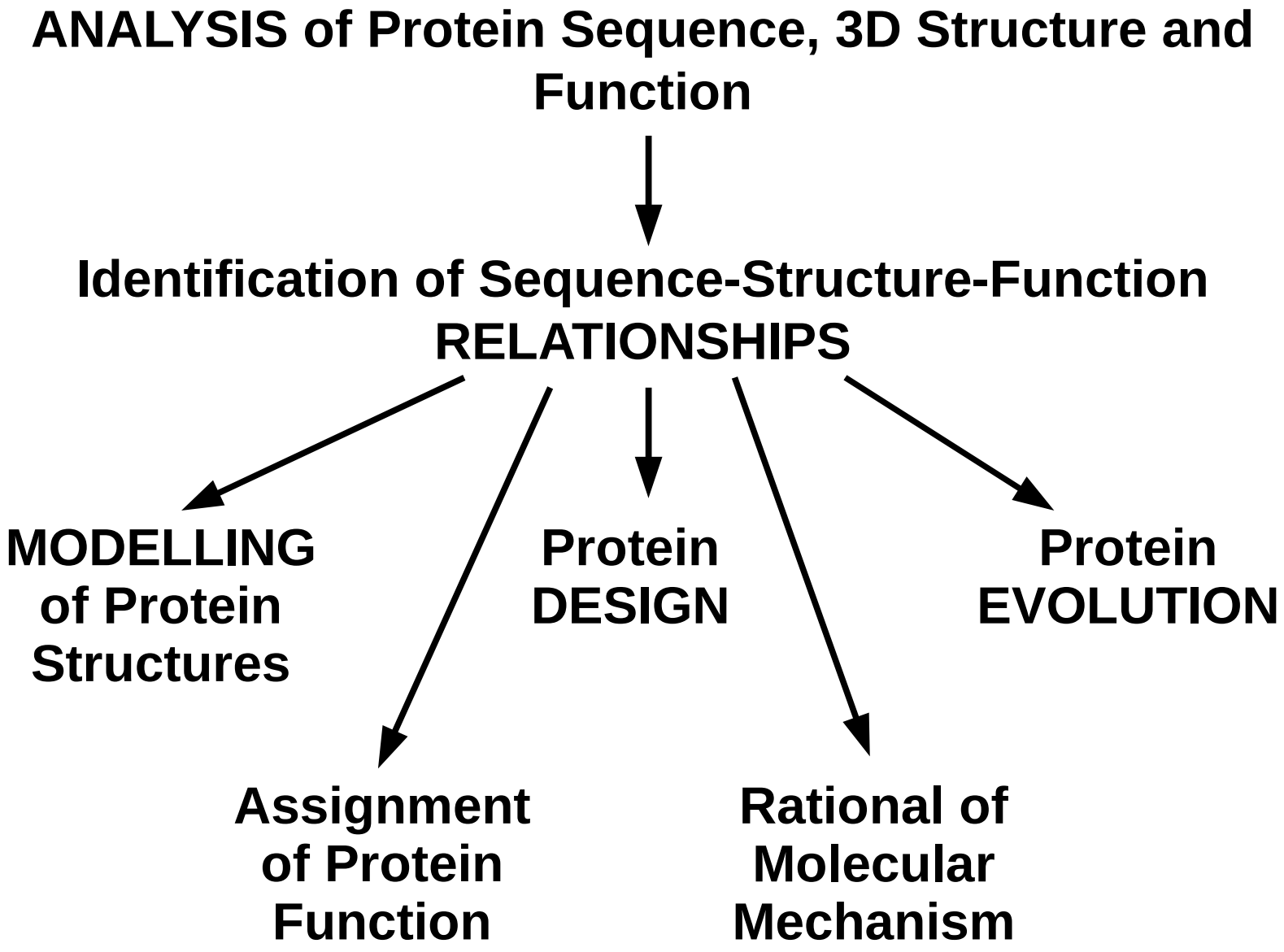

ANALYSIS of Protein Sequence, 3D Structure and Function
Identification of Sequence-Structure-Function RELATIONSHIPS
MODELLING of Protein Structures
Protein DESIGN
Protein EVOLUTION
Assignment of Protein Function
Rational of Molecular Mechanism

WHERE are the Structures?

THE PROTEIN DATA BANK

www.pdb.org

WHERE are the Structures?

THE PROTEIN DATA BANK

www.pdb.org

WHAT IS A PROTEIN?

http://pdb101.rcsb.org/learn/resource/what-is-a-protein-video

WHAT IS A PROTEIN?

BIOLOGICAL MACRO MOLECULE

SEQUENCE

- building-blocks: amino acids
- 20 amino acid types

STRUCTURE

- primary structure
- secondary structure
- tertiary structure
- quaternary structure

FUNCTION

- molecular
- cellular
- pathway

PROTEIN RULES

AMINO ACID SEQUENCE

↓

3D STRUCTURE

↓

FUNCTION

WHAT IS A PROTEIN?

SEQUENCE

- 20 building-blocks: amino acids

A	Ala	Alanine	M	Met	Metionine
C	Cys	Cysteine	N	Asn	Asparagine
D	Asp	Aspartic Acid	P	Pro	Proline
E	Glu	Glutamic Acid	Q	Gln	Glutamine
F	Phe	Phenylalanine	R	Arg	Arginine
G	Gly	Glycine	S	Ser	Serine
H	His	Histidine	T	Thr	Threonine
I	Ile	Isoleucine	V	Val	Valine
K	Lys	Lysine	W	Trp	Tryptophane
L	Leu	Leucine	Y	Tyr	Tyrosine

WHAT IS A PROTEIN?

SEQUENCE

- atoms in amino acids

C (green)
H (white)
N (blue)
O (red)
S (yellow)
P (magenta)

The Periodic table

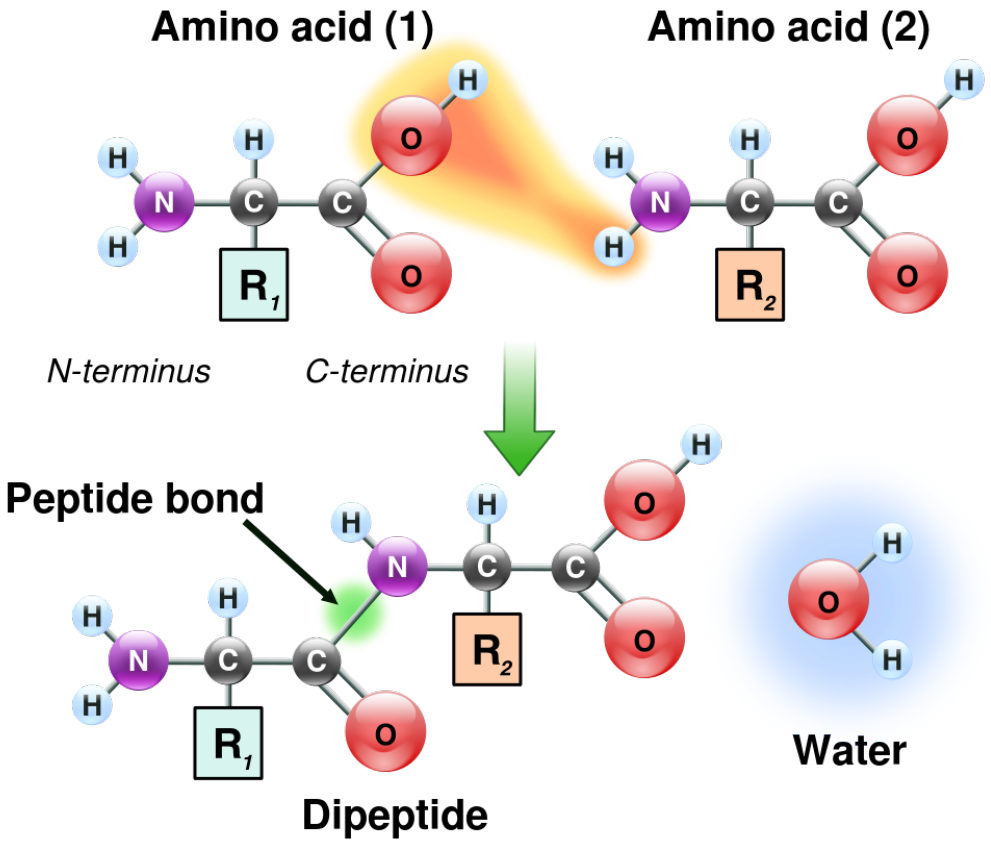
Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Period 1	1 H 1.008																	2 He 4.003
2	3 Li 6.94	4 Be 9.012											5 B 10.81	6 C 12.01	7 N 14.01	8 O 16.00	9 F 19.00	10 Ne 20.18
3	11 Na 22.99	12 Mg 24.31											13 Al 26.98	14 Si 28.09	15 P 30.97	16 S 32.06	17 Cl 35.45	18 Ar 39.95
4	19 K 39.10	20 Ca 40.08	21 Sc 44.96	22 Ti 47.88	23 V 50.94	24 Cr 52.00	25 Mn 54.94	26 Fe 55.85	27 Co 58.93	28 Ni 58.69	29 Cu 63.55	30 Zn 65.39	31 Ga 69.72	32 Ge 72.64	33 As 74.92	34 Se 78.96	35 Br 79.90	36 Kr 83.79
5	37 Rb 85.47	38 Sr 87.62	39 Y 88.92	40 Zr 91.22	41 Nb 92.91	42 Mo 95.96	43 Tc (98)	44 Ru 101.1	45 Rh 102.9	46 Pd 106.4	47 Ag 107.9	48 Cd 112.4	49 In 114.8	50 Sn 118.7	51 Sb 121.8	52 Te 127.6	53 I 126.9	54 Xe 131.3
6	55 Cs 132.9	56 Ba 137.3	*	72 Hf 178.5	73 Ta 180.9	74 W 183.9	75 Re 186.2	76 Os 190.2	77 Ir 192.2	78 Pt 195.1	79 Au 197.0	80 Hg 200.5	81 Tl 204.38	82 Pb 207.2	83 Bi 209.0	84 Po (209)	85 At (210)	86 Rn (222)
7	87 Fr (223)	88 Ra (226)	**	104 Rf (265)	105 Db (268)	106 Sg (271)	107 Bh (270)	108 Hs (277)	109 Mt (276)	110 Ds (281)	111 Rg (280)	112 Cn (285)	113 Uut (284)	114 Fl (289)	115 Uup (288)	116 Lv (293)	117 Uus (294)	118 Uuo (294)

Lanthanide Series*	57 La 138.9	58 Ce 140.1	59 Pr 140.9	60 Nd 144.2	61 Pm (145)	62 Sm 150.4	63 Eu 152.0	64 Gd 157.2	65 Tb 158.9	66 Dy 162.5	67 Ho 164.9	68 Er 167.3	69 Tm 168.9	70 Yb 173.0	71 Lu 175.0
Actinide Series**	89 Ac (227)	90 Th 232	91 Pa 231	92 U 238	93 Np (237)	94 Pu (244)	95 Am (243)	96 Cm (247)	97 Bk (247)	98 Cf (251)	99 Es (252)	100 Fm (257)	101 Md (258)	102 No (259)	103 Lr (262)

WHAT IS A PROTEIN?

SEQUENCE

- main-chain, side-chain

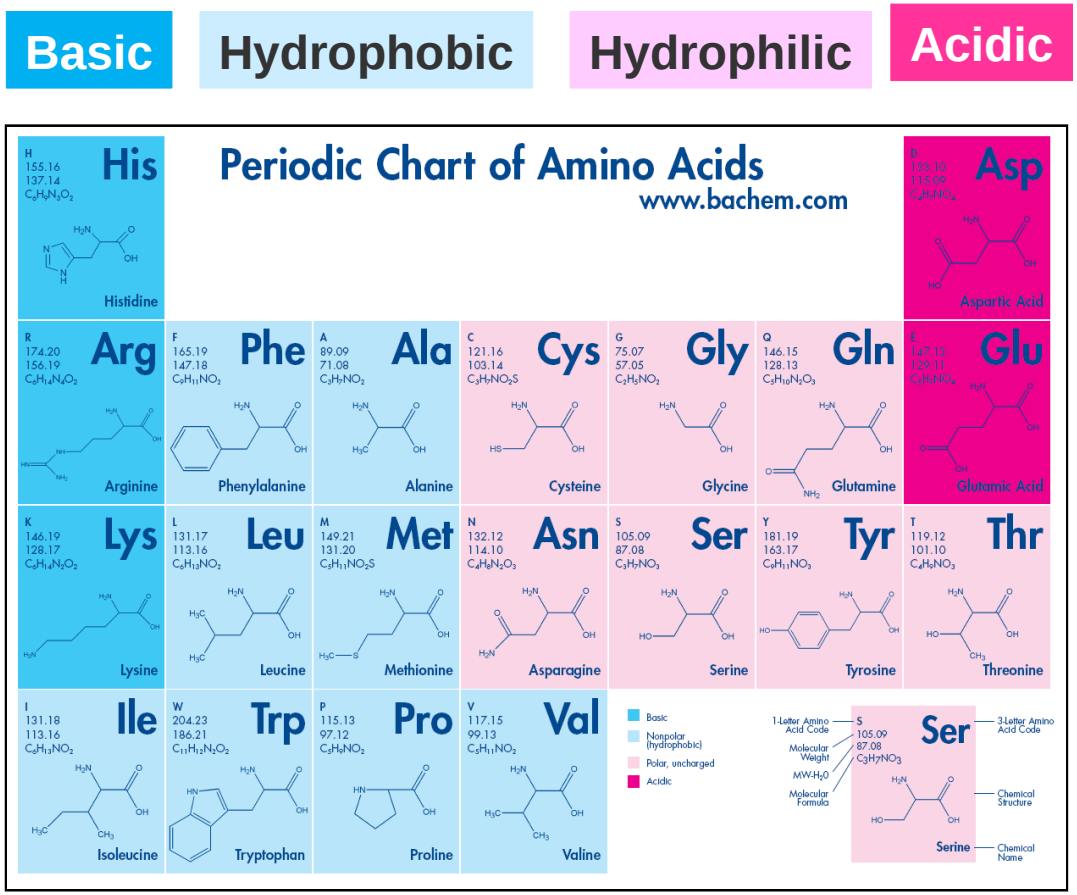


WHAT IS A PROTEIN?

SEQUENCE

- amino acid types

WHEN ARE TWO AMINO ACIDS "SIMILAR"?

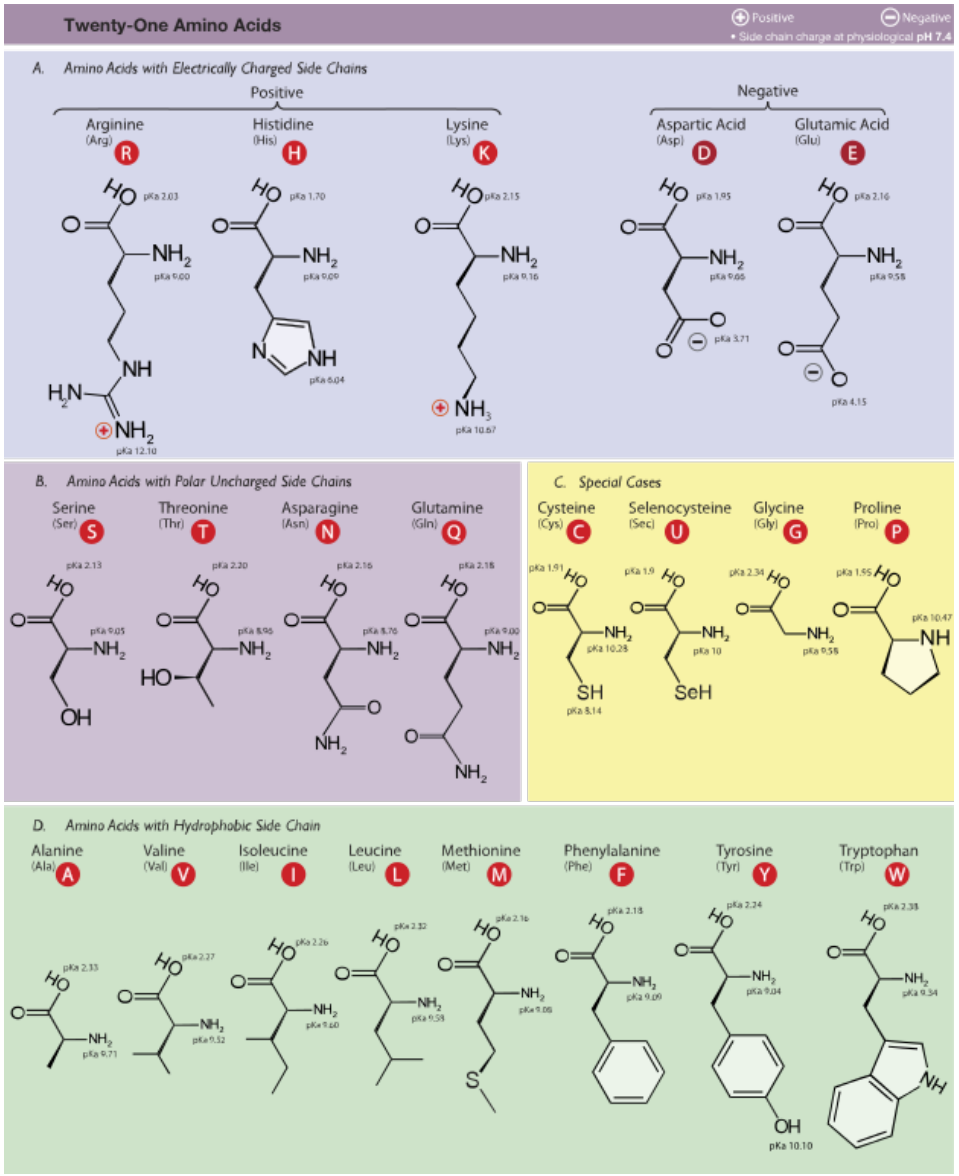


WHAT IS A PROTEIN?

SEQUENCE

- amino acid types

WHEN ARE TWO AMINO ACIDS "SIMILAR"?

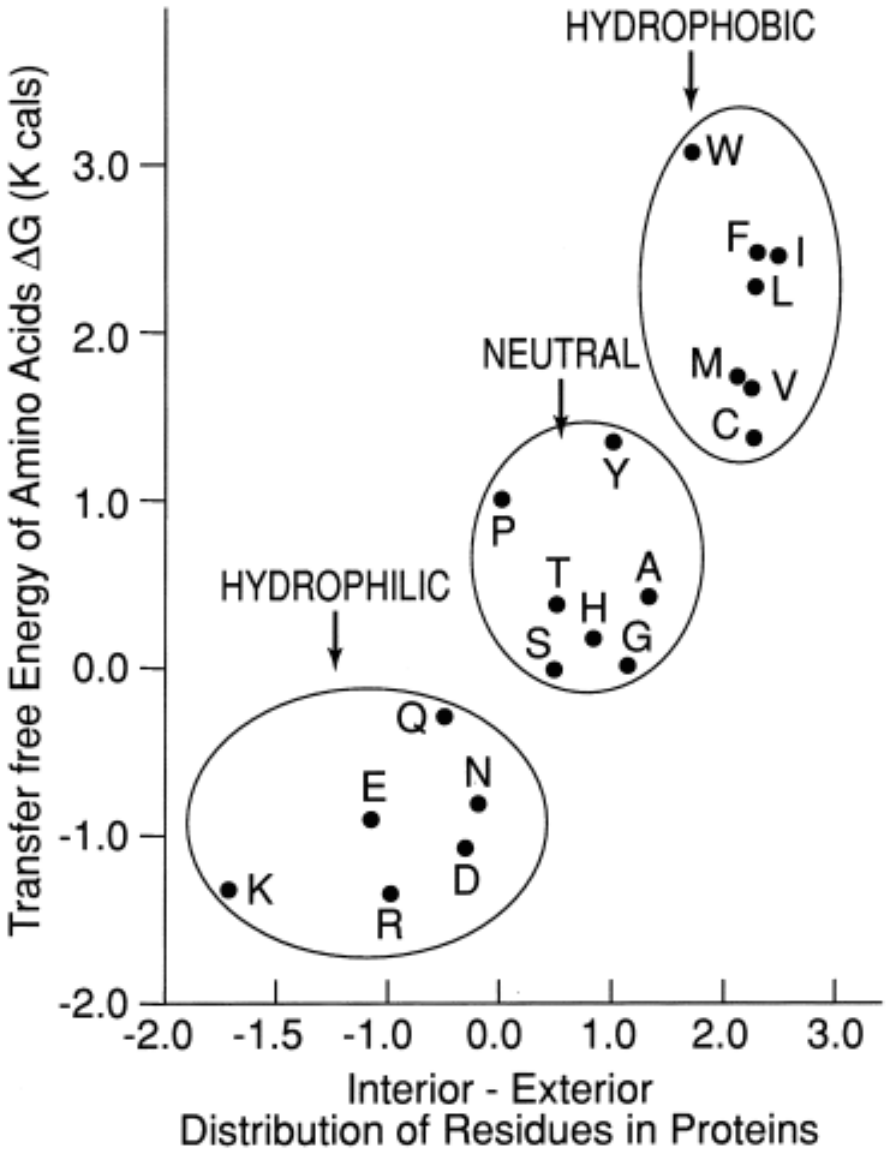


WHAT IS A PROTEIN?

SEQUENCE

- amino acid types

WHEN ARE TWO AMINO ACIDS "SIMILAR"?



WHAT IS A PROTEIN?

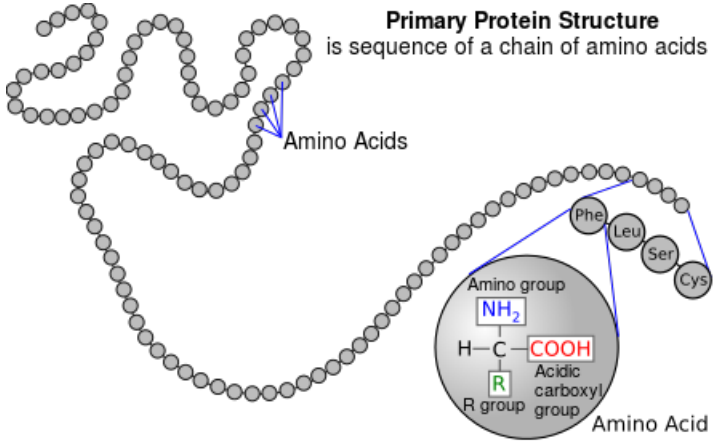
STRUCTURE

- primary structure

SEQUENCE OF AMINO ACID RESIDUES

```
>gi|4503079|ref|NP_000750.1|
MAGPATQSPMKLMALQLLLWHSALWTVQEA
TPLGPASSLPQSFLLKCLEQVRKIQGDGAA
LQEKLVSECATYKLCHPEELVLLGHSLGIP
WAPLSSCPSQALQLAGCLSQLHSGLFLYQG
LLQALEGISPELGPTLDTLQLDVADFATTI
WQQMEELGMAPALQPTQGAMPAFASAFQRR
AGGVLVASHLQSFLEVSYRVLRHLAQP
```

TPEEKSAVTALWGKV



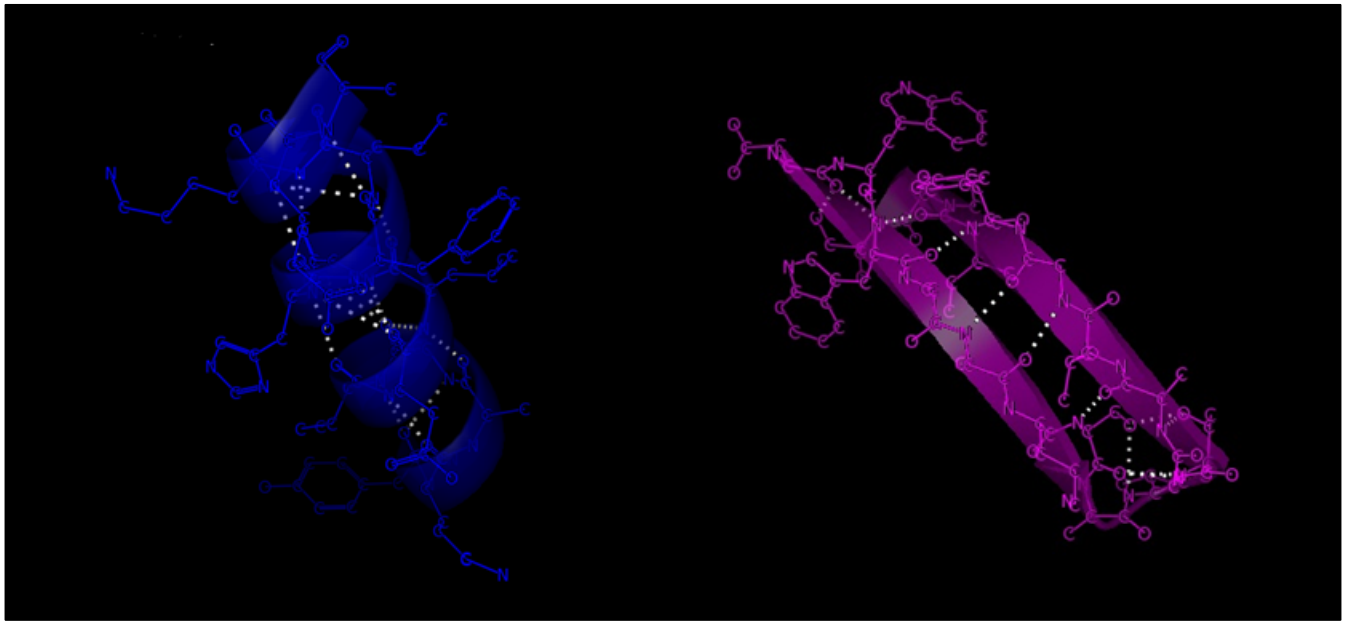
WHAT IS A PROTEIN?

STRUCTURE

- secondary structure

α-HELICES

β–SHEET



WHAT IS A PROTEIN?

STRUCTURE

- tertiary structure

SPATIAL ARRANGEMENT (FOLD)
OF THE AMINO ACID SEQUENCE
IN THE 3D SPACE

WHAT IS A PROTEIN?

STRUCTURE

- quaternary structure

PROTEIN CHAINS INTERACTING
TO FORM STABLE COMPLEXES

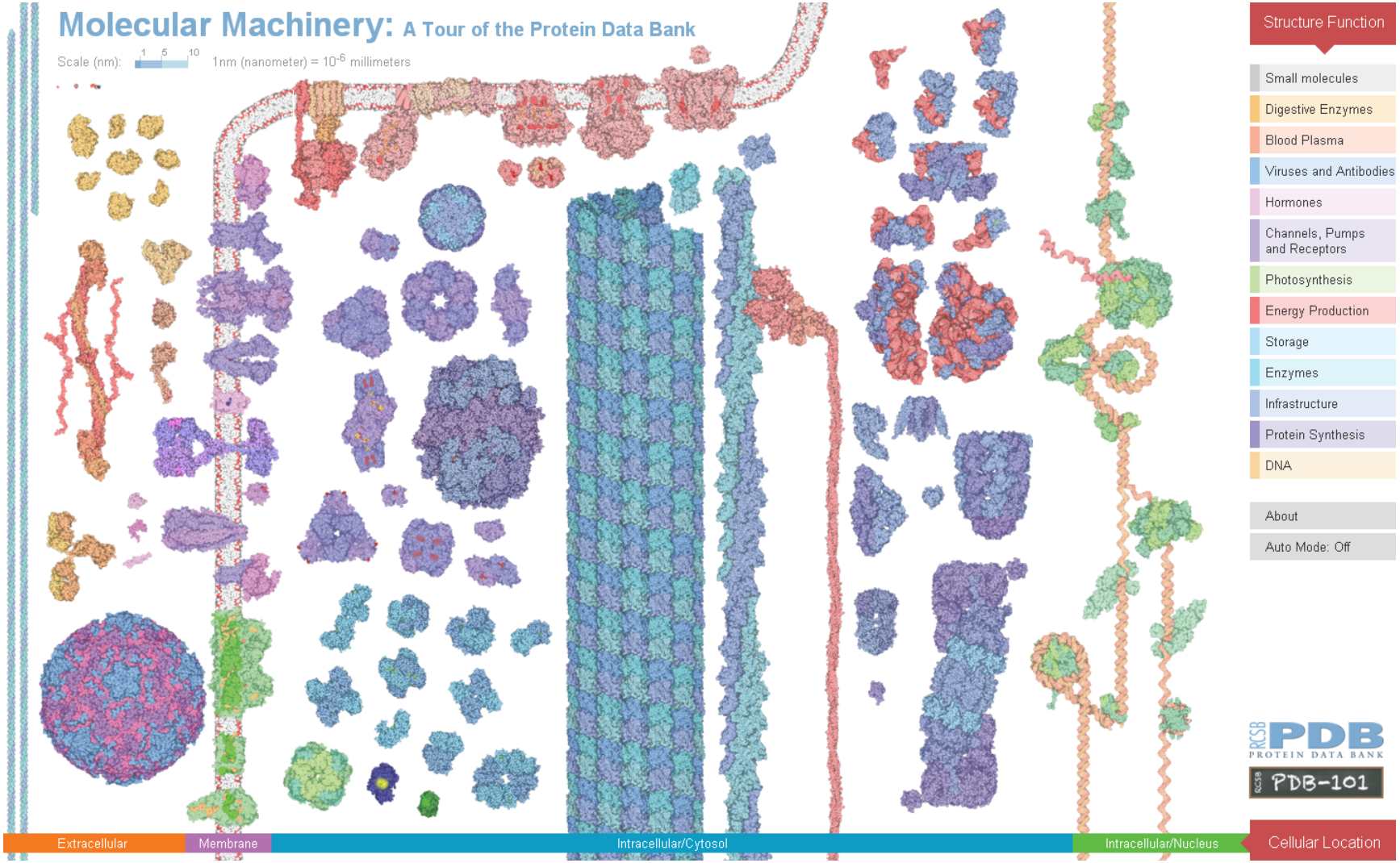
WHAT IS A PROTEIN?

FUNCTION

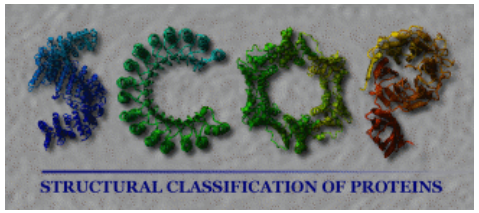
	PROTEIN	3D STRUCTURE IDENTIFIER (PDB ID)
STRUCTURAL	COLLAGEN	1CAG
DEFENSE	ANTIBODY	1IGT, 1VFB
RESPIRATION	HAEMOGLOBIN	2HHB/1HHO
STORAGE	FERRITIN	1FHA
COMMUNICATION	INSULIN	4INS, 3W11
CATALYSIS	ALPHA-AMYLASE	1PPI
TRANSPORT	CALCIUM PUMP	1IWO, 1SU4

PROTEIN STRUCTURE-FUNCTION

http://mm.rcsb.org/



PROTEIN STRUCTURE CLASSIFICATION DBs

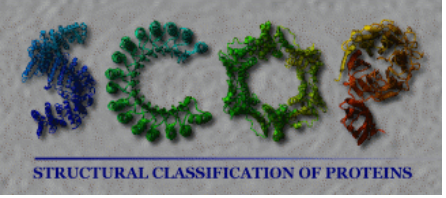


http://scop.mrc-lmb.cam.ac.uk/scop/

http://www.cathdb.info/

DALI

http://ekhidna.biocenter.helsinki.fi/dali/start



Structural Classification Of Proteins

(scop.mrc-lmb.cam.ac.uk/scop/)

CLASS:	**Secondary Structure content (α; β; α/β; α+β; …)**
FOLD:	**3D Structure Similarity**
SUPERFAMILY:	**3D Structure Similarity & Distant evolutionary relationship**
FAMILY:	**Amino Acid Sequence Similarity & Close evolutionary relationship**

DOMAINS:	**Structural, functional, folding and evolutionary protein units**

PROTEIN FUNCTION CLASSIFICATION DBs

WHAT IS PROTEIN FUNCTION?

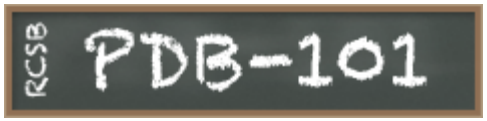
Gene Ontology Consortium

http://geneontology.org/

Three aspects:

1. **molecular function**
   molecular activities of gene products
1. **cellular component**
   where gene products are active
1. **biological process**
   pathways and larger processes made up of the activities of multiple gene products

PROTEIN STRUCTURE-FUNCTION

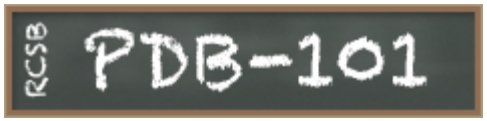


Educational portal of

Browse by category

1. **Health and Disease**
2. **Molecules of Life**
3. **Biotech and Nanotech**
4. **Structures and Structure Determination**

PROTEIN STRUCTURE-FUNCTION



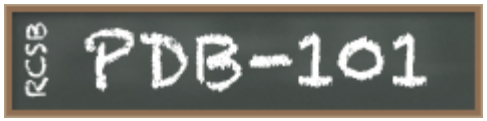
Educational portal of

Browse by category

1. Health and Disease

a) You and Your Health
b) Immune System
c) HIV and AIDS
d) Diabetes
e) Cancer
f) Viruses
g) Toxins and Poisons
h) Drug Action
i) Drug Resistance

PROTEIN STRUCTURE-FUNCTION



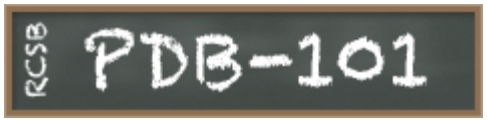
Educational portal of RCSB PDB PROTEIN DATA BANK

Browse by category

2. Molecules of life

a) DNA, RNA, and Protein Synthesis
b) Enzymes
c) Molecular Infrastructure
d) Transport
e) Biological Energy
f) Molecules and the Environment
g) Photosynthesis
h) Molecular Motors
i) Cellular Signaling

PROTEIN STRUCTURE-FUNCTION



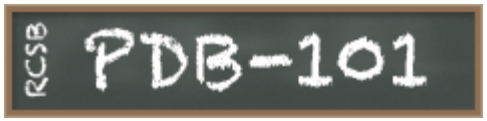
Educational portal of

Browse by category

3. Biotech and Nanotech

a) Recombinant DNA
b) Biotechnology
c) using biology in industry
d) Nanotechnology
e) Renewable Energy

PROTEIN STRUCTURE-FUNCTION



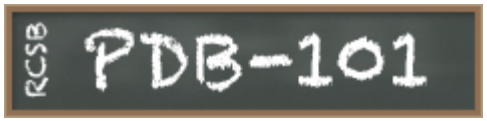
Educational portal of RCSB PDB PROTEIN DATA BANK

Browse by category

4. Structures and Structure Determination

a) Biomolecules
b) Biomolecular Structural Biology
c) Integrative/Hybrid Methods
d) PDB Data
e) Visualizing Molecules

PROTEIN STRUCTURE-FUNCTION



Educational portal of RCSB PDB PROTEIN DATA BANK

Browse by category

1. Health and Disease

a) You and Your Health

- **How do drugs work?**

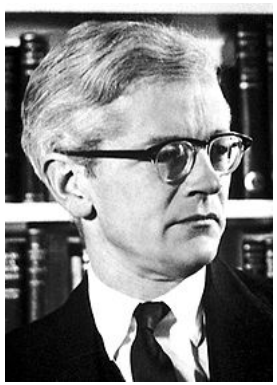
  http://cdn.rcsb.org/pdb101/learn/resources/how-do-drugs-work-flyer.pdf

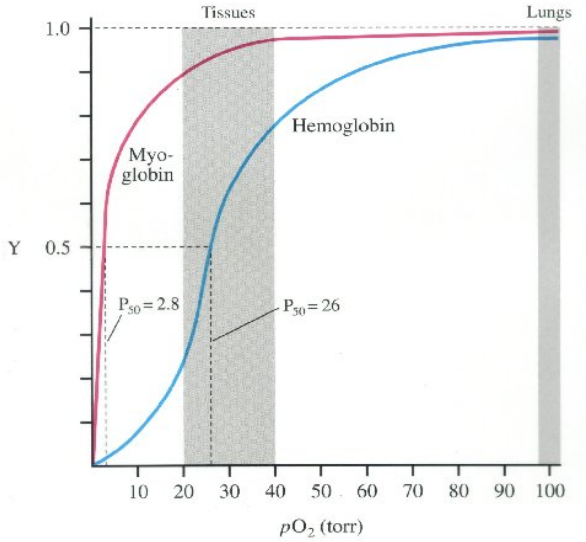
- Alcohol Dehydrogenase
- Anabolic Steroids
- Hemoglobin
- Serotonin receptors
- Serum Albumin

PROTEIN STRUCTURE-FUNCTION

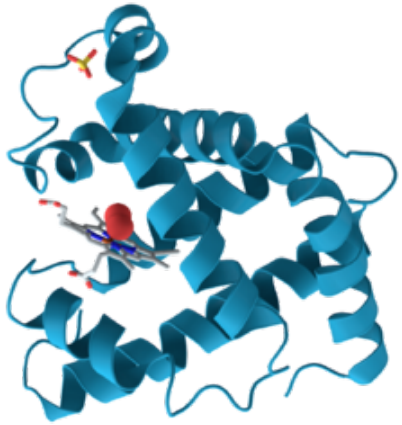
1962 Shared Nobel Prize in Chemistry

John Kendrew

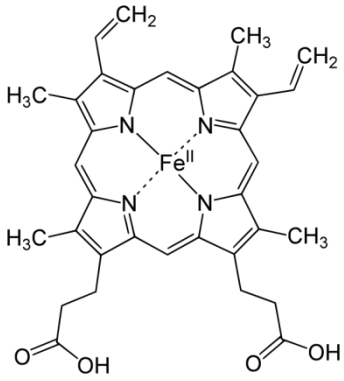


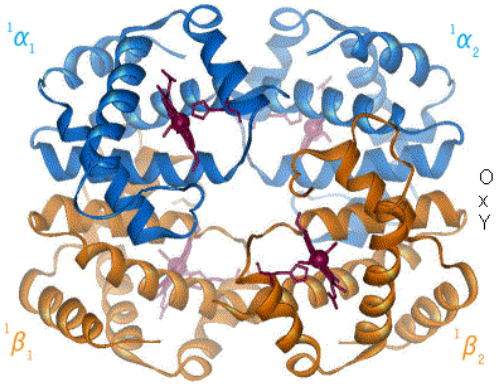


Max Perutz



myoglobin

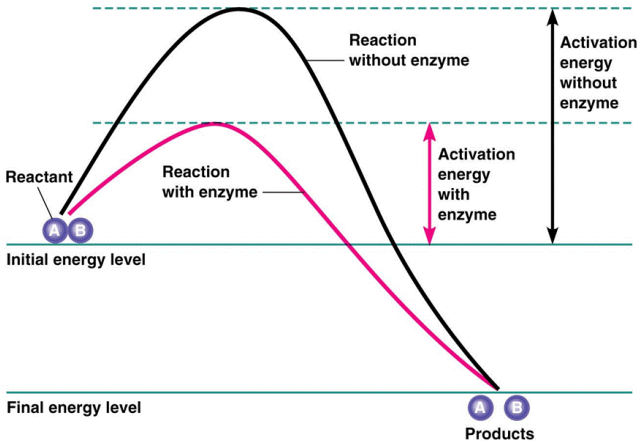




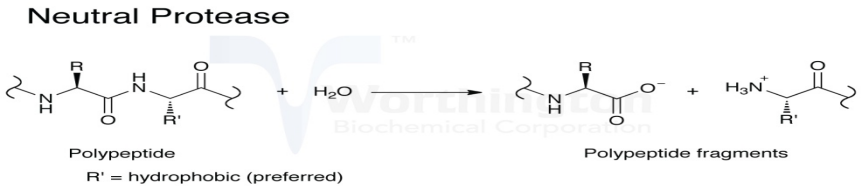
haemoglobin

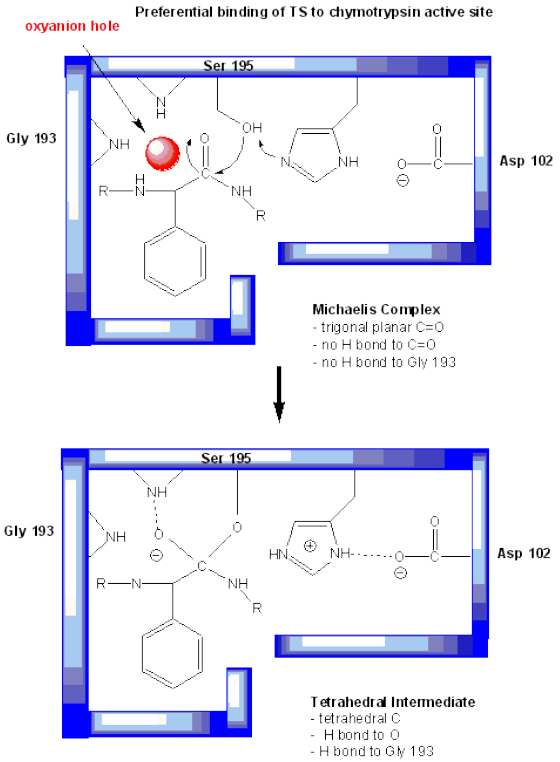
PROTEIN STRUCTURE-FUNCTION

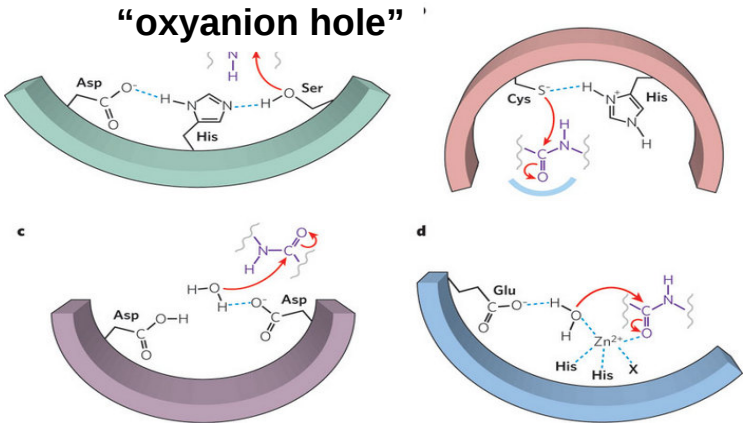
ENZYMES



PROTEOLYTIC ENZYMES



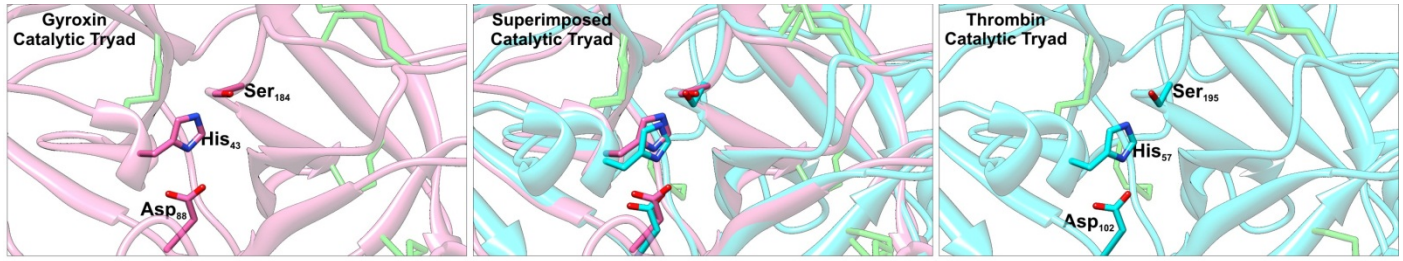


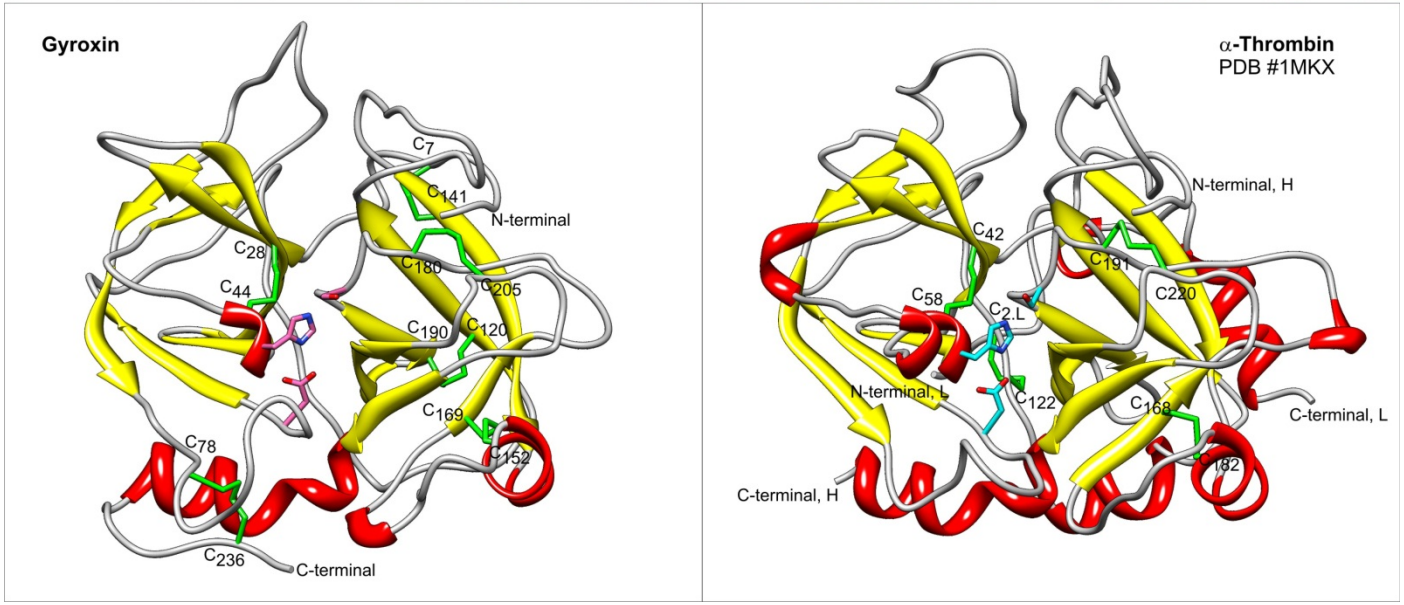


PROTEIN STRUCTURE-FUNCTION

SERINE PROTEASES

Conserved Catalytic Triad (Ser-His-Asp)





Different Structures

PROTEIN RULES

FUNCTIONAL RESIDUES
are
MORE CONSERVED
than
the rest of the structure

PROTEIN STRUCTURE ANALYSIS

ANTIBODY STRUCTURES

3D structures of intact Antibodies are available from the RCSB PDB PROTEIN DATA BANK

- B12 (PDB IDs: **1HZH**, 1IQD)
- Mab231 (PDB ID: 1IGT)
- Mab61.1.3 (PDB ID: 1IGY)
- MCG (PDB ID: 1MCO)

ANTIBODY STRUCTURES

Antibodies (Abs) or **Immunoglobulins (Igs)**

Source: plasma cells

Localization:

blood, body fluids,

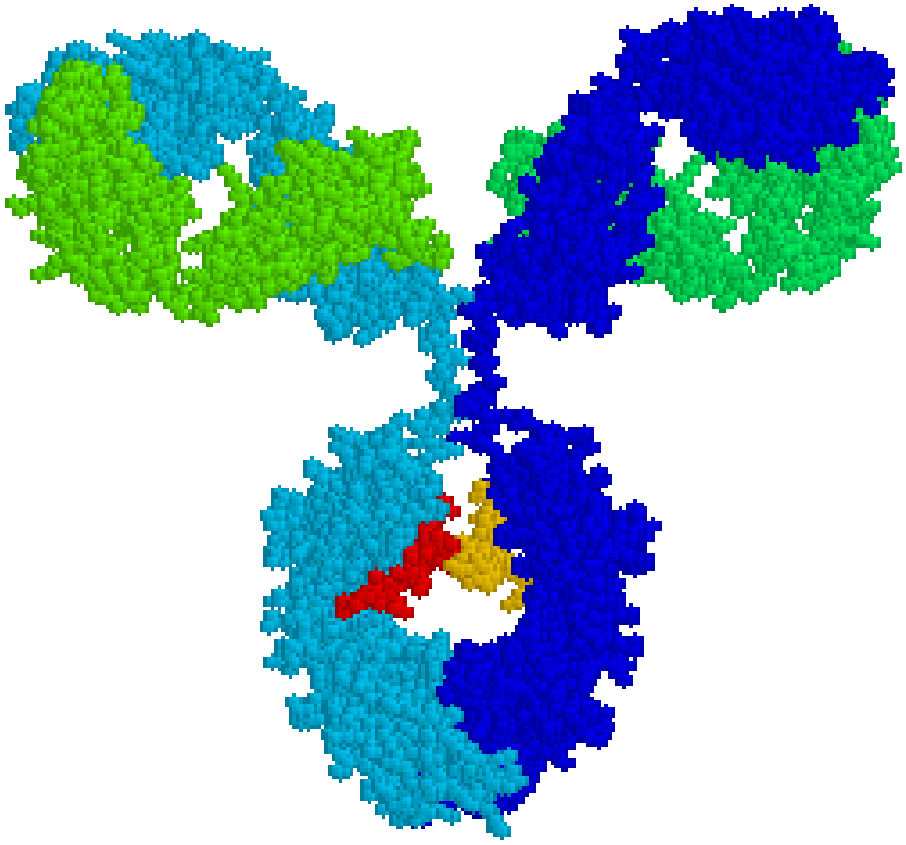
secretions, B-cell membrane

Function: identify and

neutralize foreign antigens (Ags)

Structure:

- “Y”-shaped
- large: 12 domains in IgG
- oligomeric: 4 chains (blue, green)
- glycoprotein: sugars (red, orange)



ANTIBODY STRUCTURE ANALYSIS

Download from the RCSB PDB PROTEIN DATA BANK

- **1HZH**
- **Display files:**
  - PDB file -> save 1HZH.pdb
  - Fasta sequence -> save 1HZH_fasta.txt

ANTIBODY STRUCTURE ANALYSIS

1HZH_fasta.txt: sequence file (fasta format):

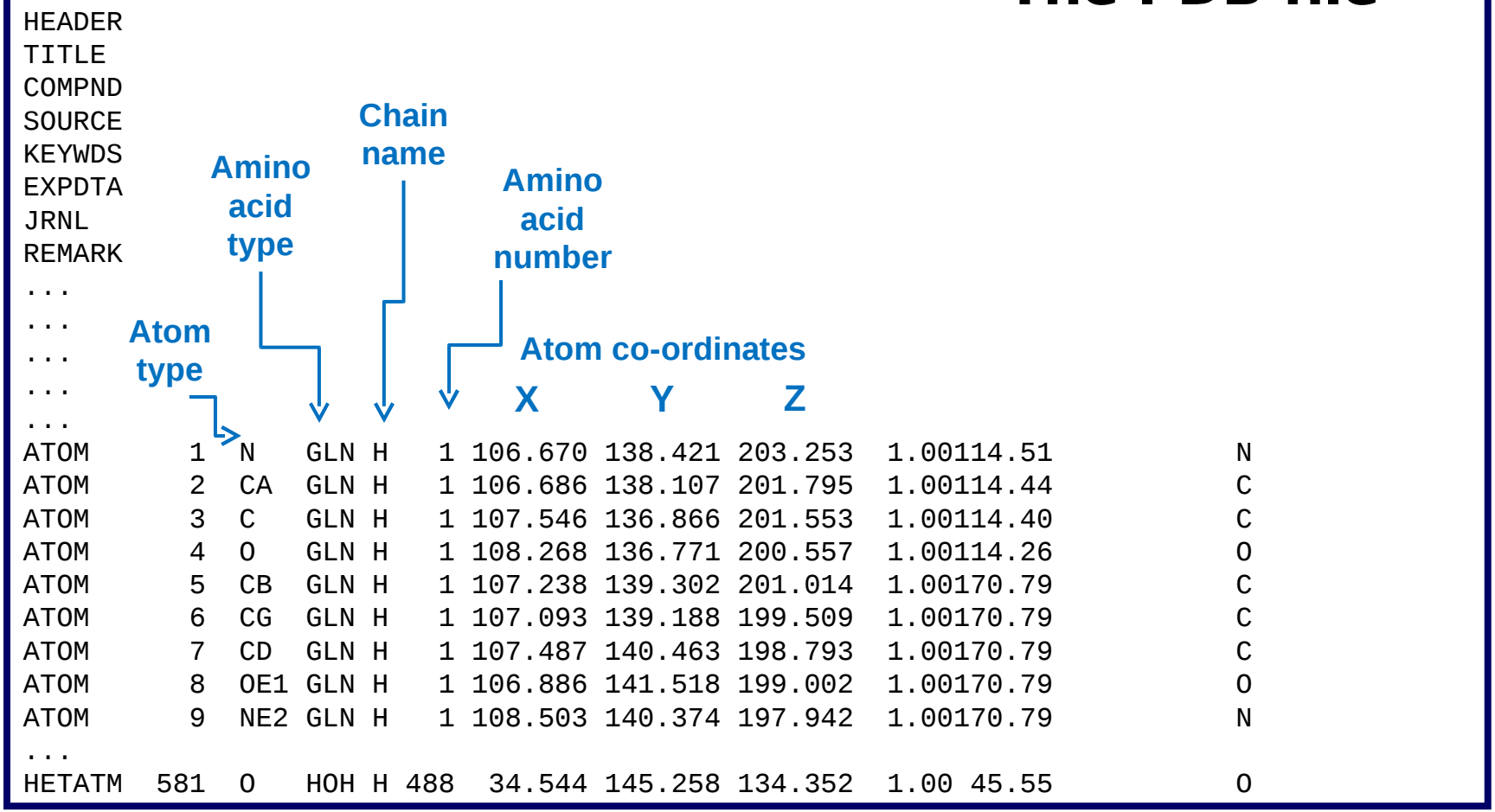
- 4 chains: 2 ‘light’ (M,L: 215 aa) and 2 ‘heavy’ (K,H: 457 aa)

```
>1HZH:M|PDBID|CHAIN|SEQUENCE
EIVLTQSPGTLSLSPGERATFSCRSSHSIRSRRVAWYQHKPGQAPRLVIHGVSNRASGISDRFSGSGSGTDFTLTITRVE
PEDFALYYCQVYGASSYTFGQGTKLERKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNS
QESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLRSPVTKSFNRGEC

>1HZH:L|PDBID|CHAIN|SEQUENCE
EIVLTQSPGTLSLSPGERATFSCRSSHSIRSRRVAWYQHKPGQAPRLVIHGVSNRASGISDRFSGSGSGTDFTLTITRVE
PEDFALYYCQVYGASSYTFGQGTKLERKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNS
QESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLRSPVTKSFNRGEC

>1HZH:K|PDBID|CHAIN|SEQUENCE
QVQLVQSGAEVKKPGASVKVSCQASGYRFSNFVIHWVRQAPGQRFEWMGWINPYNGNKEFSAKFQDRVTFTADTSANTAY
MELRSLRSADTAVYYCARVGPYSWDDSPQDNYYMDVWGKGTTVIVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYF
PEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKAEPKSCDKTHTCPPCP
APELLGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREEQYNSTYRVVSVLTVLH
QDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYTLPPSRDELTKNQVSLTCLVKGFYPSDIAVEWESNGQPENN
YKTTPPVLDSDGSFFLYSKLTVDKSRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK

>1HZH:H|PDBID|CHAIN|SEQUENCE
QVQLVQSGAEVKKPGASVKVSCQASGYRFSNFVIHWVRQAPGQRFEWMGWINPYNGNKEFSAKFQDRVTFTADTSANTAY
MELRSLRSADTAVYYCARVGPYSWDDSPQDNYYMDVWGKGTTVIVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYF
PEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKAEPKSCDKTHTCPPCP
APELLGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREEQYNSTYRVVSVLTVLH
QDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYTLPPSRDELTKNQVSLTCLVKGFYPSDIAVEWESNGQPENN
YKTTPPVLDSDGSFFLYSKLTVDKSRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK
```

ANTIBODY STRUCTURE ANALYSIS

1HZH.pdb: co-ordinate file (PDB format):

The PDB file



PROTEIN STRUCTURE ANALYSIS

Download software: Swiss PDB Viewer (spdbv)

Home page: http://spdbv.vital-it.ch/

Download: http://spdbv.vital-it.ch/disclaim.html

- **Click on: “I agree…”**
- **Download the most recent version for: Macintosh or Windows**
  - **Show in folder**
  - **Extract (compressed file)**
  - **Open folders to find spdbv icon**
  - **Double-click on spdbv icon to open**
  - **Move the program to folder of choice and create link**
- **Download User Guide**
- **FollowTutorial**

ANTIBODY STRUCTURE ANALYSIS

Swiss-Pdb Viewer

- **Main menu:**
  - File: open PDB file -> 1HZH.pdb
  - Move molecules: 4 icons on the left
  - Colour -> Chains; Secondary structure; Type; CPK; Accessibility; B-factor; Chains
- **Control Panel:**
  - Select sugars –> colour orange
  - Select residues –> colour green
- **Tools:**
  - Surface -> compute; discard

ANTIBODY STRUCTURE ANALYSIS

Swiss-Pdb Viewer

- Display: side-chains; main-chain C, N, O atoms; sugars
- Identify domain boundaries
- Colour domains
- Highlight –S-S- bonds
- Visualize only Fab and Fv
- Show as ribbon
- Calculate h-bonds
- Highlight conserved cysteines (L23, H22; L88, H92) and tryptophanes (L35, H36)
- Colour secondary structures: identify loops

ANTIBODY STRUCTURE ANALYSIS

Abs:

- multi-chain and multi-domain
- 4 chains: 2 light (L), 2 heavy (H)
- 12-14 domains:
  - 2 light chains: 1 variable (VL) and 1 constant (CL)
  - 2 heavy chains: 1 variable (VH) and 3 or 4 constant (CH1, CH2, CH3; or CH1, CH2, CH3 and CH4)

Domain:

- structure, function, folding, evolutionary protein unit

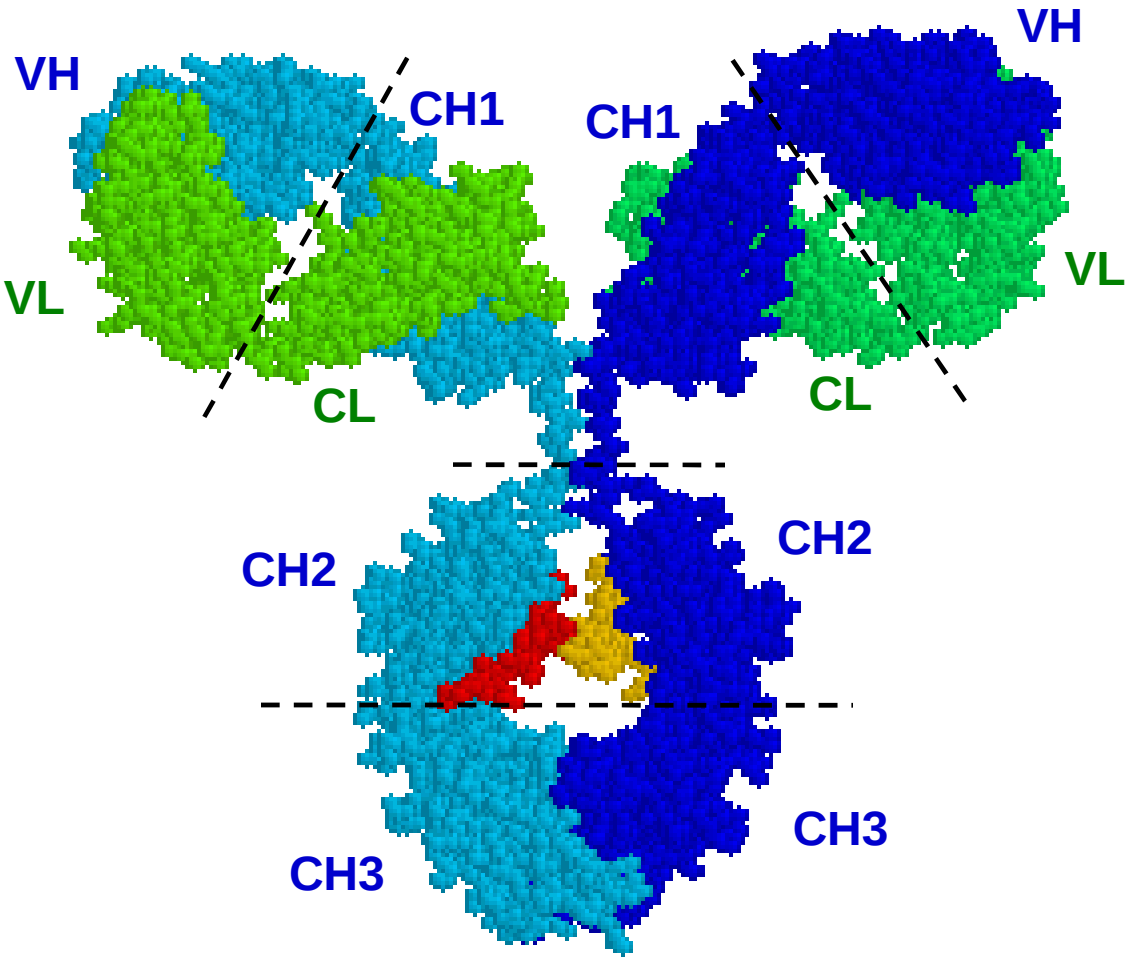
PROTEIN RULES

PROTEINS
are made of
DOMAINS

PROTEIN RULES

DOMAINS
are
STRUCTURAL
FUNCTIONAL
EVOLUTIONARY
(FOLDING)
UNITS

ANTIBODY STRUCTURE ANALYSIS



Domains:

structure

function

evolutionary

(folding)

protein units

Light (L) chain
VL, CL

Heavy (H) chain
VH, CH1, CH2, CH3

ANTIBODY STRUCTURE ANALYSIS

Swiss-Pdb Viewer: Highlight domains

- Control Panel:

- side -> undisplay side-chains
- show -> undisplay sugars

- Main menu:

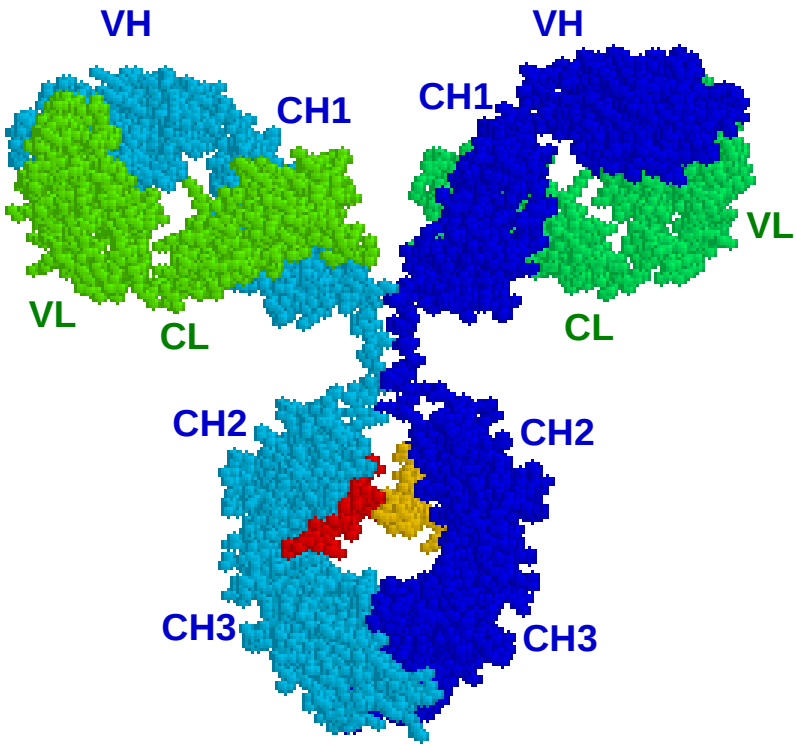
- Display: show backbone as alpha carbon trace
- Colour: chain

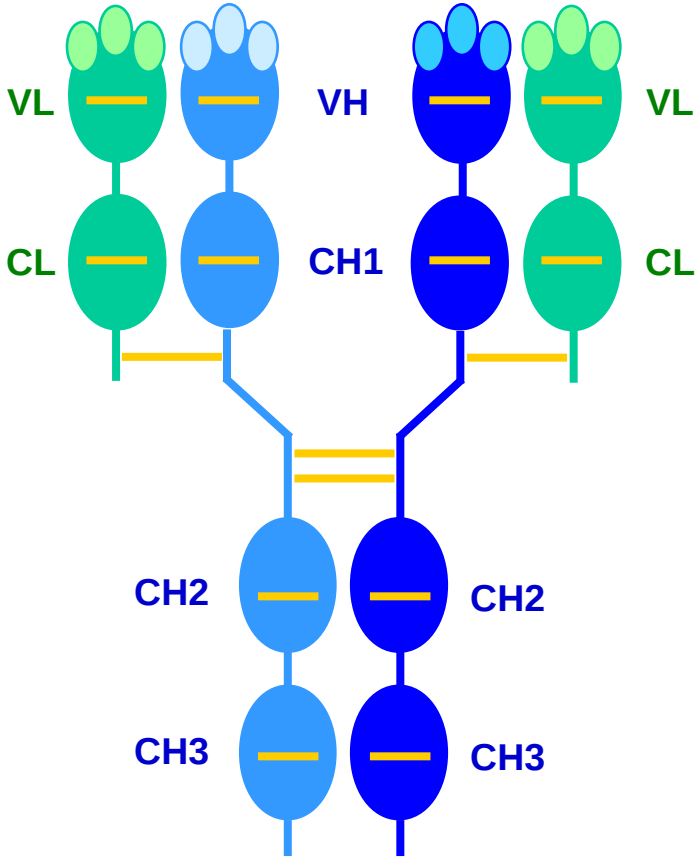
- Icons: question mark

- Select domain boundaries
- Colour CH1-CH4 orange
- Colour CL dark green

ANTIBODY STRUCTURE ANALYSIS

Schematic Ab structure





ANTIBODY STRUCTURE ANALYSIS

Swiss-Pdb Viewer: Highlight disulfide bonds

- Main menu:

- Color: chain
- Select: group kind -> SS bonds

- Control Panel:

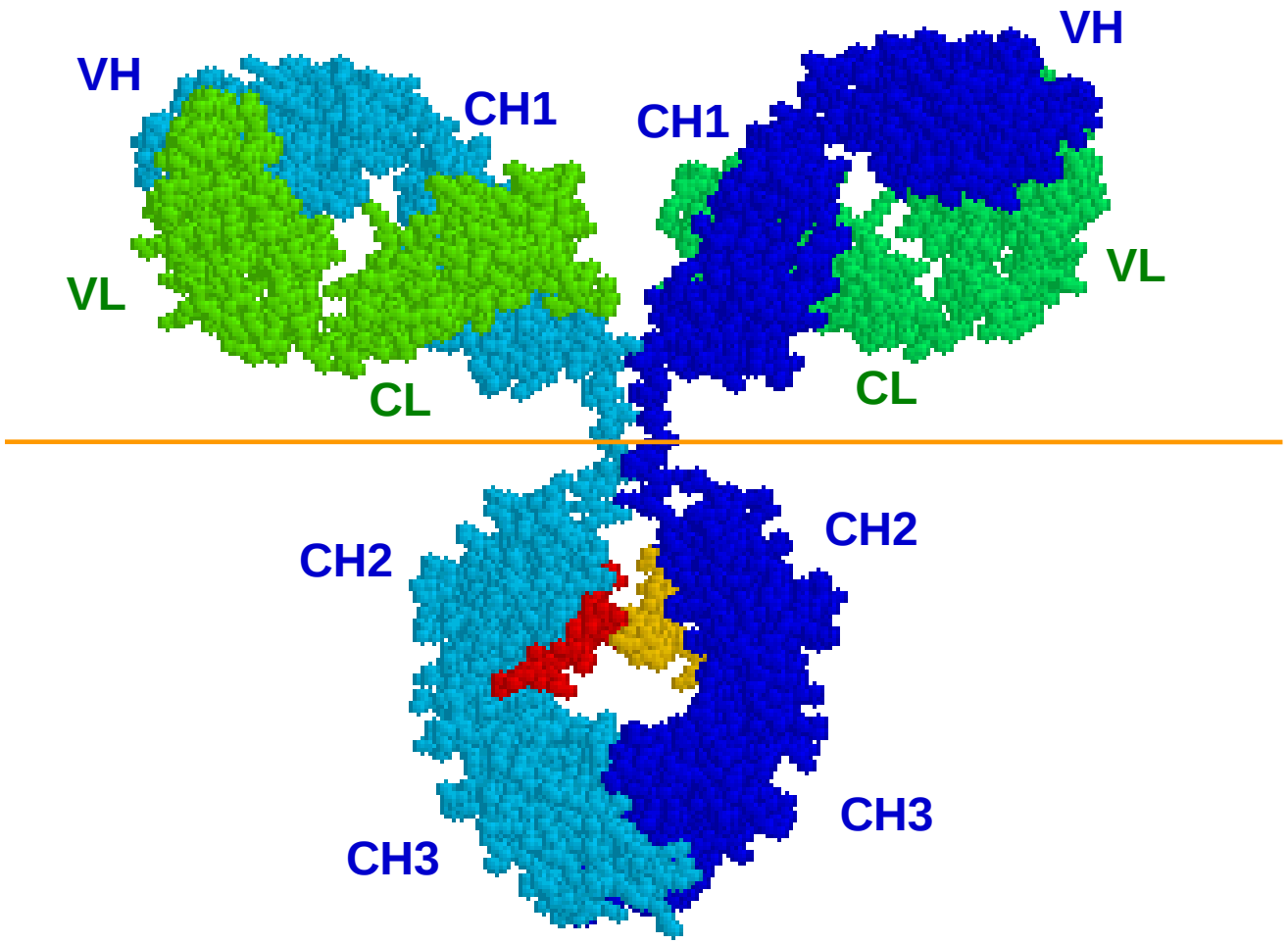
- side -> display side-chains of selected residues (cysteines)

- Main menu:

- Color: selection

ANTIBODY STRUCTURE ANALYSIS

Ab fragments can be generated by proteolytic cleavage

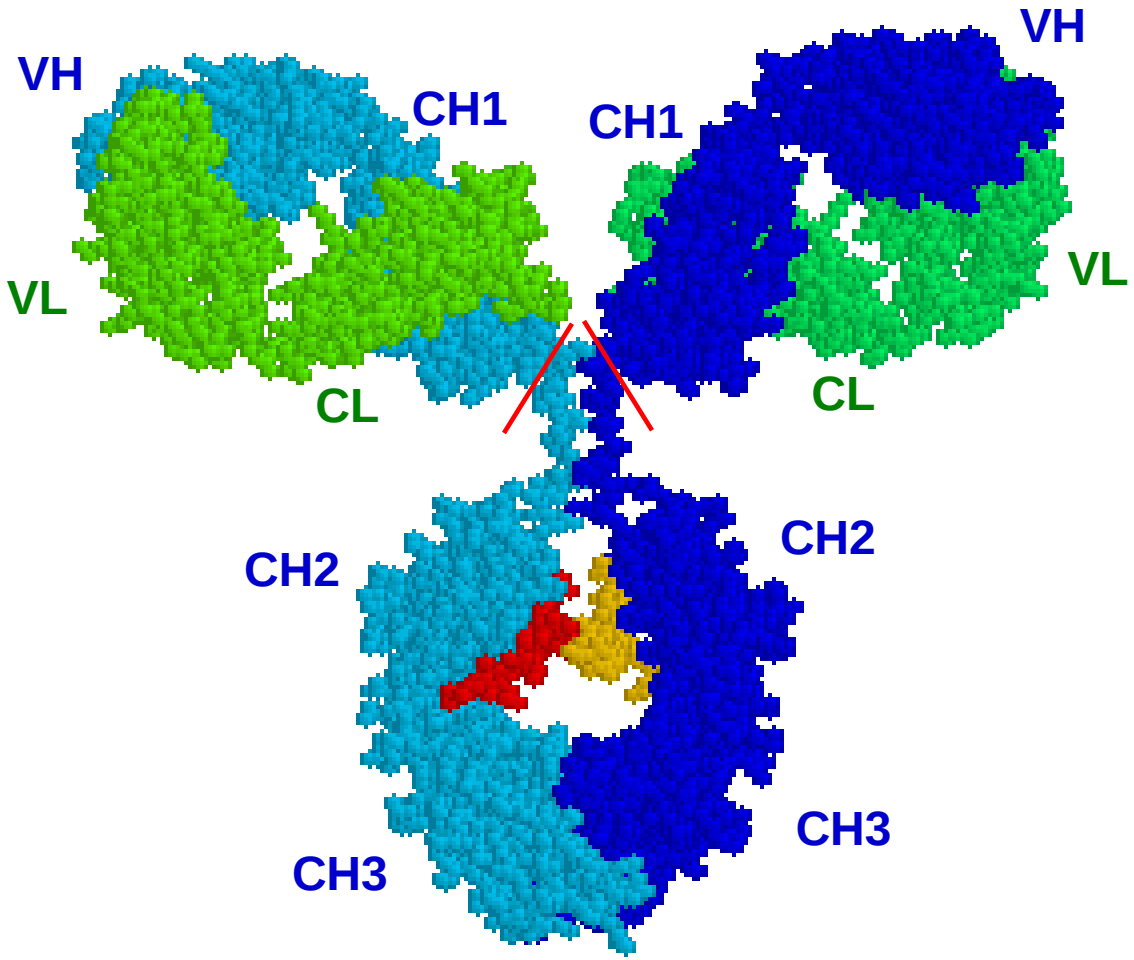


F(ab)2

Fc

ANTIBODY STRUCTURE ANALYSIS

Ab fragments can be generated by proteolytic cleavage



2 Fab

ANTIBODY STRUCTURE ANALYSIS

Swiss-Pdb Viewer: Analyse secondary structure

- Control Panel:

- Show -> Deselect: (H) 233-478; (K); (M)

- Main Menu:

- Display -> Labels –> clear user labels
- Display -> deselect ‘show backbone as carbon alpha traces’
- Tools -> compute H-bonds
- File -> Open -> 1HZH.pdb

- Control Panel:

- select 1HZH (coloured) -> toggle off

- Main Menu:

- Color -> secondary structure; secondary structure succession

ANTIBODY STRUCTURE ANALYSIS

Swiss-Pdb Viewer: Analyse secondary structures and loops

- Control Panel:

- Show -> Deselect: whole chains (K, M)
- Show -> Deselect: CH2-CH3 (Fc) domains of chain (H): 233-478

- Main Menu:

- Display -> deselect ‘show backbone as alpha carbon traces’
- Tools -> compute H-bonds

- Control Panel:

- Show -> Deselect: CL domain of chain (L): 108-
- Show -> Deselect: CH1 domain of chain (H): 114-

- Main Menu:

- Select “?” -> pick residues between b-strands

ANTIBODY STRUCTURE ANALYSIS

Swiss-Pdb Viewer: Analyse secondary structures and loops

Antigen Combining Site (ACS):

- Six hypervariable loops (L1, L2, L3, H1, H2, H3)
- CDR residues

Framework (Fw) Regions:

- ‘scaffold’ supporting the ACS

ACS and Fw regions:

- Very well known sequence-structure-function relationships

ANTIBODY STRUCTURE ANALYSIS

Swiss-Pdb Viewer: Highlight conserved cysteines and tryptophan

- Control Panel:

- Show -> Side: (L) 23, 88, 35
- Colour Dark Green
- Show -> Side: (H) 22, 92, 36
- Colour Red

- Main Menu

- Display -> deselect ‘show backbone as alpha carbon traces’
- Tools -> compute H-bonds

ANTIBODY STRUCTURE ANALYSIS

Swiss-Pdb Viewer: Highlight conserved cysteines and tryptophan

Control Panel:

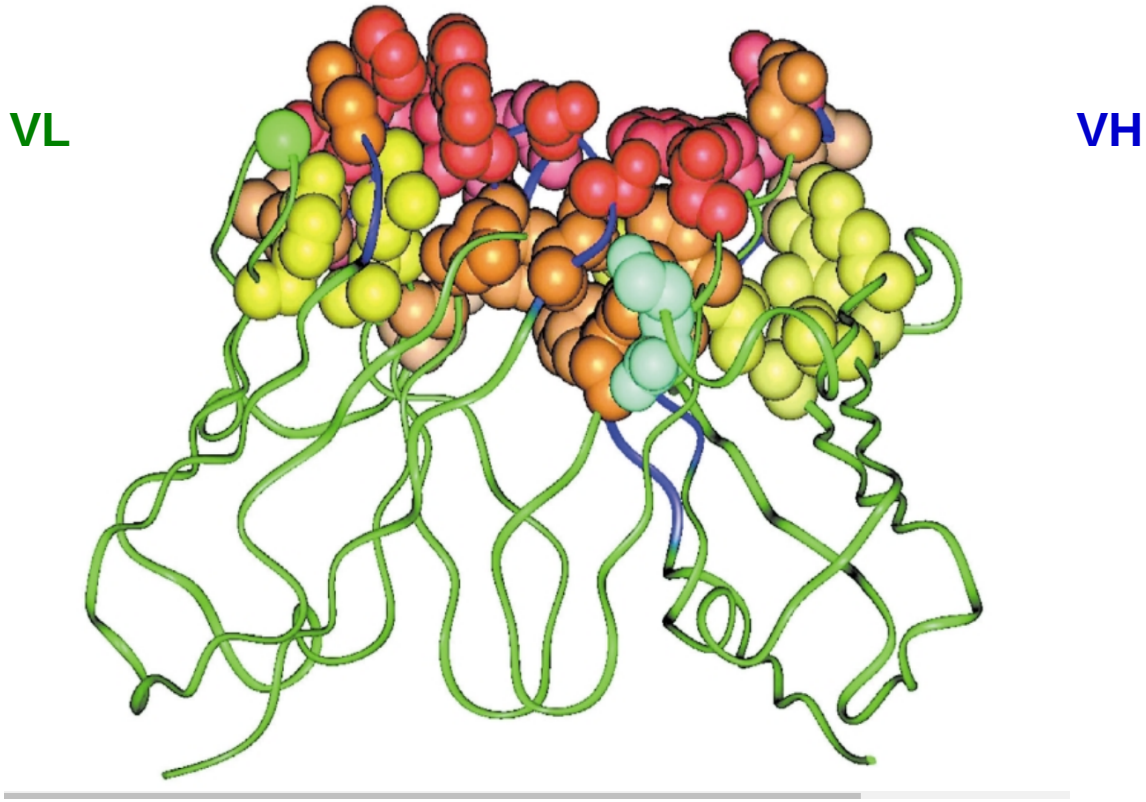
- Select -> Residue: (L) 109-
- Select -> Chain: (K)
- Select -> Chain: (H) 115-

Main Menu:

- Build -> remove selected residues

ANTIBODY STRUCTURE ANALYSIS

1VFB: variabile domains (VL, VH) in complex with Ag



Fv

ANTIBODY STRUCTURE ANALYSIS

Antigen Combining Site (ACS)

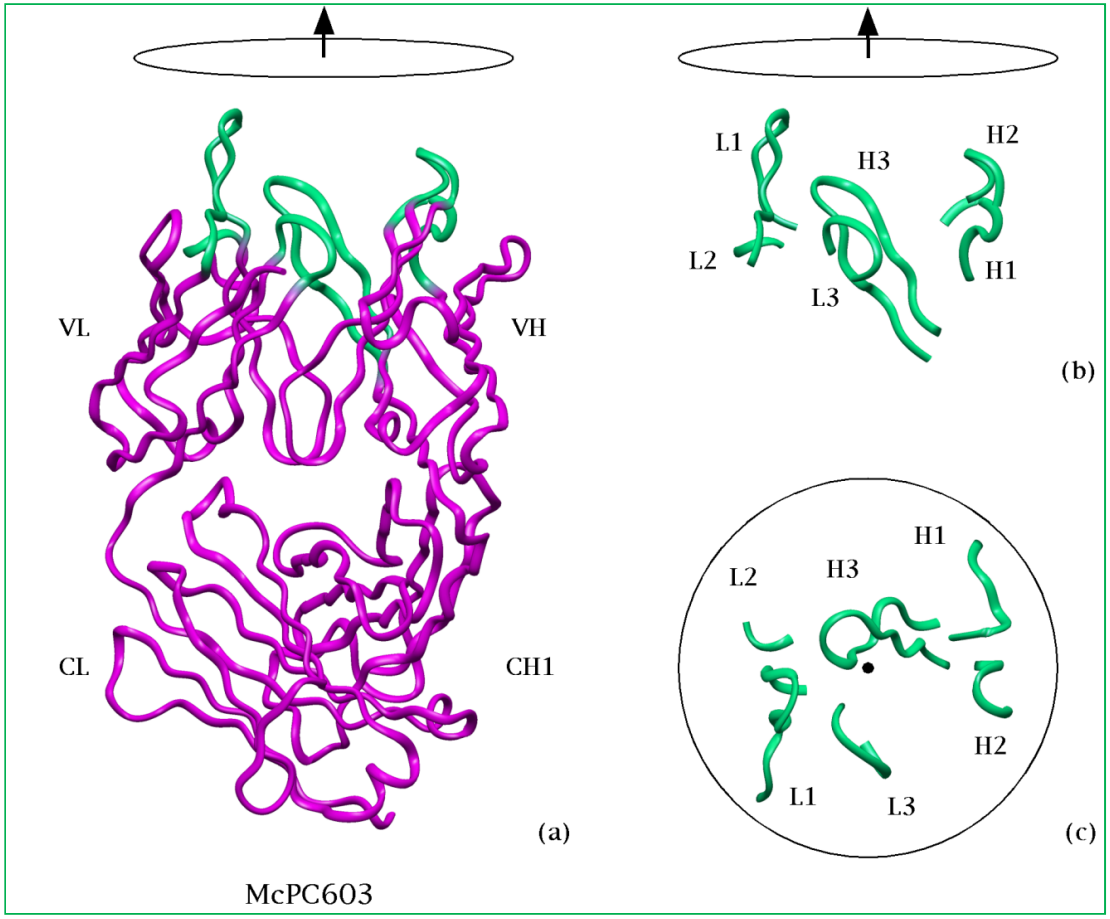
6 loops: 3 in VL, 3 in VH

Hypervariable in sequence (Kabat):

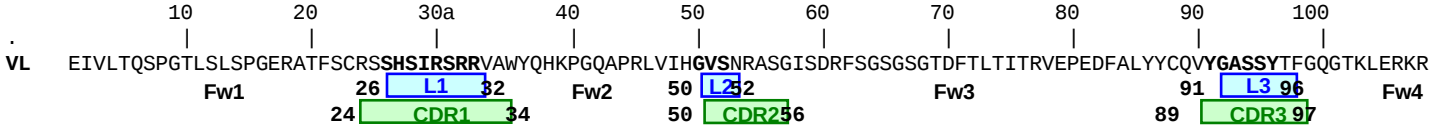
- CDR1, CDR2, CDR3

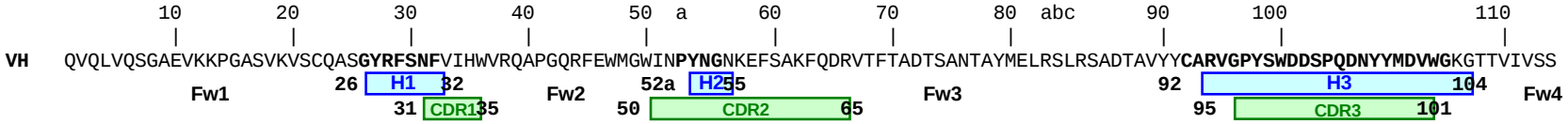
Out of the **β-sheet** framework (Chothia & Lesk):

- L1, L2, L3
- H1, H2, H3



ANTIBODY STRUCTURE ANALYSIS





1VFB	Standard (specific) Loops Nb	Standard (specific) CDRs Nb
	L1: 26-32 (26-32)	CDR1: 24-34 (24-34)
	L2: 50-52 (50-52)	CDR2: 50-56 (50-56)
	L3: 91-96 (91-96)	CDR3: 89-97 (89-97)
	H1: 26-32 (26-32)	CDR1: 31-35 (31-35)
	H2: 52a-55 (52a-55)	CDR2: 50-65 (50-65)
	H3: 92-104 (92-104)	CDR3: 95-101 (95-101)

ANTIBODY STRUCTURE ANALYSIS

Swiss-Pdb Viewer: Colour hypervariable loops and CDRs

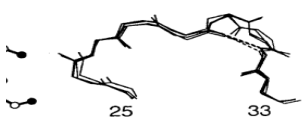
- Control Panel:

- Select: 26-32 (L1), 50-52 (L2), 91-96 (L3)
- Color magenta
- Select: 26-32 (H1), 52a-55 (H2), 92-104 (H3)
- Color orange
- Select: 24, 25, 33, 34; 53-56; 89, 90, 97 (VL-CDRs outside L1, L2, L3)
- Color cyan
- Select: 33-35; 50, 51, 52, 56-65 (VH-CDRs outside H1, H2)
- Color red

ANTIBODY STRUCTURE ANALYSIS

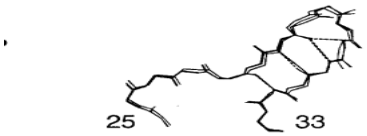
Canonical structures of Ab hypervariable loops

L1



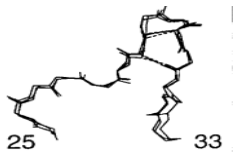
sion angles:

	S.D.	Angle	Mean	S.D.



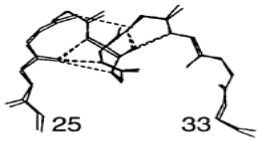
Torsion angles:

ean	S.D.	Angle	Mean	S.D.
-75	4	Ψ	-15	4
-36	1	Ψ	157	3



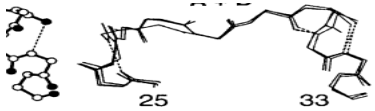
Torsion angles:

ean	S.D.	Angle	Mean	S.D.
	8	Ψ	-22	22



Torsion angles:

Mean	S.D.	Angles	Mean	S.D.
-132	11	Ψ	-148	8



Torsion angles:

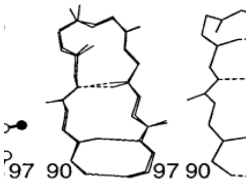
Mean	S.D.	Angle	Mean	S.D.
-99	1	Ψ	137	5
-67	1	Ψ	-31	19

L2

Torsion angles:

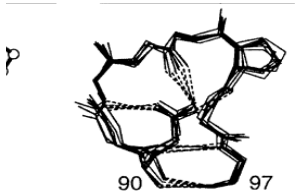
Angle

L3



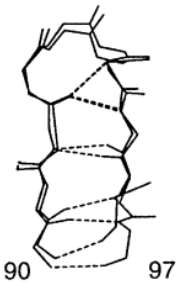
Torsion angles

	A	B	C	Ang
-128	1	-141	-89	Ψ
-96	4	139	142	...



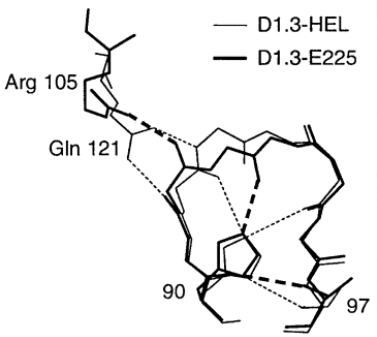
rsion angles:

s.d.	Angle	Mean	s.d.
19	Ψ	131	16
13	Ψ	25	12
23	ω	-39	12



on angles

S.D.	Angle	Mean	S.D.
4	Ψ	150	3



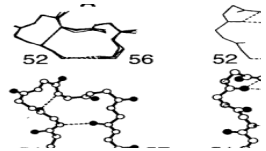
ANTIBODY STRUCTURE ANALYSIS

Canonical structures of Ab hypervariable loops

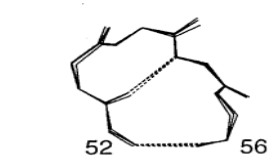
Torsion angles:

n	S.D.	Angle	Mean	S.D.

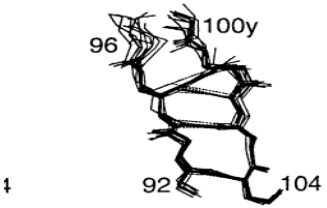
H1



H2

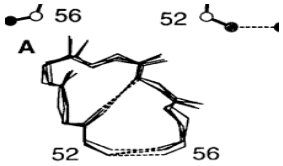


H3



Torsion Angles:

S.D.	Angle	Mean	S.D.
10	ψ	147	6
7	ψ	152	13

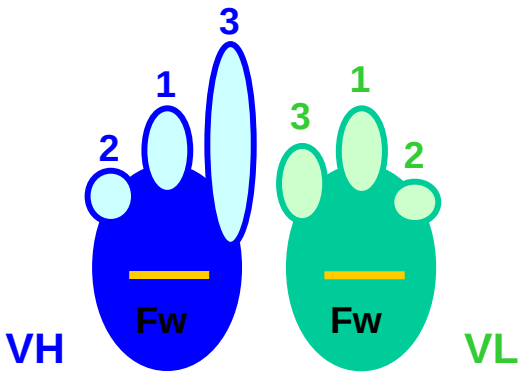


Torsion angles :

A		B		A
an	S.D.		Angle	mean
95	15	-114	Ψ	111

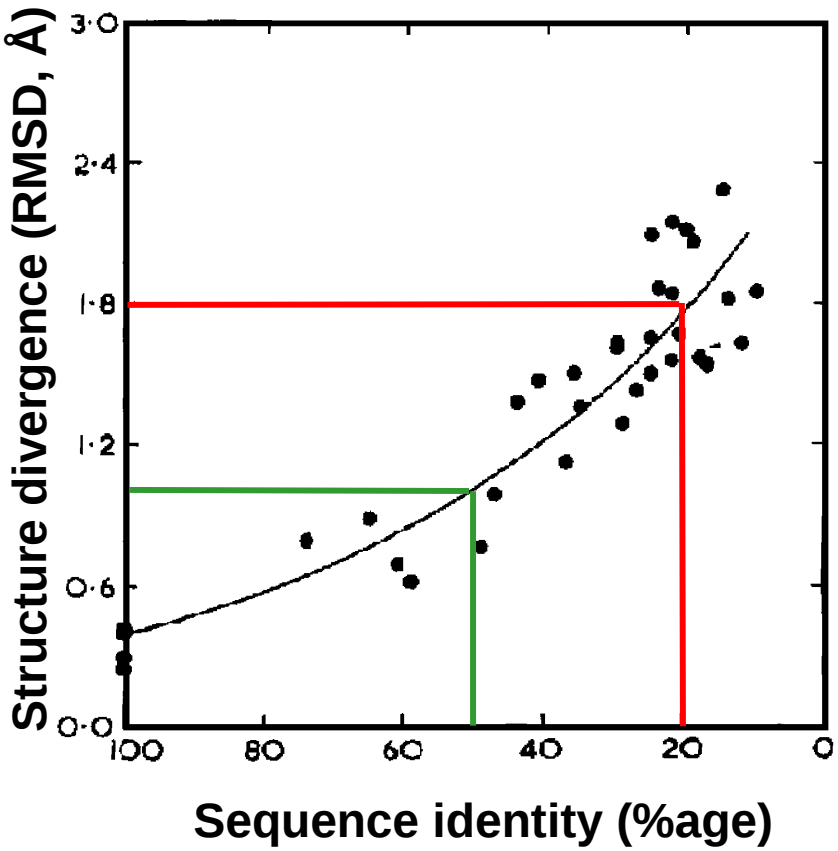
ANTIBODY STRUCTURE ANALYSIS

Fw regions: Structure-Sequence Relationships



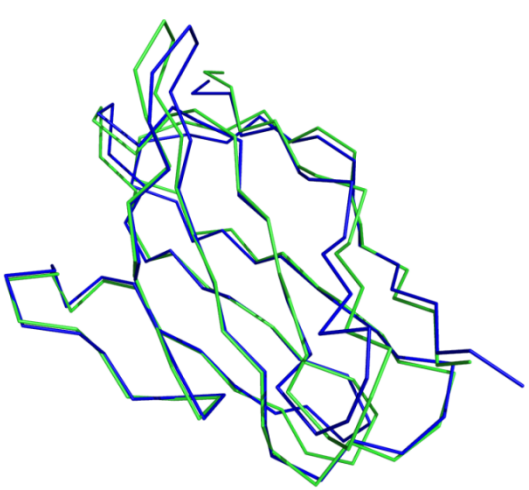
Fw regions:

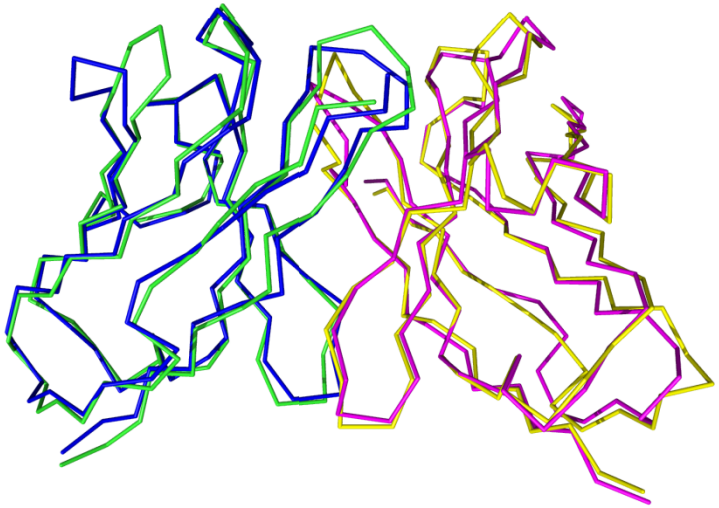
- highly conserved sequences (>50% sequence identity)
- Highly conserved structures (RMSD domains < 1.0 Å)
- Accurate structure predictions (molecular models)



ANTIBODY STRUCTURE ANALYSIS

Fw regions: Sequence-Structure Relationships





VL:

68% sequence identity

RMSD Cα 0.69 Å

VH:

66% sequence identity

RMSD Cα 0.83 Å

PROTEIN RULES

HIGH SEQUENCE IDENTITY

↓

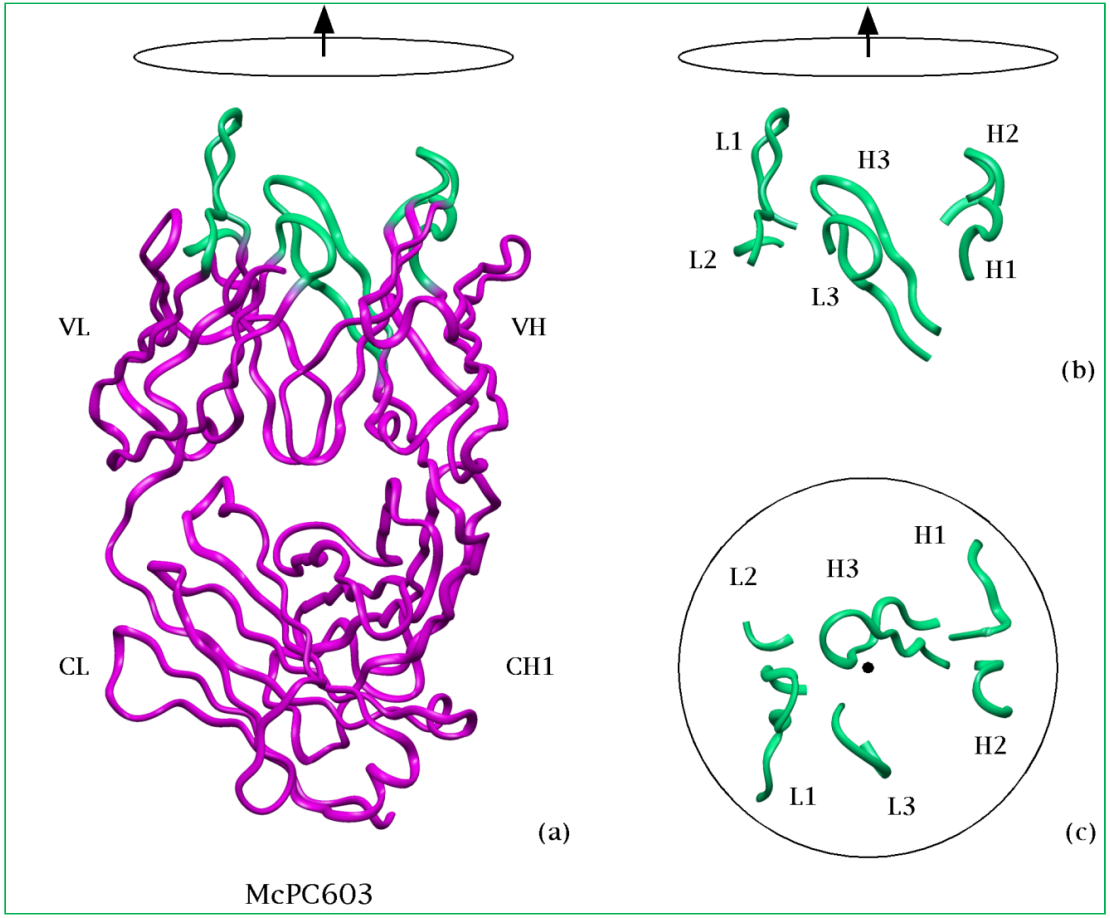
HIGH STRUCTURE SIMILARITY

(a.a. Sequence –> 3D structure)

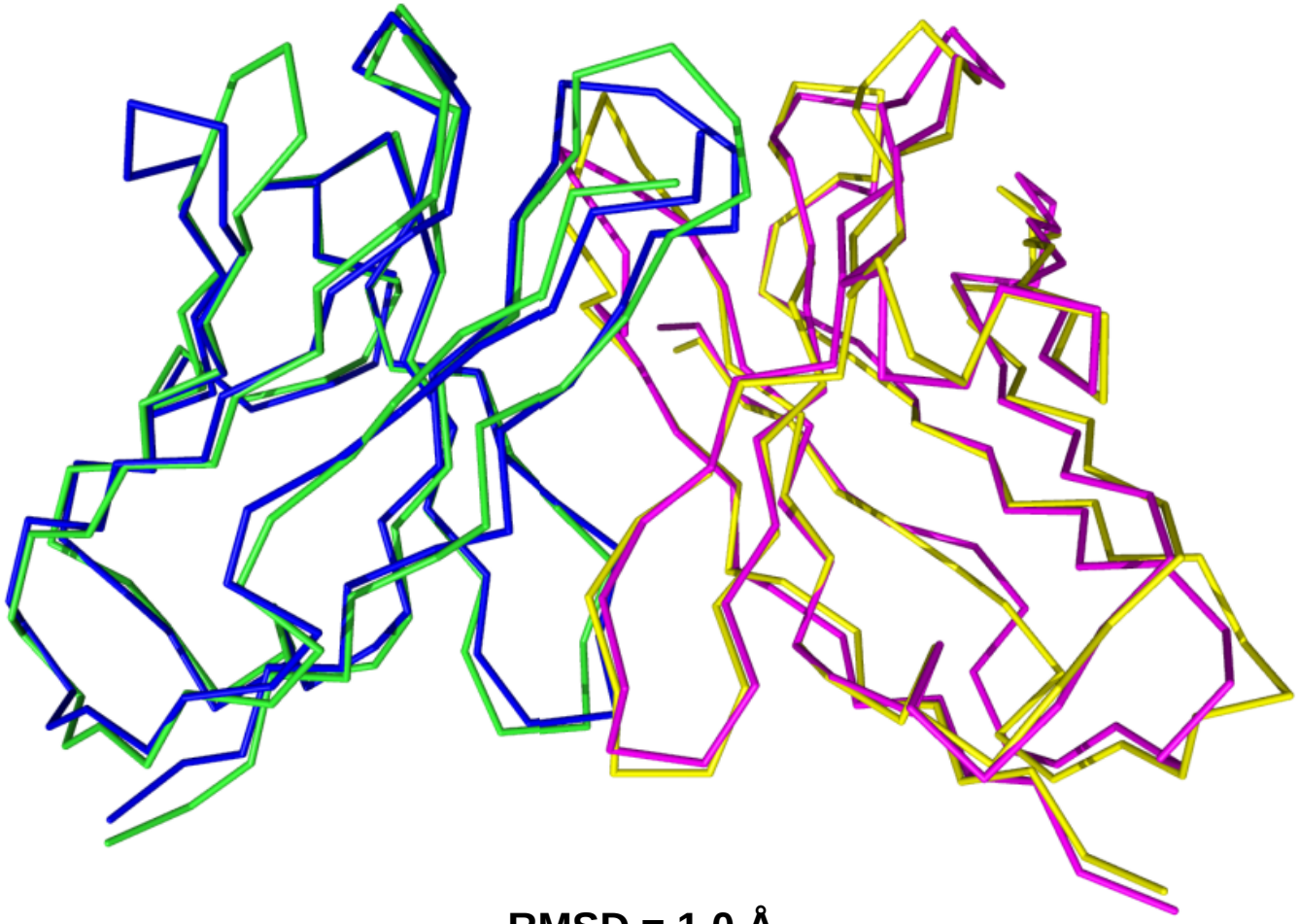
ANTIBODY STRUCTURE ANALYSIS

ACS: Structure-Sequence Relationships

- **6 loops**: 3 in VL, 3 in VH
- **Hypervariable** sequences
- **H3**: variable structure
- **L1, L2, L3, H1, H2**: "Canonical Structures" (Chothia & Lesk)



ANTIBODY STRUCTURE ANALYSIS



RMSD = 1.0 Å

PROTEIN STRUCTURE ANALYSIS

HEMOGLOBIN

3D structures of Hemoglobin are available from the RCSB PDB PROTEIN DATA BANK

2HHB: Human Deoxyhaemoglobin at 1.74 Å Resolution

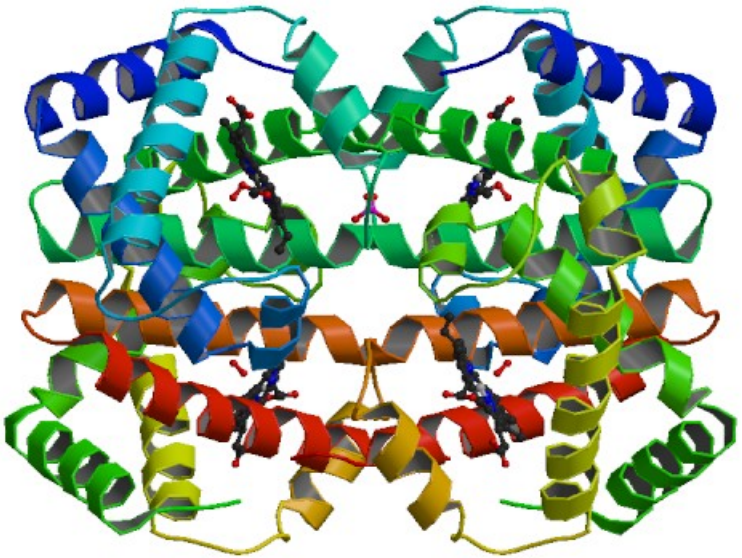
1HHO: Human Oxyhaemoglobin at 2.1 Å Resolution

HAEMOGLOBIN STRUCTURE

Bound to oxygen

Human oxyhaemoglobin

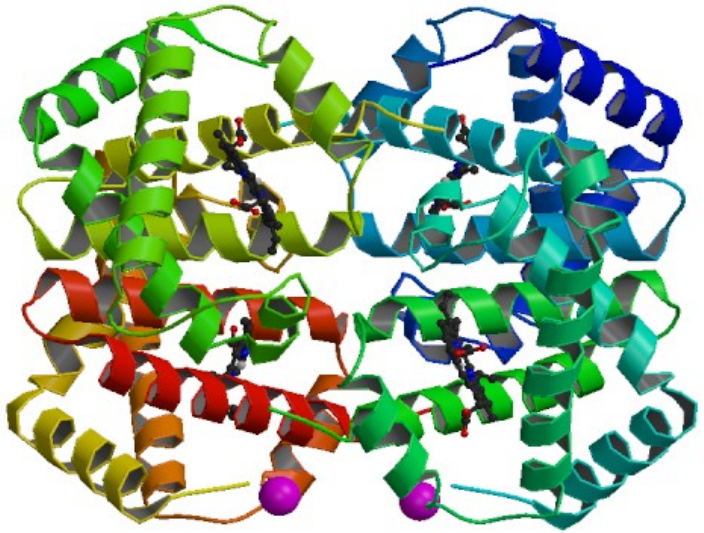
1HHO



Not bound to oxygen

Human deoxyhaemoglobin

2HHB



HAEMOGLOBIN STRUCTURE ANALYSIS

Download from the RCSB PDB PROTEIN DATA BANK

- **1HHO, 2HHB**
- **Download files:**
  - PDB file
  - Fasta sequence

HAEMOGLOBIN STRUCTURE ANALYSIS

1HHO fasta file:

```
>1HHO:A|PDBID|CHAIN|SEQUENCE
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNAL
SALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR

>1HHO:B|PDBID|CHAIN|SEQUENCE
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN
LKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
```

2HHB fasta file:

```
>2HHB:A|PDBID|CHAIN|SEQUENCE
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNAL
SALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
>2HHB:B|PDBID|CHAIN|SEQUENCE
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN
LKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
>2HHB:C|PDBID|CHAIN|SEQUENCE
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNAL
SALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
>2HHB:D|PDBID|CHAIN|SEQUENCE
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN
LKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
```

HAEMOGLOBIN STRUCTURE ANALYSIS

1HHO, 2HHB

- **Open with SwissPDBViewer**

PROTEIN RULES

LOW SEQUENCE IDENTITY

↓

HIGH STRUCTURE SIMILARITY

Protein 3D Structures are more conserved than a.a. sequences

Protein sequence-structure-function relationships

3D-Structures are more conserved than a.a sequences

1alu & 1bgc

1.1 Å RMSD

14/71 = 20 % Seq. ID

1cnt & 1ax8

0.9 Å RMSD

8/71 = 11 % Seq. ID

1hgu &1lki

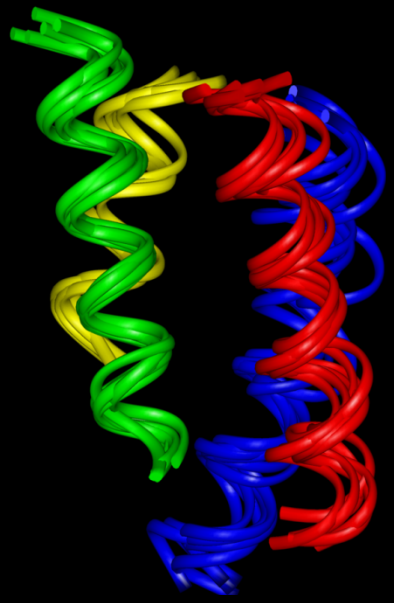
1.7 Å RMSD

9/71 = 13 % Seq. ID

Protein sequence-structure-function relationships

Just a small number of residues (“key-residues”) are required to maintain the protein fold

4-alpha-helical cytokines family

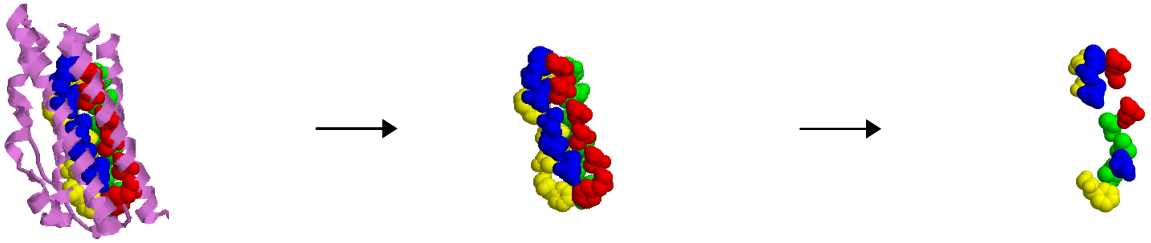


common core: 71 a.a.

11 highly conserved “key-residues” surrounded by 10 a.a. conserving a generic hydrophobic-neutral character

Protein sequence-structure-function relationships

4-helical cytokines family



Common Core

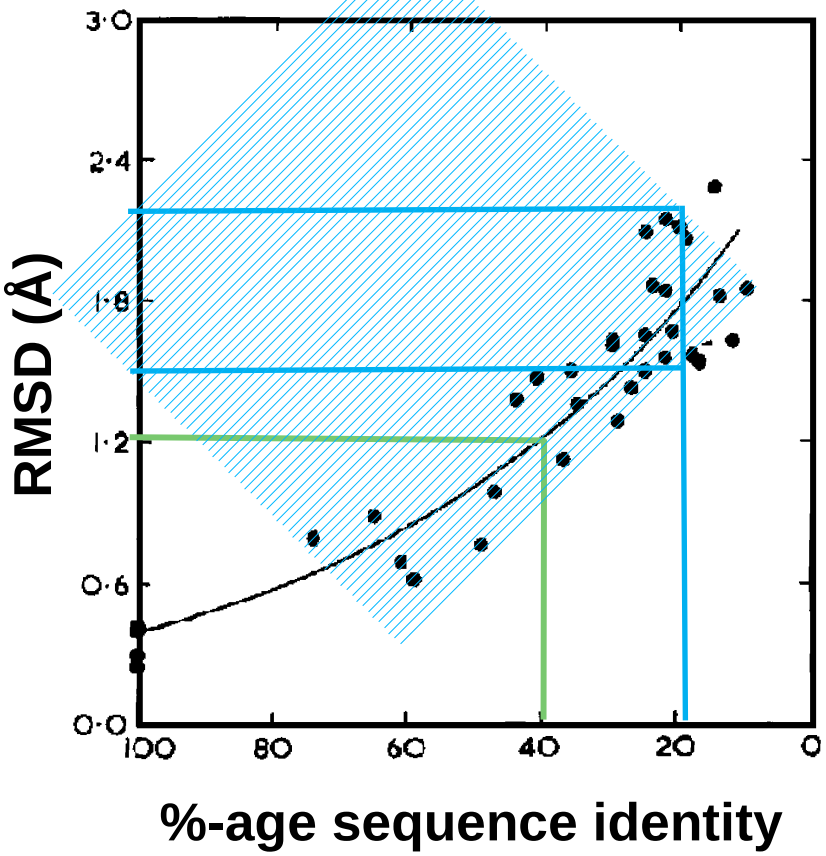
Structure similarity

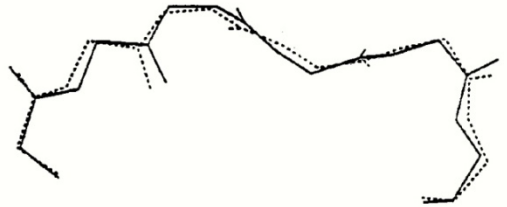
‘Key’-residues

Structure & sequence similarity

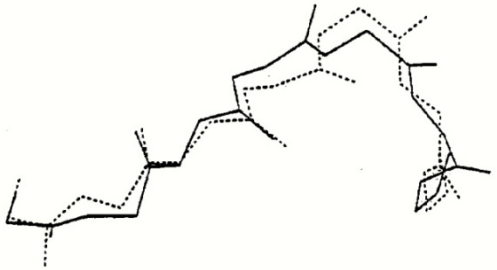
Sequence-structure relationships

Structure similarity

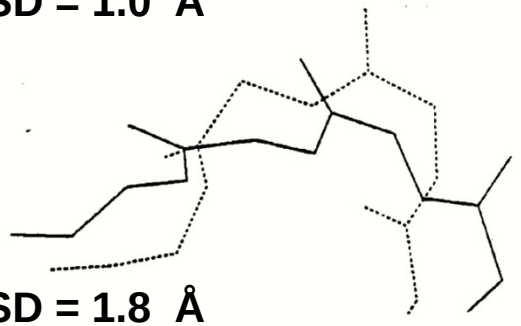




RMSD = 0.5 Å



RMSD = 1.0 Å



RMSD = 1.8 Å

Similar sequences ⟶ **Similar structures**

Different sequences may give Similar structures

PROTEIN RULES

3D STRUCTURES
are
MORE CONSERVED
than
AMINO ACID SEQUENCES

PROTEIN RULES

HIGH STRUCTURE SIMILARITY

↓

HIGH OR LOW SEQUENCE IDENTITY

PROTEIN RULES

STRUCTURALLY (as well as FUNCTIONALLY) IMPORTANT RESIDUES

are

MORE CONSERVED

than

the rest of the structure

PROTEIN RULES

“KEY”
FUNCTIONAL or STRUCTURAL
RESIDUES
are
MORE CONSERVED
than
the rest of the structure

HOW DO I FIND MY PROTEIN?

The Protein Data Bank: www.pdb.org

The National Center for Biological Information (NCBI):
www.ncbi.nlm.nih.gov

The Uniprot Knowledgebase:
http://www.uniprot.org/

HOW DO I FIND MY PROTEIN?

The Protein Data Bank: www.pdb.org

Essential things to do:

1) provide PDB ID (4 characters, 1st is always a number) to go to the page of a specific structure

2) in the structure page, download:
   - Fasta sequence
   - PDB file (atomic co-ordinates)

HOW DO I FIND MY PROTEIN?

The National Center for Biological Information (NCBI):
www.ncbi.nlm.nih.gov

Essential resources:

1) PubMed

2) Gene

PubMed

http://www.ncbi.nlm.nih.gov/pubmed

1) Click on “Advanced”
2) Search by field:
   - Author
   - Author – First
   - Date - Publication
   - Title
   - Title/Abstract
   - Text Word
   - ...

Gene

http://www.ncbi.nlm.nih.gov/gene/

1) Hemoglobin human: HBB

2) Vascular endothelial growth factor receptor 1 human

3) Granulocyte colony-stimulating factor human

Gene page: Hemoglobin (HBB)

Many information, e.g.:

- Interactions
- GO function, process, component
- NCBI Reference Sequences (RefSeq)
  - Genomic: **NG_000007.3**
  - mRNA: **NM_000518.4**
  - Protein: **NP_000509.1**

Click on Protein sequence identifier **(NP_000509.1)**

Protein page: Hemoglobin subunit beta (NP_000509.1)

Many information

- at the end sequence in “human friendly” GenPept format
- at the beginning, click on “FASTA” to get the sequence in Fasta format
- on the right: Run BLAST

Blast

https://blast.ncbi.nlm.nih.gov/Blast.cgi

Pairwise Sequence Comparison Method

- compares the **Input** or **Query sequence** with **each sequence** in the chosen **database**
- produces **pairwise alignments** and **parameters** to evaluate each alignment

Versions

- **blastn: nucleotide** sequence
- **blastp: protein** sequence

Blast input

https://blast.ncbi.nlm.nih.gov/Blast.cgi

Input or **Query** sequence in **Fasta format**

```
>gi|4503079|ref|NP_000750.1| granulocyte
colony-stimulating factor isoform a
precursor [Homo sapiens]
MAGPATQSPMKLMALQLLLWHSALWTVQEATPLGPASSLPQSF
LLKCLEQVRKIQGDGAALQEKLVSECATYKLCHPEELVLLGHS
LGIPWAPLSSCPSQALQLAGCLSQLHSGLFLYQGLLQALEGIS
PELGPTLDTLQLDVADFATTIWQQMEELGMAPALQPTQGAMPA
FASAFQRRAGGVLVASHLQSFLEVSYRVLRHLAQP
```

Blast input

https://blast.ncbi.nlm.nih.gov/Blast.cgi

Sequence Database

PDB: sequences of proteins whose 3D structure is known

NR: all known protein sequences

RefSeq: curated set of sequences from NR

SwissProt: curated set of sequences from UniProt

Organism: only sequences from the selected species

Exclude: sequences from all but the selected species

Blast input

https://blast.ncbi.nlm.nih.gov/Blast.cgi

Algorithm parameters

Max target sequences: 5000

Show results in a new window

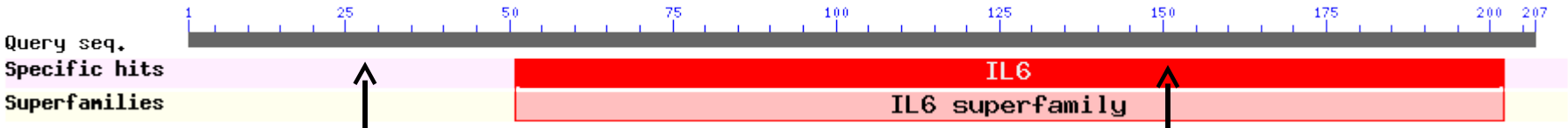
BLAST

Blast output

https://blast.ncbi.nlm.nih.gov/Blast.cgi

1) Conserved domains (CDD)

Putative conserved domains have been detected, click on the image below for detailed results.



The N-terminal region (1-50) is not similar to known protein domains

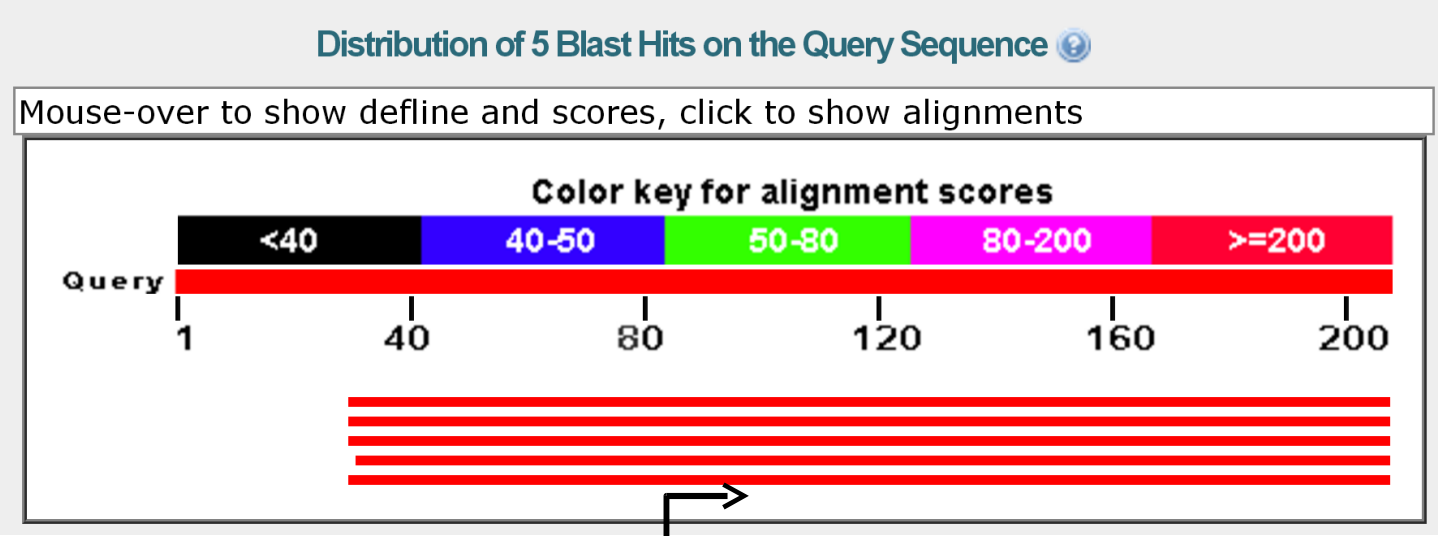
The C-terminal region (50-200) is assigned to the interleukin 6 family

Clicking on the domain picture opens a page with information about the selected domain

Blast output

https://blast.ncbi.nlm.nih.gov/Blast.cgi

2) Graphic view of matched sequences

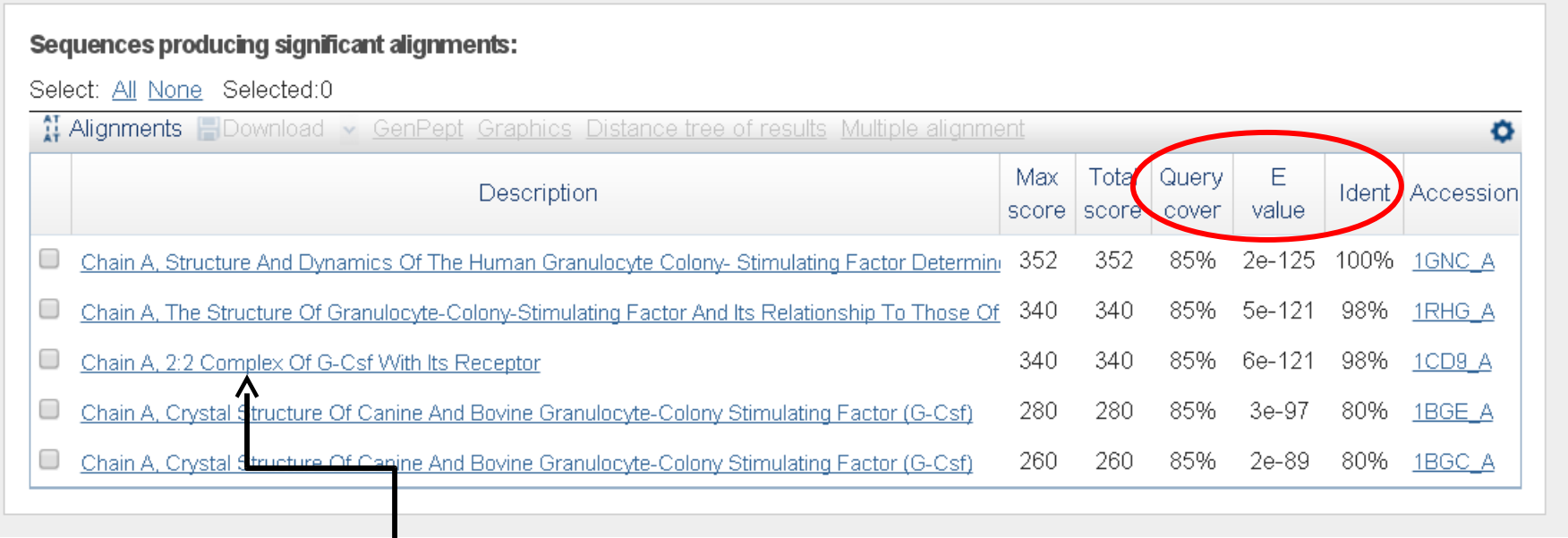


Only the query region corresponding to the IL6 domain is similar to proteins of known 3D structures

Blast output

https://blast.ncbi.nlm.nih.gov/Blast.cgi

3) Summary of sequences similar to the query



Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
Chain A, Structure And Dynamics Of The Human Granulocyte Colony- Stimulating Factor Determin	352	352	85%	2e-125	100%	1GNC_A
Chain A, The Structure Of Granulocyte-Colony-Stimulating Factor And Its Relationship To Those Of	340	340	85%	5e-121	98%	1RHG_A
Chain A, 2:2 Complex Of G-Csf With Its Receptor	340	340	85%	6e-121	98%	1CD9_A
Chain A, Crystal Structure Of Canine And Bovine Granulocyte-Colony Stimulating Factor (G-Csf)	280	280	85%	3e-97	80%	1BGE_A
Chain A, Crystal Structure Of Canine And Bovine Granulocyte-Colony Stimulating Factor (G-Csf)	260	260	85%	2e-89	80%	1BGC_A

Clicking on sequence description takes you to the corresponding pairwise alignment

Alignment parameters

Blast output

https://blast.ncbi.nlm.nih.gov/Blast.cgi

4) Pairwise alignment

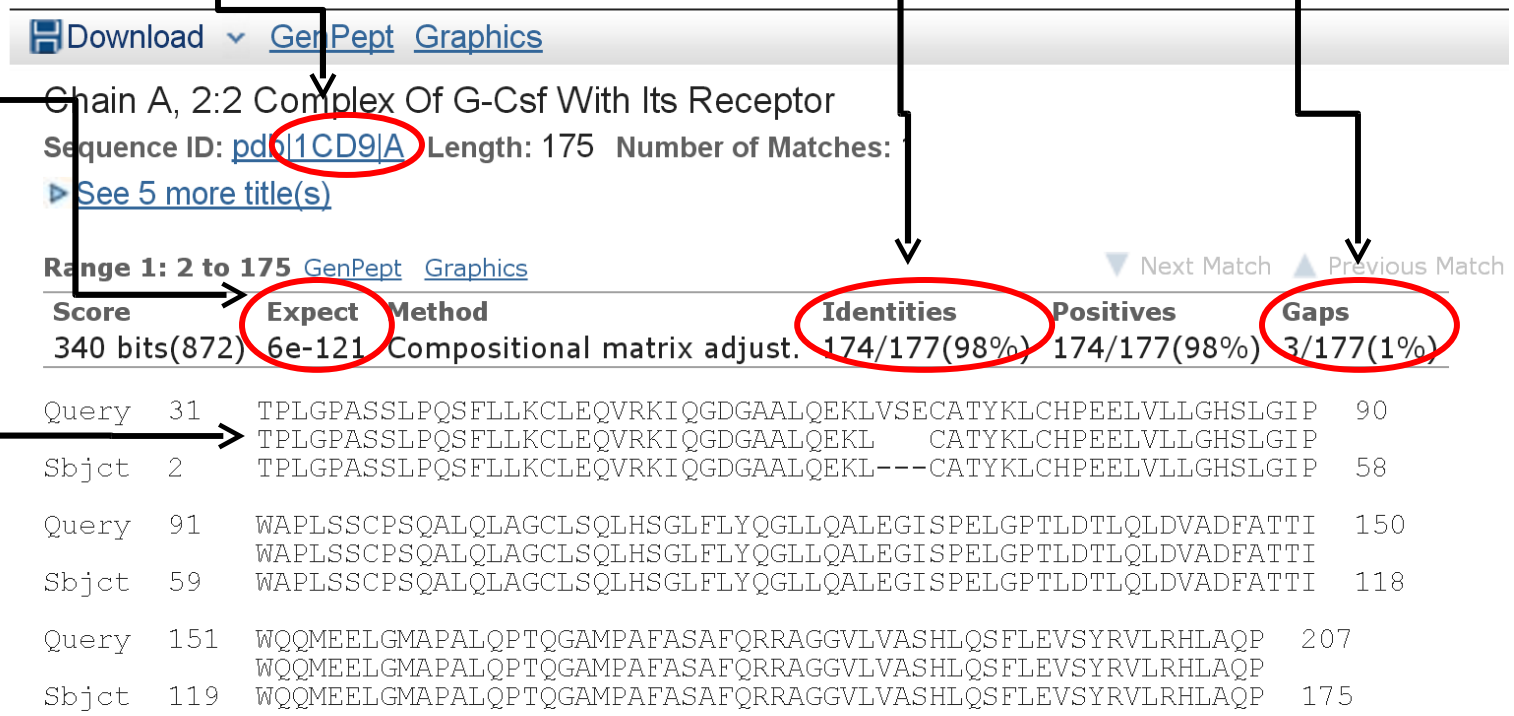
PDB code (1CD9) and chain name (A)

%age of identical residues between "query" and "sbjct"

%age of insertions and deletions (dashes) between "query" and "sbjct")

E value

Identical residues between "query" and "sbjct" (or "hit" or "match")



Blast output

https://blast.ncbi.nlm.nih.gov/Blast.cgi

4) Pairwise alignment

Ident → % of sequence identity in the aligned region

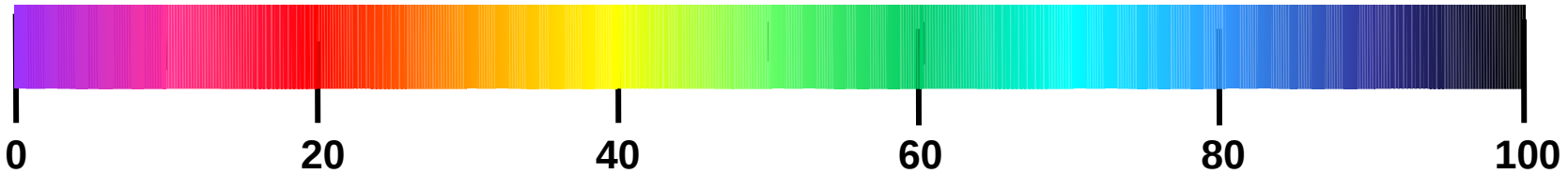
Query cover → %age of input sequence matched

E-value → probability that the matched sequence is **not** homologous

Accession → page with protein description

%age sequence identity

Homologous or **Not-homologous?**



Random (~ 20%)

‘Twilight’

(Close) Homology (> 40%)

Length ~ 100 a.a.

Homology:

< Length, > %_ID (e.g., 10 a.a.)

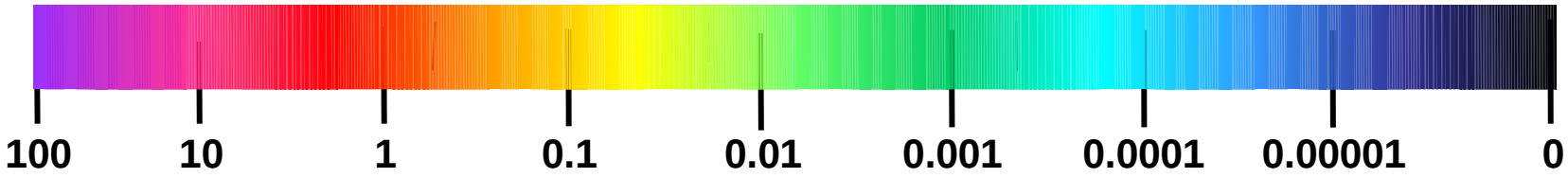
> Length, < %_ID

Not ‘low-complexity’

‘low complexity’ can be ‘masked’

Expect (E) value

Homologous or **Not-homologous?**



Uncertainty **Homology**

E-value
< 0.001-0.0001

Blast output

https://blast.ncbi.nlm.nih.gov/Blast.cgi

3) Summary of sequences similar to the query

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	Chain A, Structure And Dynamics Of The Human Granulocyte Colony- Stimulating Factor Determin	352	352	85%	2e-125	100%	1GNC_A
☐	Chain A, The Structure Of Granulocyte-Colony-Stimulating Factor And Its Relationship To Those Of	340	340	85%	5e-121	98%	1RHG_A
☐	Chain A, 2:2 Complex Of G-Csf With Its Receptor	340	340	85%	6e-121	98%	1CD9_A
☐	Chain A, Crystal Structure Of Canine And Bovine Granulocyte-Colony Stimulating Factor (G-Csf)	280	280	85%	3e-97	80%	1BGE_A
☐	Chain A, Crystal Structure Of Canine And Bovine Granulocyte-Colony Stimulating Factor (G-Csf)	260	260	85%	2e-89	80%	1BGC_A

>80% sequence identity over >150 residues

E-values < 2e-89

All matched sequences are homologous to the "query" (EASY!)

Blast output

https://blast.ncbi.nlm.nih.gov/Blast.cgi

4) Pairwise alignment

%age of sequence identity well below the homology threshold (35-40%)

Expect (E) value above the homology threshold (~0.001)

```
>dbj|BAG62733.1|  unnamed protein product [Homo sapiens]
Length=198

 GENE ID: 3569 IL6 | interleukin 6 (interferon, beta 2) [Homo sapiens]
(Over 100 PubMed links)

 Score = 33.5 bits (75),  Expect = 0.40, Method: Compositional matrix adjust.
 Identities = 19/77 (24%), Positives = 37/77 (48%), Gaps = 1/77 (1%)

Query  3    EQVRKIQDDGAALQEKLCATYKLCHPEELVLLGHSLGIP-WAPLSSCPSQALQLAGCLSQ  61
            +Q+R I D  +AL+++ C    +C   +  L  ++L +P  A    C          CL +
Sbjct  55   KQIRYILDGISALRKETCNKSNMCESSKEALAENNLNLPKMAEKDGCFQSGFNEETCLVK  114

Query  62   LHSGLFLYQGLLQALEG  78
            + +GL  ++  L+ L+
Sbjct  115  IITGLLEFEVYLEYLQN  131
```

DIFFICULT!!!

Homologous or **Not-homologous?**

Expect (E) value: number of matches with a given score

"expected to be found merely by chance"

Threshold problem:

Matches above (better than) threshold

Homologues
True Positives, TP

E-value threshold

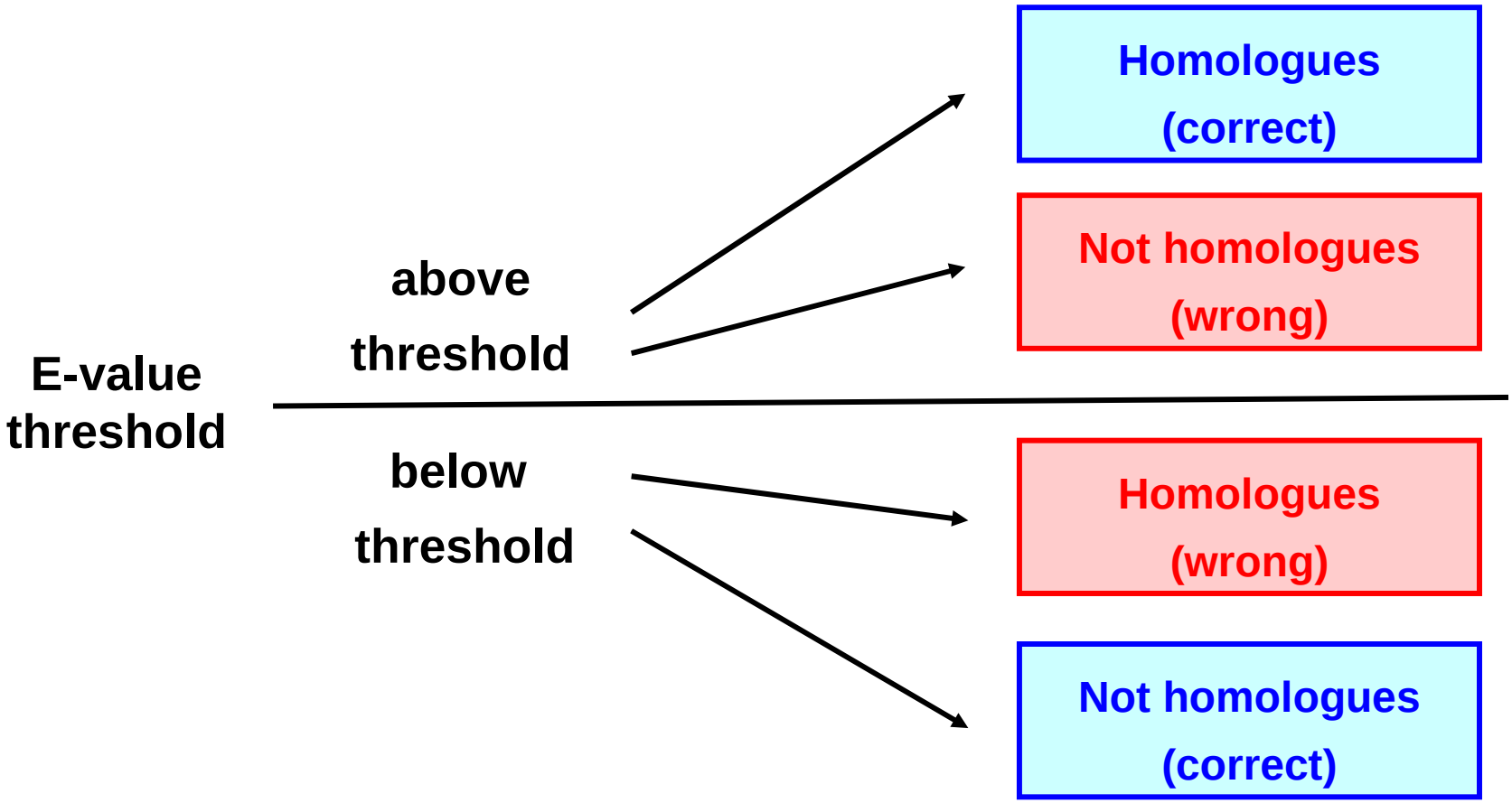
Matches below (worse than) threshold

Not homologues
True Negatives, TN

Homologous or **Not-homologous?**

Expect (E) value: number of matches with a given score

"expected to be found merely by chance"



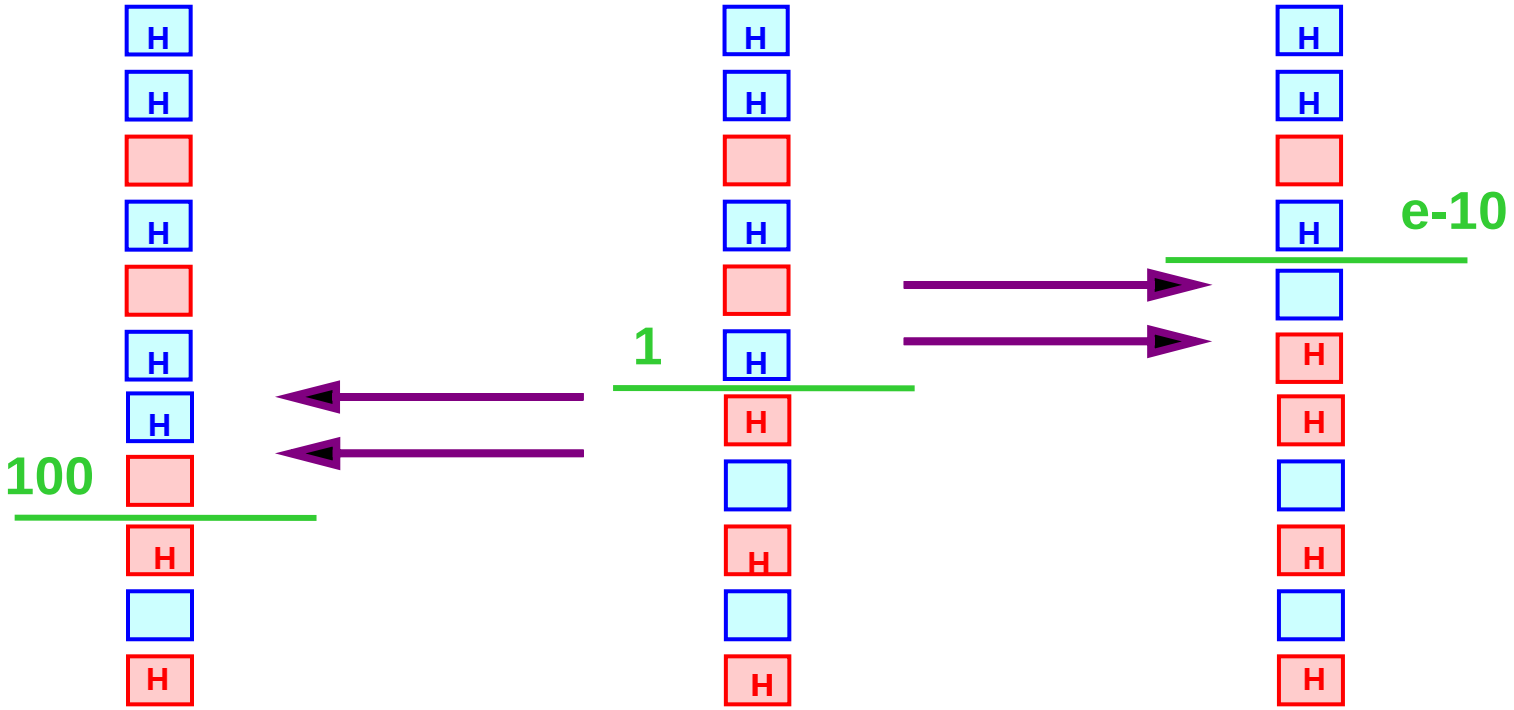
Threshold problem

- lower the threshold: decrease false positives, increase false negatives
- increase the threshold: increase false positives, decrease false negatives
- it is rarely possible to eliminate both false positives and false negatives

H: Homologue

Correct Result
(H above threshold,
Not H below threshold)

Wrong Result
(H below threshold,
Not H above threshold)



Homology

Homology = Evolutionary relationship = Common ancestor

'*High*' or '*Low*' Homology	**No!!!**
'**Close**' or '**Distant**' Homology	**Yes**
'*Measured*' Homology	**No**

Homology is inferred from measurable parameters:

- %-age of sequence identity
- structure similarity (RMSD)

Homologous or **Not-homologous?**

“key-residues” conservation

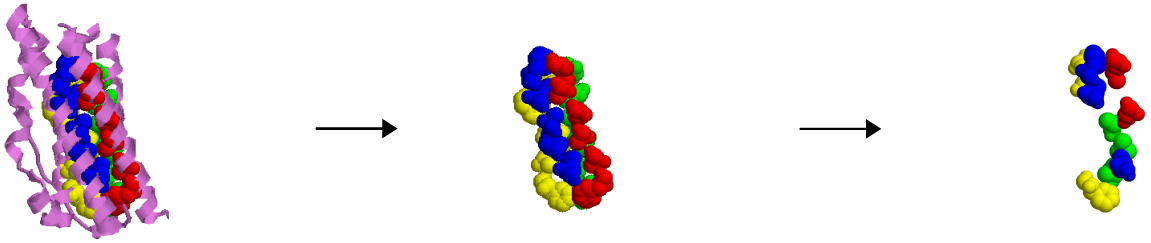
Pair-wise sequence comparison methods do not recognize “key-residues” for protein structure/function

All positions of the alignment are the same and have the same weight on the computed parameters (i.e., %_ID, E-value, etc.)

```
>pdb|1ALU|A Chain A, Human Interleukin-6
Length=186

 Score = 32.7 bits (73), Expect = 0.55, Method: Compositional matrix adjust.
Identities = 19/77 (24%), Positives = 37/77 (48%), Gaps = 1/77 (1%)

Query  3    EQVRKIQDDGAALQEKLCATYKLCHPEELVLLGHSLGIP-WAPLSSCPSQALQLAGCLSQ  61
            +Q+R I D  +AL+++ C    +C   +  L  ++L +P  A    C          CL +
Sbjct  29   KQIRYILDGISALRKETCNKSNMCESSKEALAENNLNLPKMAEKDGCFQSGFNEETCLVK  88

Query  62   LHSGLFLYQGLLQALEG  78
            + +GL  ++  L+ L+
Sbjct  89   IITGLLEFEVYLEYLQN  105
```

Protein sequence-structure-function relationships

4-helical cytokines family



Common Core

Structure similarity

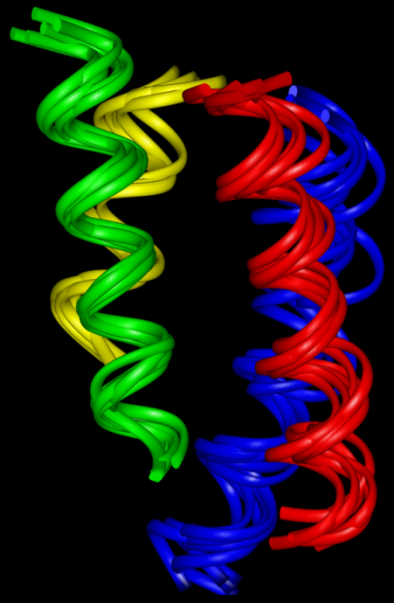
‘Key’-residues

Structure & sequence similarity

Protein sequence-structure-function relationships

Just a small number of residues (“key-residues”) are required to maintain the protein fold

4-alpha-helical cytokines family



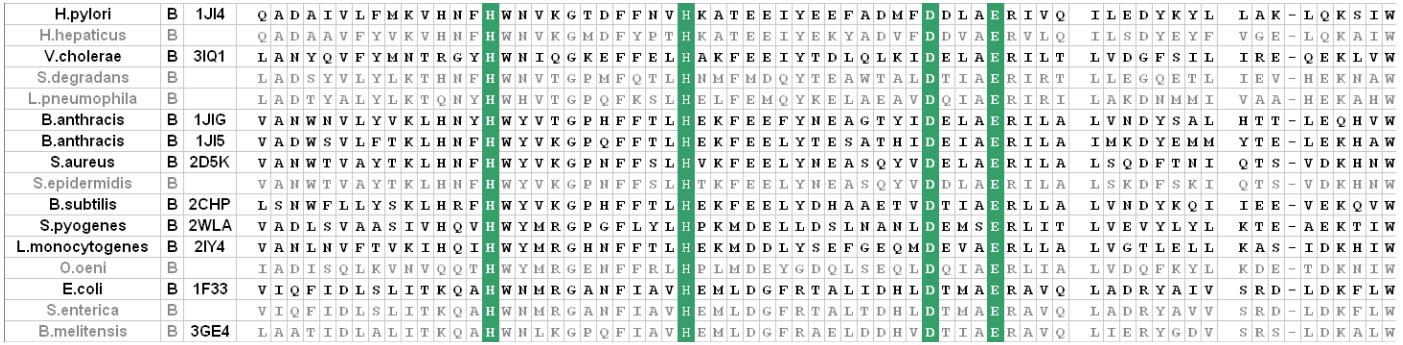
common core: 71 a.a.

11 highly conserved “key-residues” surrounded by 10 a.a. conserving a generic hydrophobic-neutral character

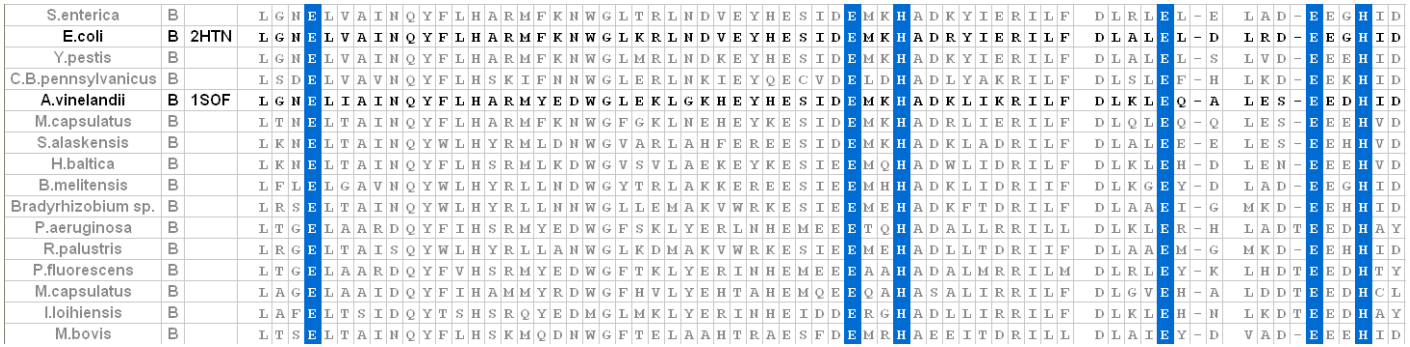
Multiple sequence alignments (MSA)

- **Different positions have different conservation**
- **May allow to recognize “key-residues” for protein structure/function**

Dps proteins



Bacterioferritins



Multiple sequence alignments (MSA)

From Blast:

COBALT alignment, OR:

1) Get sequences to align:

- putative homologs detected from a Blast search (saved as text)

2) Align all sequences to one another

- **Clustal:** http://www.ebi.ac.uk/Tools/msa/clustalo/

Multiple sequence alignments (MSA)

Edit/Visualize MSA:

- **ClustalX (ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/)**
- **JalView (http://www.jalview.org/download.html)**
- **BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html)**
- **WebLogo (http://weblogo.berkeley.edu/logo.cgi)**

HOW DO I FIND MY PROTEIN?

The Uniprot Knowledgebase:
http://www.uniprot.org/

Key-sites (1)

1) **PDB** www.pdb.org

2) **National Center for Biotechnology Information** (NCBI): www.ncbi.nlm.nih.gov

3) **European Bioinformatics Institute** (EBI): www.ebi.ac.uk

4) **Swiss Institute of Bioinformatics** (SIB): www.isb-sib.ch

5) **UniProt** www.uniprot.org/

Key-sites (2)

6) **Nucleic Acid Research** (NAR): nar.oxfordjournals.org

- Database Issue → Compilation paper (Tables)
- Web Server Issue → Editorial

7) **Wikipedia**: www.wikipedia.org

8) **Google**: www.google.com

If you have any questions...

Veronica Morea

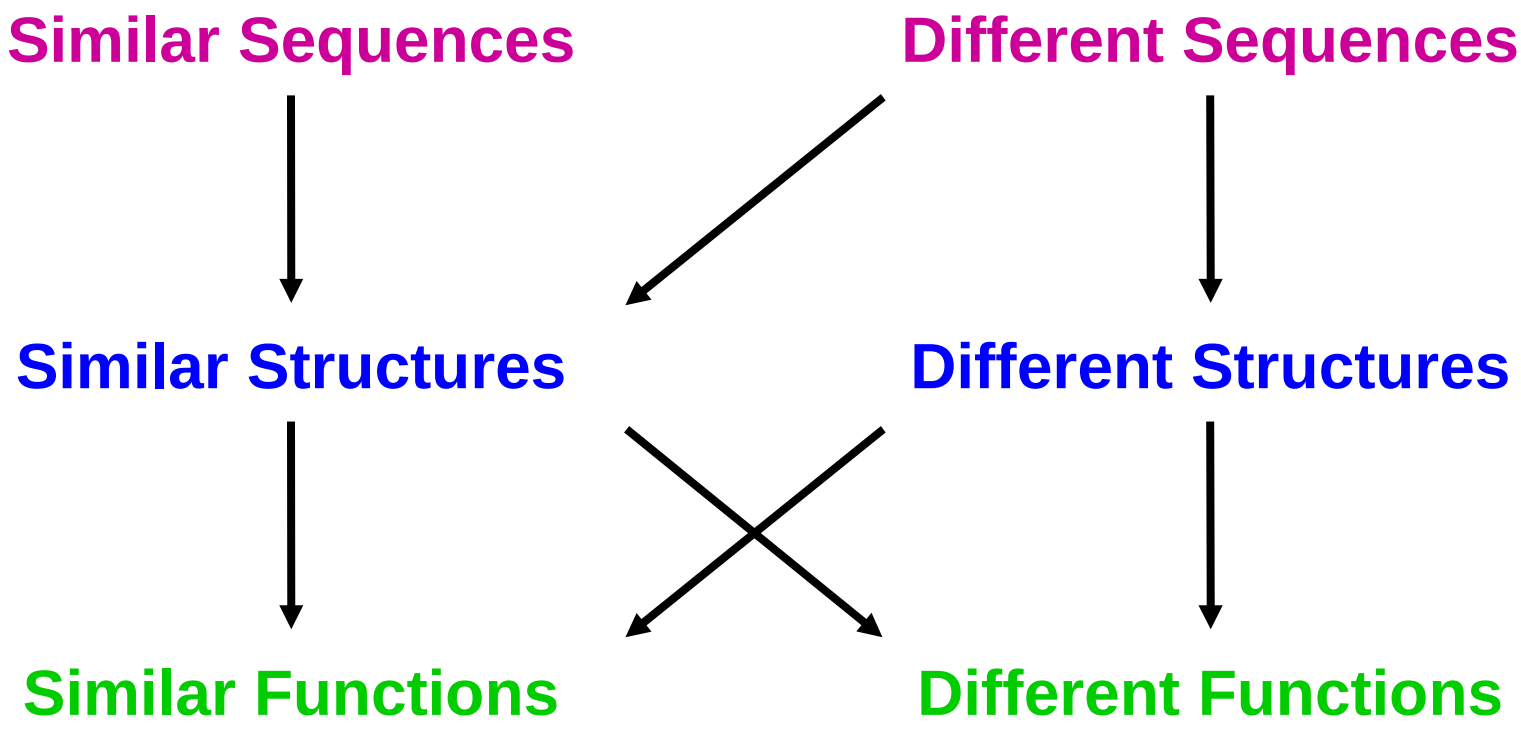
CNR-National Research Council of Italy
Institute of Molecular Biology and Pathology (IBPM)

c/o Department of Biochemical Sciences "A. Rossi Fanelli“
"Sapienza" University of Rome, P.le Aldo Moro 5
00185 - Rome, Italy
Room S26

Tel. +39-06-49910556
Fax: +39-06-4440062

e-mail: veronica.morea@uniroma1.it

Protein sequence-structure-function relationships



Protein sequence-structure-function relationships