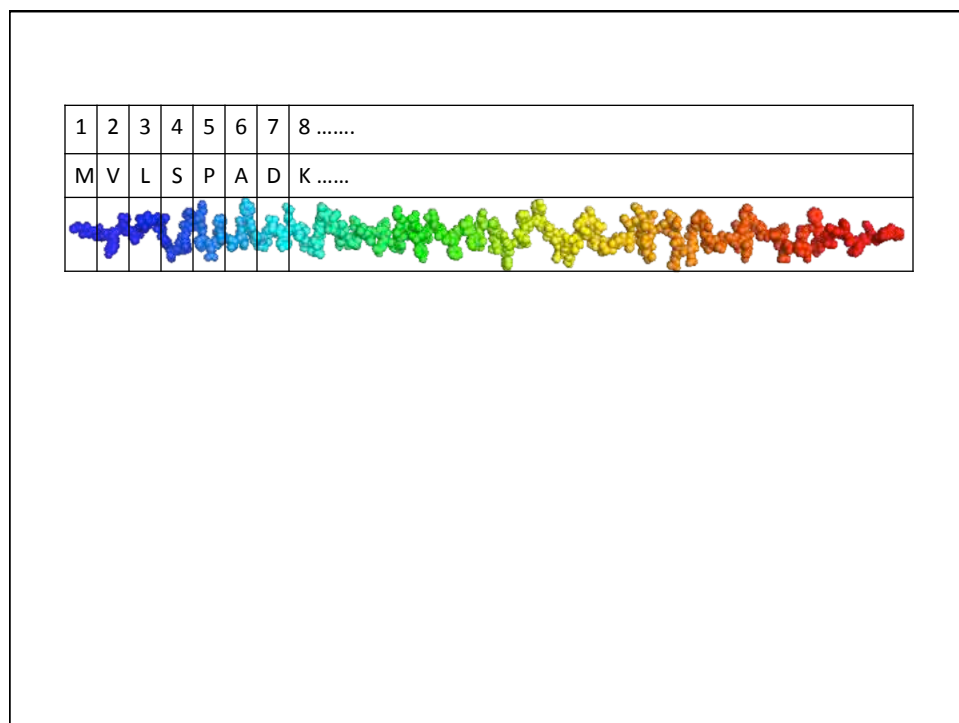
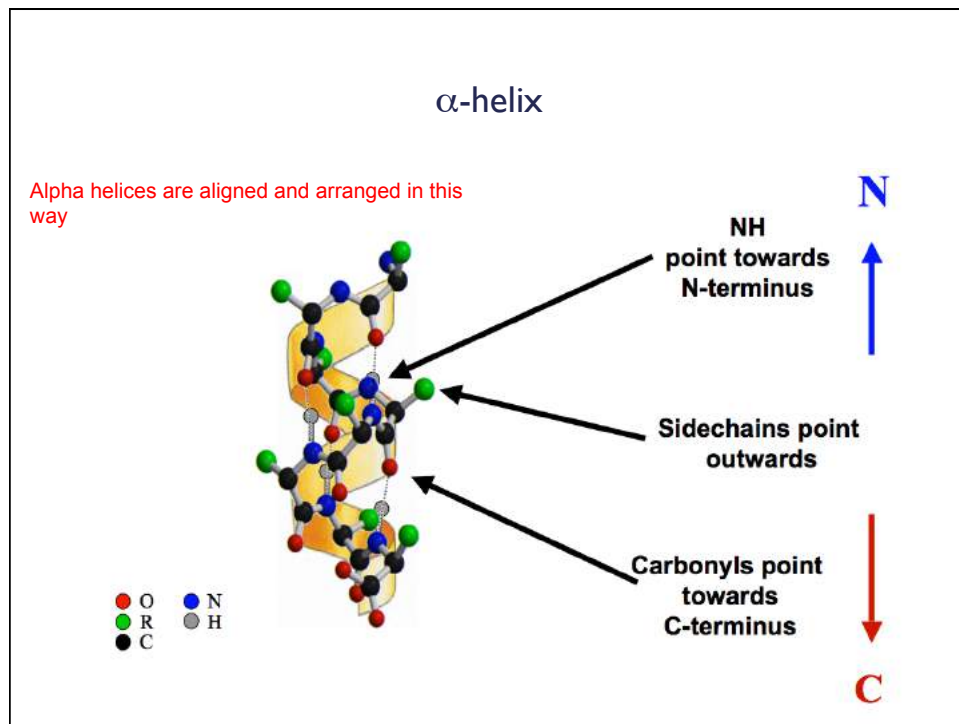
We can define them using Newman projections

Definitions of ϕ and ψ

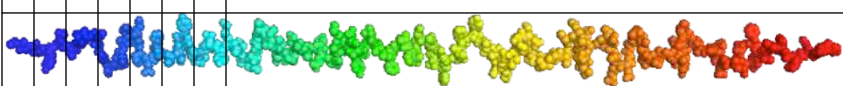


Ramachandran plots





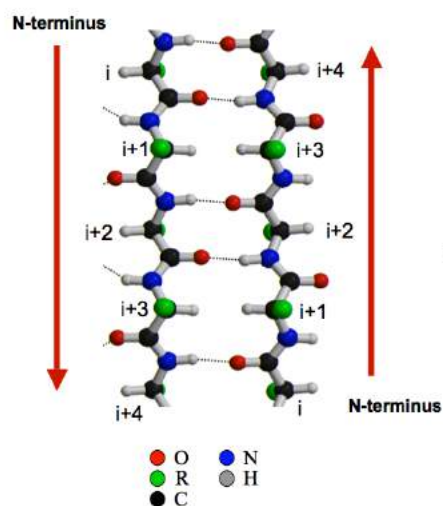
1	2	3	4	5	6	7	8	
M	V	L	S	P	A	D	K	



Polymer self-assembles into
unique 3D structure



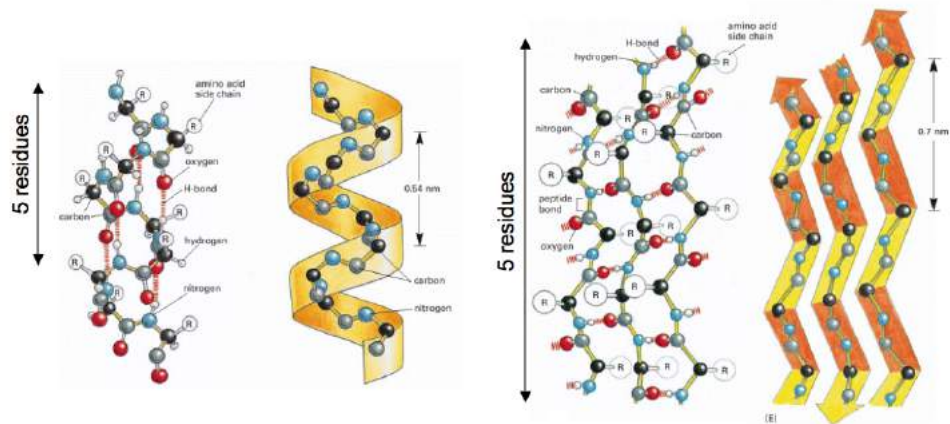
Antiparallel β -sheet



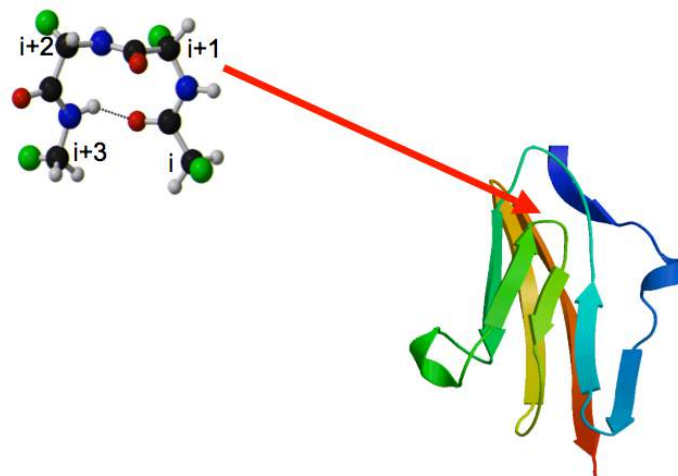
2 hydrogen bonds every other residue
across the sheet

Alternate pairing - residues i, i+2, i+4
are H-bonded across the sheet; i+1, i+3
can form H bonds with other strands

Comparison of α -helix and β -sheet



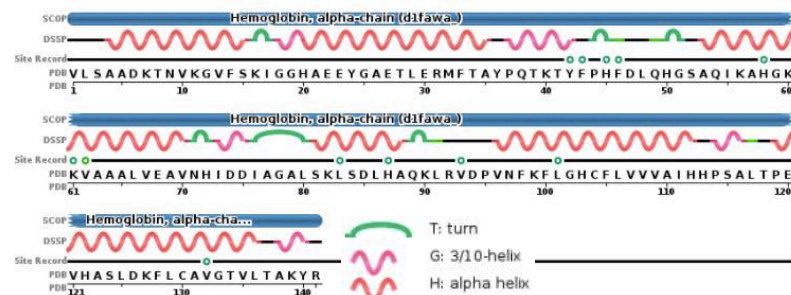
Turns link other features



DSSP classification

We can identify secondary structure from solved 3D protein structures. The following code is used:

- H = alpha helix
- B = residue in isolated beta-bridge
- E = extended strand, participates in beta ladder
- C or L or ' ' = none of the above, coil or loop



We can also predict secondary structure from sequence with good accuracy.

1	2	3	4	5	6	7	8	
M	V	L	S	P	A	D	K	

Polymer self-assembles into unique 3D structure

This is the protein folding problem.



20 July 1973, Volume 181, Number 4096

SCIENCE**Principles that Govern the Folding of Protein Chains**

Christian B. Anfinsen



Anfinsen CB (1973). "Principles that govern the folding of protein chains". Science 181 (4096): 223–230.

20 July 1973, Volume 181, Number 4096

SCIENCE**Principles that Govern the Folding of Protein Chains**

Christian B. Anfinsen

Empirical considerations of the large amount of data now available on correlations between sequence and three-dimensional structure (48), together with an increasing sophistication in the theoretical treatment of the energetics of polypeptide chain folding (49) are beginning to make more realistic the idea of the a priori prediction of protein conformation. It is certain that major advances in the understanding of cellular organization, and of the causes and control of abnormalities in such organization, will occur when we can predict, in advance, the three-dimensional phenotypic consequences of a genetic message.



Anfinsen CB (1973). "Principles that govern the folding of protein chains". Science 181 (4096): 223–230.

Energy and forces

- How do proteins hold together?
- Interplay between entropy (disorder) and enthalpy

$$\Delta G = \Delta H - T\Delta S$$

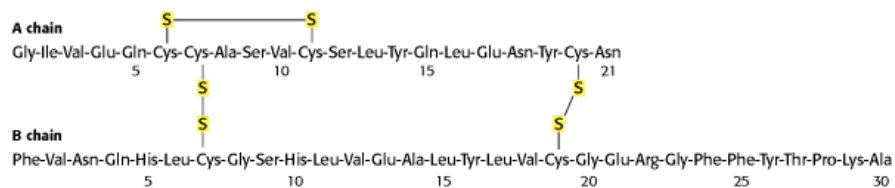
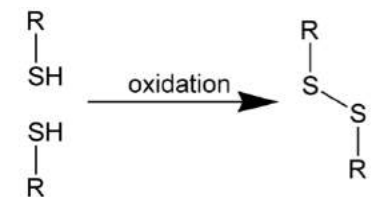
Making bonds is favourable enthalpically

Constraining motion is unfavourable entropically

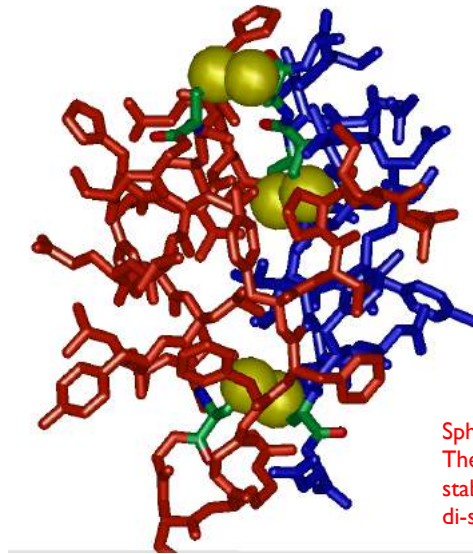
See, http://www.wiley.com/college/pratt/0471393878/student/review/thermodynamics/7_relationship.html
and http://rbaldwin.stanford.edu/PDFs/energetics_of_protein_folding.pdf

Types of interaction

- Covalent
 - Disulphide bonds



Bovine Insulin - 2bn3



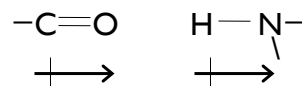
Spheres are Sulphur atoms
The two chains are
stabilised by intermolecular
di-sulphide bridges.

Types of interaction

- Covalent
 - Disulphide bonds
- Ionic
 - Dipole interactions



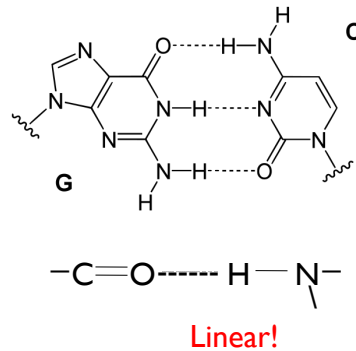
These are called partial charges
 δ^- δ^+



the dipoles are aligned too

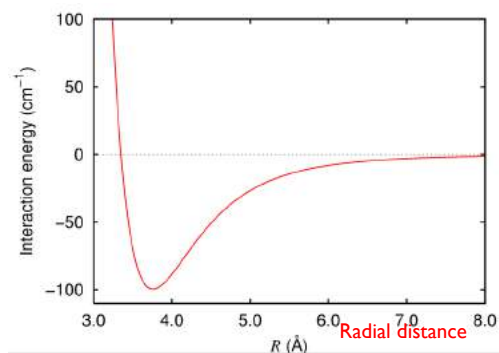
Types of interaction

- Covalent
 - Disulphide bonds
- Ionic
 - Dipole interactions
- Hydrogen bonds



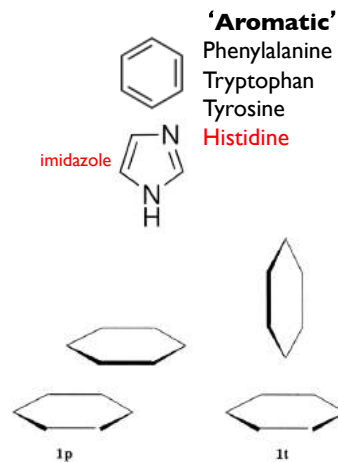
Types of interaction

- Covalent
 - Disulphide bonds
- Ionic
 - Dipole interactions
- Hydrogen bonds
- Van der Waals



Types of interaction

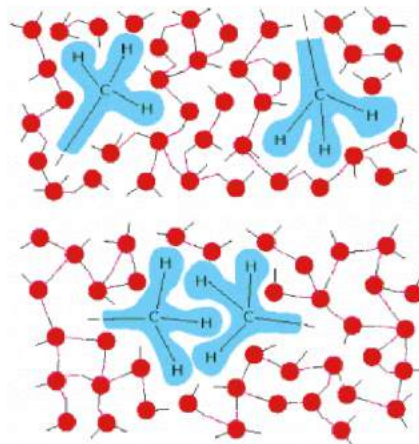
- Covalent
 - Disulphide bonds
- Ionic
 - Dipole interactions
- Hydrogen bonds
- Van der Waals
- π -stacking



Types of interaction

- Covalent
 - Disulphide bonds
- Ionic
 - Dipole interactions
- Hydrogen bonds
- Van der Waals
 - π -stacking
- Hydrophobic interactions

Hydrophobic effect



Displacement due to the mass action of water reducing its exposure to surfaces where it is forced to be ordered.

- Effect not driven by enthalpy
- Driven by entropy - releasing solvent (water)
- Surprisingly powerful



1	2	3	4	5	6	7	8	
M	V	L	S	P	A	D	K	

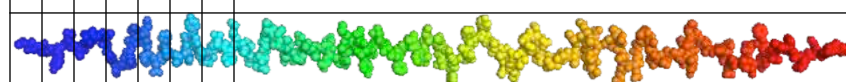
Polymer self-assembles into unique 3D structure

This is the protein folding problem.



Haemoglobin: carries oxygen from the respiratory organs to the rest of the body.

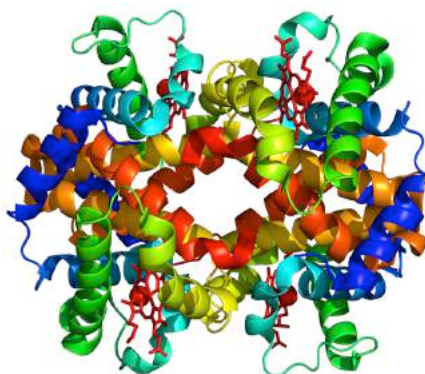
1	2	3	4	5	6	7	8
M	V	L	S	P	A	D	K



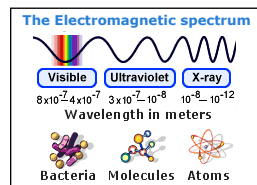
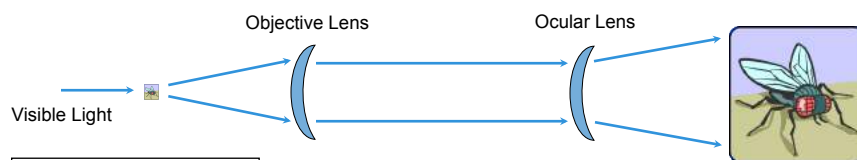
Polymer self-assembles into
unique 3D structure



Four copies interact
specifically to make
a functional complex



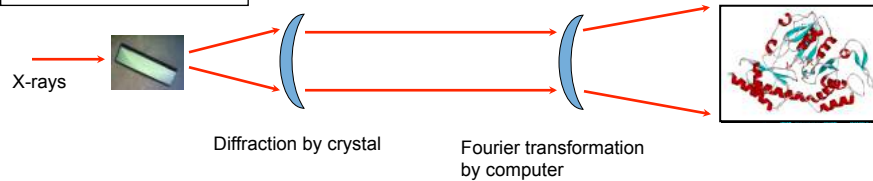
X-ray crystallography simplified



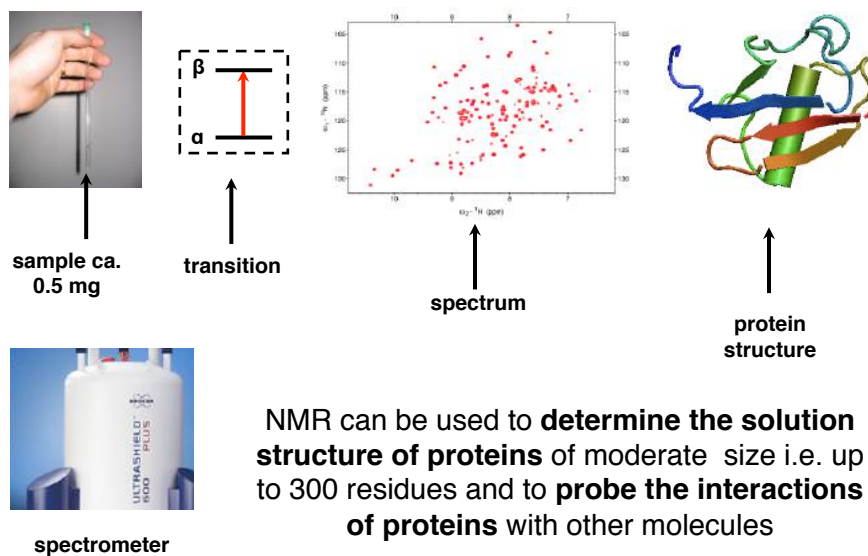
Optical microscope



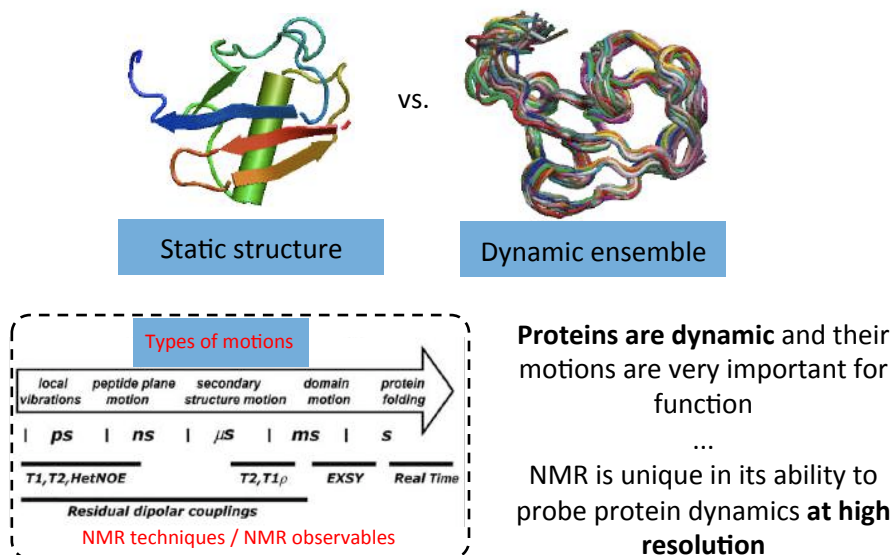
X-ray crystallography



NMR spectroscopy of proteins



NMR and protein dynamics



Protein Data Bank stores 3D structures

RCSB PDB
PROTEIN DATA BANK

135767 Biological Macromolecular Structures
Enabling Breakthroughs in Research and Education

Search by PDB ID, author, macromolecule, sequence, or ligands **Go**

Advanced Search | Browse by Annotations

PDB-101 | PDB | EMBL/GenBank | NCBI | Wellcome Protein Data Bank

Facebook | Twitter | YouTube | RSS

Welcome

- Deposit
- Search
- Visualize
- Analyze
- Download
- Learn

A Structural View of Biology

This resource is powered by the Protein Data Bank archive—information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

New Video: What is a Protein?

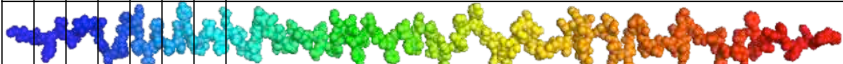
December Molecule of the Month

Biodegradable Plastic

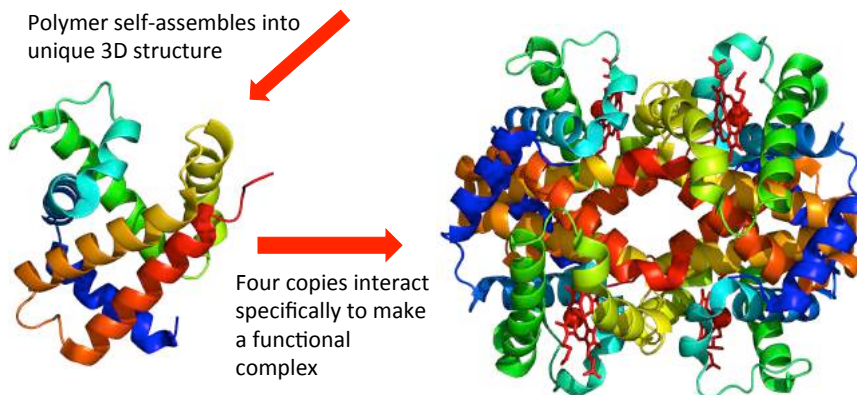
<http://www.rcsb.org/pdb/home>

Haemoglobin: carries oxygen from the respiratory organs to the rest of the body.

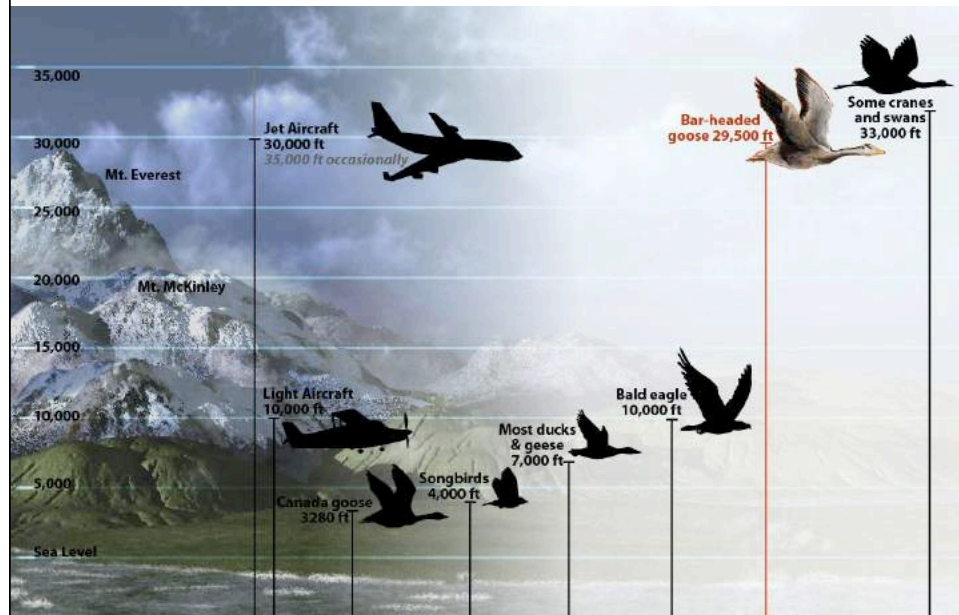
1	2	3	4	5	6	7	8	
M	V	L	S	P	A	D	K	



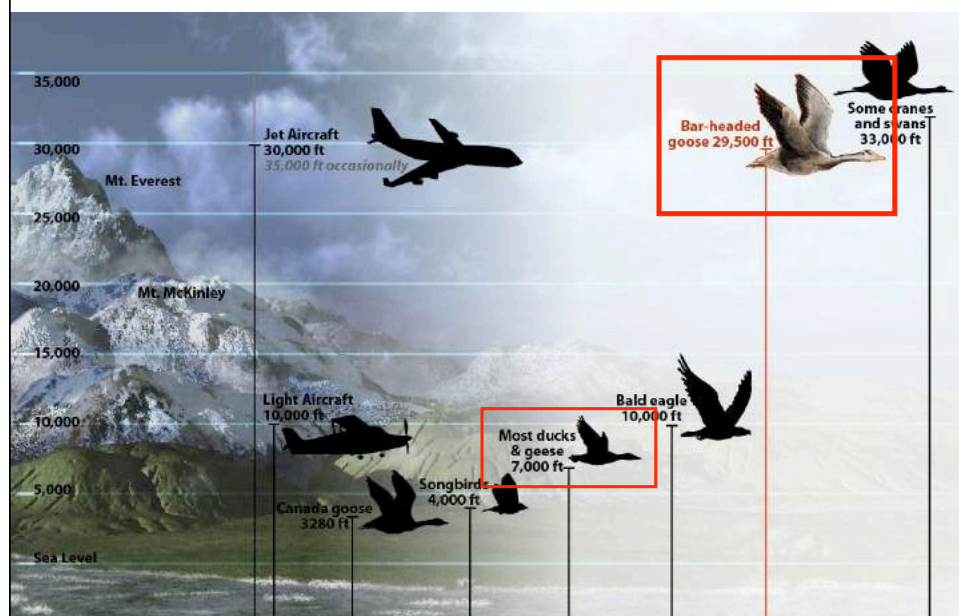
Polymer self-assembles into
unique 3D structure



Haemoglobin: carries oxygen from the respiratory organs to the rest of the body.



Haemoglobin: carries oxygen from the respiratory organs to the rest of the body.



Bar-headed
goose,
29,500 feet



Graylag
goose,
7,000 feet



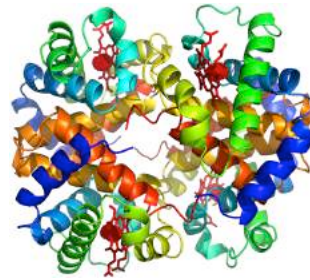
Bar-headed
goose,
29,500 feet



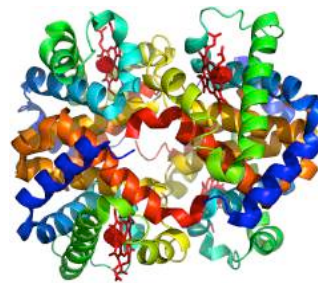
Graylag
goose,
7,000 feet



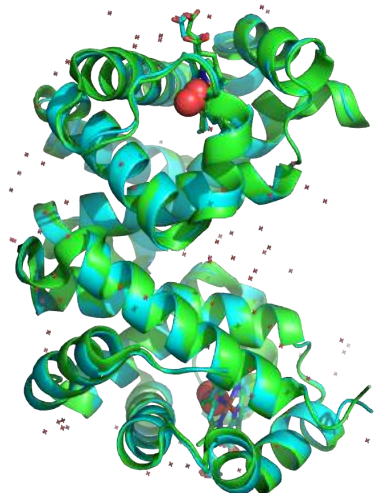
Bar-headed
goose,
29,500 feet



Graylag
goose,
7,000 feet



Bar-headed
goose,
29,500 feet



Graylag
goose,
7,000 feet



Bar-headed goose, 29,500 feet



Just 4 amino acid changes, of which one could be functional.

Graylag goose, 7,000 feet



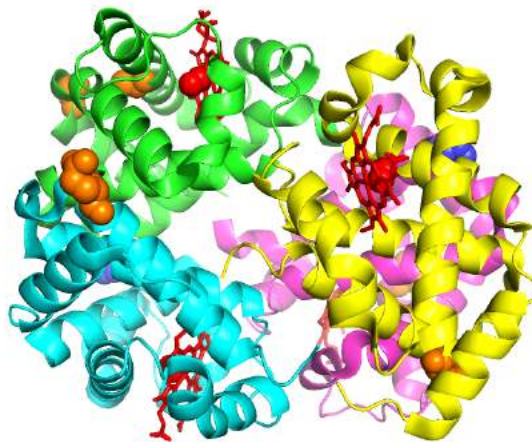
```
>sp|P01990|HBA_ANSIN Hemoglobin subunit alpha-A
OS=Anser indicus GN=HBAA PE=1 SV=2
MVLAAADKTNVKGVFSKIGHAEYGAETLERMFTAYPQTKTYFPHFDL
QHGSQAQIKAHGKVAAALVEAVNHIDDITAGALSKLDLHAQKLRVDPVN
FKFLGHCFLVVVAIHHPSALTAEVHASLDKFLCAVGTVLTAKYR

>sp|P02118|HBB_ANSIN Hemoglobin subunit beta
OS=Anser indicus GN=HBB PE=1 SV=1
VHMSAEEKQLITGLWGKVNVADCGAEALARLLIVYPWTQRFFSSFGNLS
SPTAILGNPMVRAHGKVLTSFGDAVKNLDNIKNTFAQLSELHCDKLHV
DPENFRLLGDILIIIVLAAHFAKEFTPECQAAMQKLVRVVAHALARKYH

>sp|P01989|HBA_ANSAN Hemoglobin subunit alpha-A
OS=Anser anser anser GN=HBAA PE=1 SV=2
MVLAAADKTNVKGVFSKIGHAEYGAETLERMFTAYPQTKTYFPHFDL
QHGSQAQIKAHGKVAAALVEAVNHIDDITAGALSKLDLHAQKLRVDPVN
FKFLGHCFLVVVAIHHPSALTAEVHASLDKFLCAVGTVLTAKYR

>sp|P02117|HBB_ANSAN Hemoglobin subunit beta
OS=Anser anser anser GN=HBB PE=1 SV=1
VHMSAEEKQLITGLWGKVNVADCGAEALARLLIVYPWTQRFFSSFGNLS
SPTAILGNPMVRAHGKVLTSFGDAVKNLDNIKNTFAQLSELHCDKLHV
DPENFRLLGDILIIIVLAAHFAKEFTPECQAAMQKLVRVVAHALARKYH
```

How to evaluate the functional effect of mutations? To start, where are the four mutations in the structure?



Hypothesis that the mutation in blue leads to higher oxygen affinity...

Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 6519–6522, August 1991
Evolution

Adaptation of bird hemoglobins to high altitudes: Demonstration of molecular mechanism by protein engineering

(physiology/ $\alpha_1\beta_1$ subunit contact/x-ray structure)

TIMM-H. JESSEN^{*†‡}, ROY E. WEBER[§], GIULIO FERMI[‡], JEREMY TAME^{‡¶}, AND GERHARD BRAUNITZER^{*||}

^{*}Max-Planck-Institut für Biochemie, Abteilung Proteinchemie, W-8033 Martinsried, Federal Republic of Germany; [†]Zoophysiology Laboratory, Aarhus University, DK-8000 Aarhus C, Denmark; and [‡]Medical Research Council, Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England

Communicated by M. F. Perutz, April 19, 1991 (received for review February 4, 1991)

Table 1. Oxygen affinities of human and goose Hbs

Hb	P ₅₀ , mmHg	log P ₅₀	Δ log P ₅₀
Human HbA	5.8	0.76	
Human HbPa119A	3.3	0.53	0.23
Human HbM β 55S	3.4	0.54	0.22
Greylag goose	2.8	0.45	
Bar-headed goose	2.0	0.30	0.15

mutations give rise to very similar increases in oxygen affinity. Table 1 compares our data with the known oxygen affinity values of geese Hbs. The increase in oxygen affinity (Δ log P₅₀) produced by either of the two complementary substitutions in human Hb is actually greater than the difference in oxygen affinity between the Hbs of the bar-headed and the greylag geese. This confirms that the gap left by these substitutions at the $\alpha_1\beta_1$ contact is a sufficient cause for the increase in oxygen affinity observed in nature.

What is the effect of changing the sequence?

	20	30	40	50	60	70	80	90	100																																																																						
Raccoon	V	E	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	K	N	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Ring-tailed coati	V	E	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	K	N	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Red fox	V	E	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	K	N	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Weddell seal	V	E	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	K	N	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Harbor seal	V	E	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	K	N	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Eurasian badger	V	E	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	K	N	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Red panda	V	E	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	K	N	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Ferret	V	E	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	K	N	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Fur seal	V	E	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	K	N	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
River otter	V	E	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	K	N	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Beach marten	V	E	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	K	N	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
European mink	V	E	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	K	N	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Ratel	V	E	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	K	N	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Colobus monkey	V	D	A	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	A	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Lowland gorilla	V	D	A	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	A	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Human	V	D	A	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	A	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Spider monkey	V	D	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	A	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Brown spider monkey	V	D	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	A	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Tamarin	V	D	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	A	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Moustached tamarin	V	D	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	A	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Capuchin	V	D	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	T	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
White Capuchin	V	D	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	T	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
White sapsajou	V	D	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	T	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
squirrel monkey	V	D	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	A	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
red colobus	V	D	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	A	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Collared titi	V	D	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	A	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Common marmoset	V	D	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	T	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Howler monkey	V	D	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	A	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Black spider monkey	V	D	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	A	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Goeldi's marmoset	V	D	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T	C	R	F	F	E	S	F	G	D	L	S	S	P	A	I	M	G	N	E	K	V	K	A	H	G	K	K	V	L	N	S	F	S	E	G	L	A	H	L	D	N	L	K	G	T	F	A	K	L	S	E	L	H	C	D	K	L	H	V	D	P	E
Golden lion tamarin	V	D	E	V	G	G	E	A	L	G	R	L	L	V	V	P	T																																																														

What information is stored in sequences?

Sequences record the outcome of **millions of evolutionary experiments** which are **constrained** by phenotype.

Ligand binding
Structure
Oligerimization
Enzyme Kinetics
Interaction specificity
Indirect constraints from RNA function

What information is stored in sequences?

Sequences record the outcome of **millions of evolutionary experiments** which are **constrained** by phenotype.

Ligand binding
Structure
Oligerimization
Enzyme Kinetics
Interaction specificity
Indirect constraints from RNA function



**Are these
distinct residue
sets?**

**How do these
constraints vary
between
proteins?**

Can we **invert sequence data** to reveal these constraints and so understand which sequence residues control phenotype?

Allosteric regulation of crocodilian haemoglobin

M. F. Perutz*, C. Bauer†, G. Gros‡, F. Leclercq§,
C. Vandecasserie§, A. G. Schnek§, G. Braunitzer||,
A. E. Friday¶ & K. A. Joysey¶

Nature Vol. 291 25 June 1981



Allosteric regulation of crocodilian haemoglobin

M. F. Perutz*, C. Bauer†, G. Gros‡, F. Leclercq§,
C. Vandecasserie§, A. G. Schnek§, G. Braunitzer||,
A. E. Friday¶ & K. A. Joysey¶

Nature Vol. 291 25 June 1981



Allosteric regulation of crocodilian haemoglobin

M. F. Perutz*, C. Bauer†, G. Gros‡, F. Leclercq§,
C. Vandecasserie§, A. G. Schnek§, G. Braunitzer||,
A. E. Friday¶ & K. A. Joysey¶

Nature Vol. 291 25 June 1981



Position in			
Structure	Sequence	Human	
NA1	β 1	Val	
NA2	β 2	His	
EF6	β 82	Lys	
H21	β 143	His	
HCl	β 144	Lys	
H14	α 131	Ser	

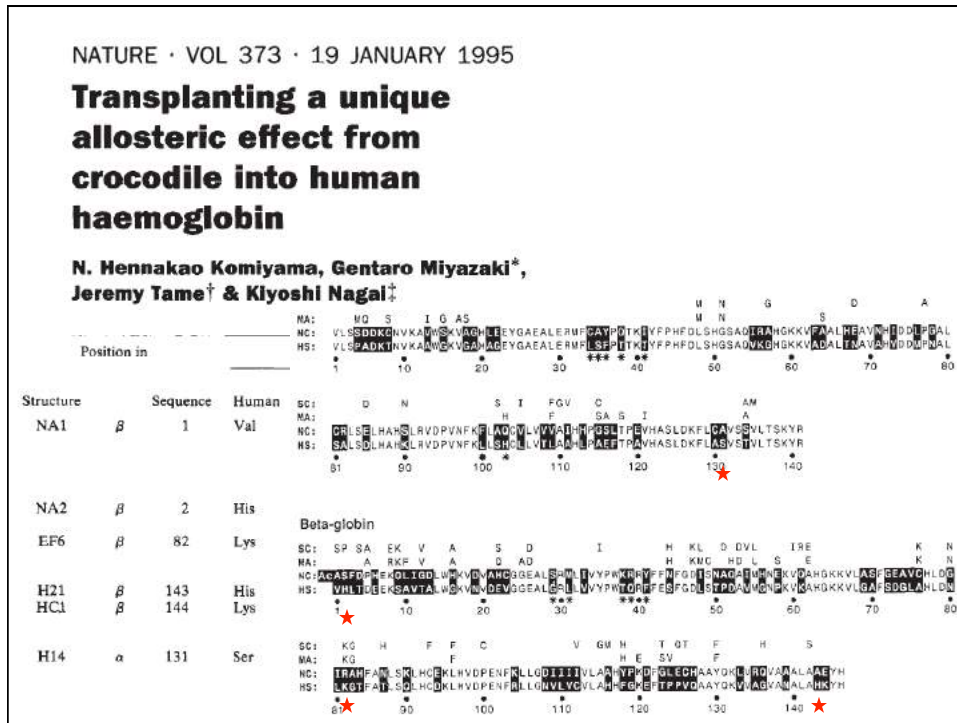
We have examined all possible sites in the model of human deoxyhaemoglobin where the amino acid replacements indicated by the three crocodilian sequences could possibly cause oxygen-linked binding of bicarbonate ions, and have found only one pair of sites with the right stereochemistry. These lie in the cavity between the two β -chains where organic phosphates or carbamino CO_2 are bound in other species, and are formed by Lys EF6(82) and Glu HCl(144) of one β -chain together with the N-terminal residue of its partner chain (Fig. 1,

NATURE · VOL 373 · 19 JANUARY 1995

Transplanting a unique allosteric effect from crocodile into human haemoglobin

N. Hennakao Komiyama, Gentaro Miyazaki*,
Jeremy Tame† & Kiyoshi Nagai‡

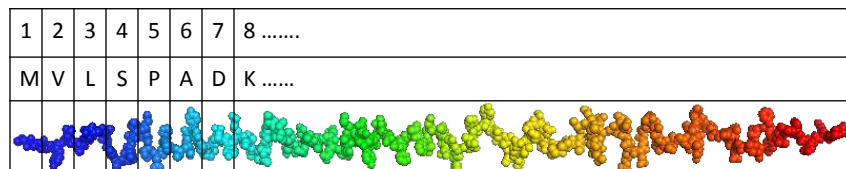
various human–crocodile chimaeric haemoglobins. Furthermore, we have been able to transplant the bicarbonate effect into human haemoglobin by replacing only a few residues, even though the amino-acid sequence identity between crocodile (*Crocodylus niloticus*) and human haemoglobins is only 68% for the α - and 51% for the β -subunit⁵. These results indicate that an entirely new function which enables species to adapt to a new environment could evolve in a protein by a relatively small number of amino-acid substitutions in key positions⁶.



How to extract useful information from protein sequences?

Each protein is constructed as a specific chain or string of amino acids. There are 20 different amino acids which are used with roughly equal frequencies across all proteins.

In a specific protein of interest, such as hemoglobin, the order of amino acids is highly important and contains all the information necessary to produce the folded, functional molecule.



We would like to find a probability model for the sequence of amino acids that corresponds to each protein of interest.

$$P(A_1, \dots, A_L) = \text{Probability that a sequence produces a folded, functional hemoglobin molecule.}$$