Protein Struktur (optional, flexible)

Andrew Torda, Wintersemester 2009 / 2010, AST

- nur für Informatiker, Mathematiker, ..
- 26 Okt, 3 Nov 2009

Proteins - who cares ?

- Most important molecules in life ? Ask the DNA / RNA people

- structural (keratin / hair)
- enzymes (catalysts)
- messengers (hormones)
- regulation (bind to other proteins, DNA, ..)
- industrial – biosensors to washing powder
- receptors
- transporters (O_2, sugars, fats)
- anti-freeze …

Proteins are easy

- data (protein data bank, www.rcsb.org)
  - 61 000 structures
- literature on function, interactions, structure
- software
  - viewers, molecular dynamics simulators, docking, ..
- nomenclature and rules

Proteins are not friendly

- one cannot take a sequence and predict structure /function
- data formats are full of surprises, mostly old formats
- data contains error and mistakes

Protein Rules

- Physics /chemistry versus rules / dogma / beliefs / folklore
- Physics / Chemistry
  - protein + water = set of interacting atoms
    - can be calculated (not really)
- Rules (not quantified)
  - proteins unfold if you heat them (exceptions ?)
  - if they contain lots of charged amino acids, they are soluble
  - if they are more than 300 residues, they have more than one domain,
  - proteins fold to a unique structure (could you prove this ?)
    - lowest free energy structure

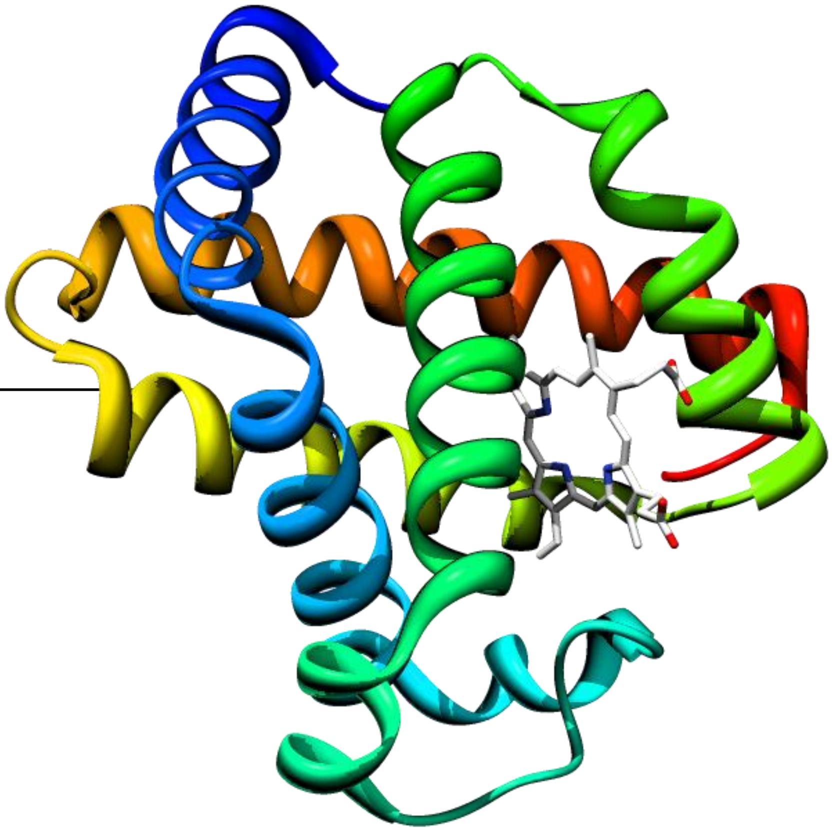
Protein chemistry

- Chemists / biochemists may sleep (quietly)
- Short version
  - proteins are sets of building blocks (amino acids, residues, Reste)
  - 20 types of residue
  - chains of length few to 10^3 (100 or 200 typical)
  - small ones ($< \approx 50$) are peptides

- Longer version

Sizes

- 1 Å = 10^{-10} m or 0.1 nm

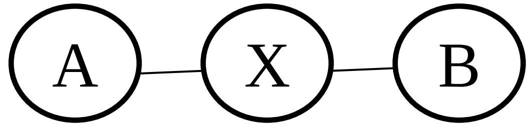


structure		size
bond	CH	1 Å
	CC	1.5 Å
protein radius		10 - 10^2 Å
α-helix spacing		5 ½ Å
C^{α}_{i} to C^{α}_{i+1}		3.8 Å

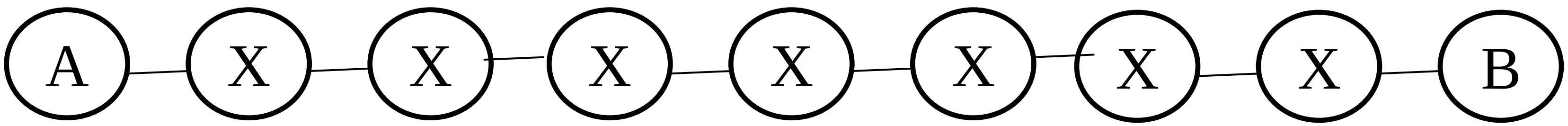
myoglobin picture 2w6w

proteins are polymers

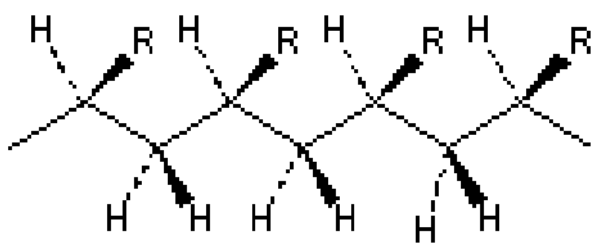
- simple polymers



many times gives



example

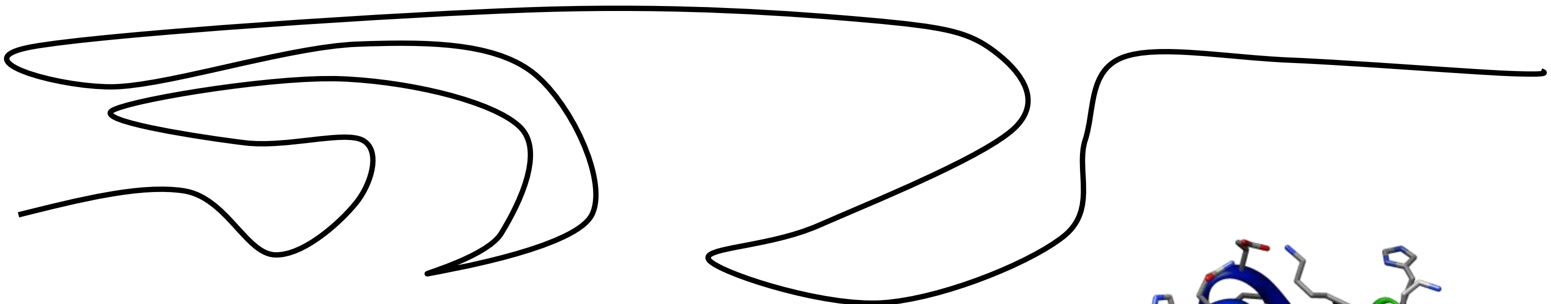


what kind of polymer would this give ?

Is it obvious what R is ?

Why are proteins interesting polymers ?

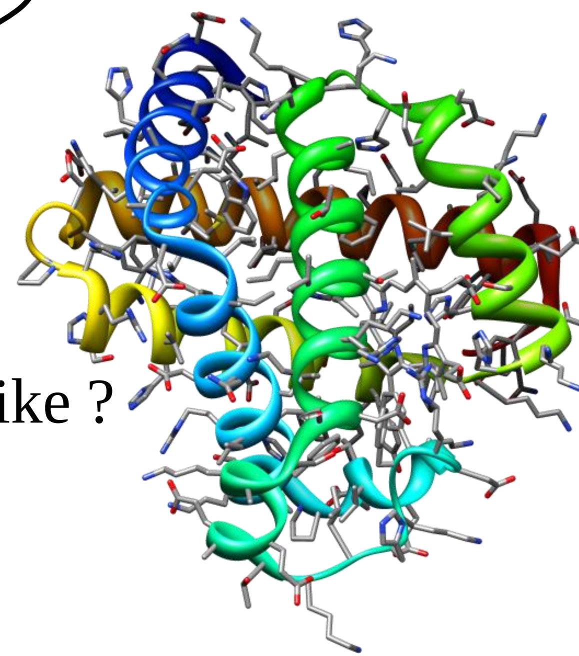
boring polymer gives uninteresting structures

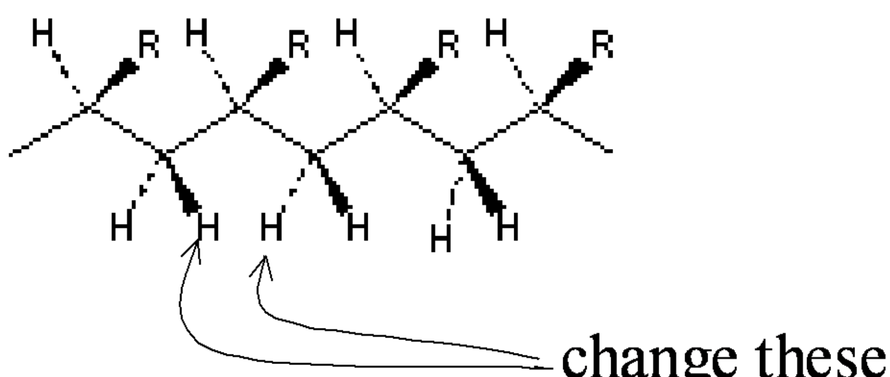


OK for plastic bags, haushaltsfolie.

Not nice regular structures..

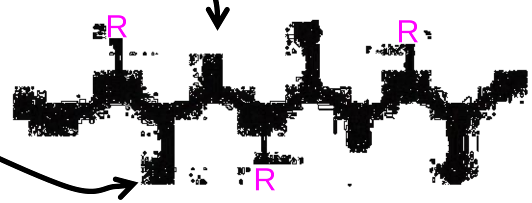
What can we do to make things more protein like ?



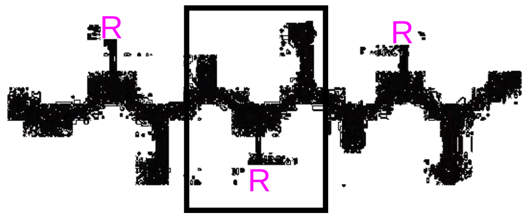


Giving proteins character 1

- more complicated backbone with H-bond
  - donor
  - acceptor



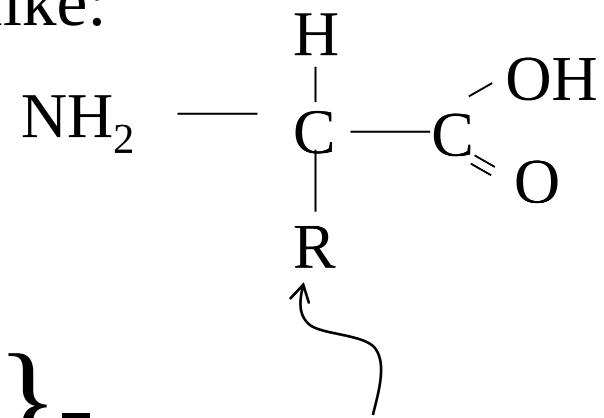
- basis of standard regular structures in proteins (secondary structure)
- repeating polymer unit:



- if this was all there was
  - all proteins would be the same

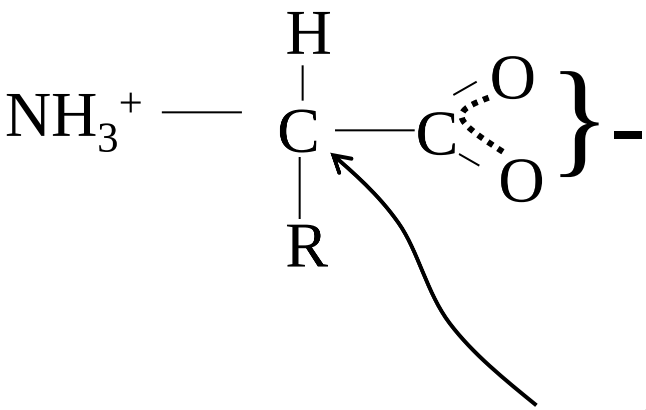
protein chemistry

amino acids (monomers) all look like:



sidechain

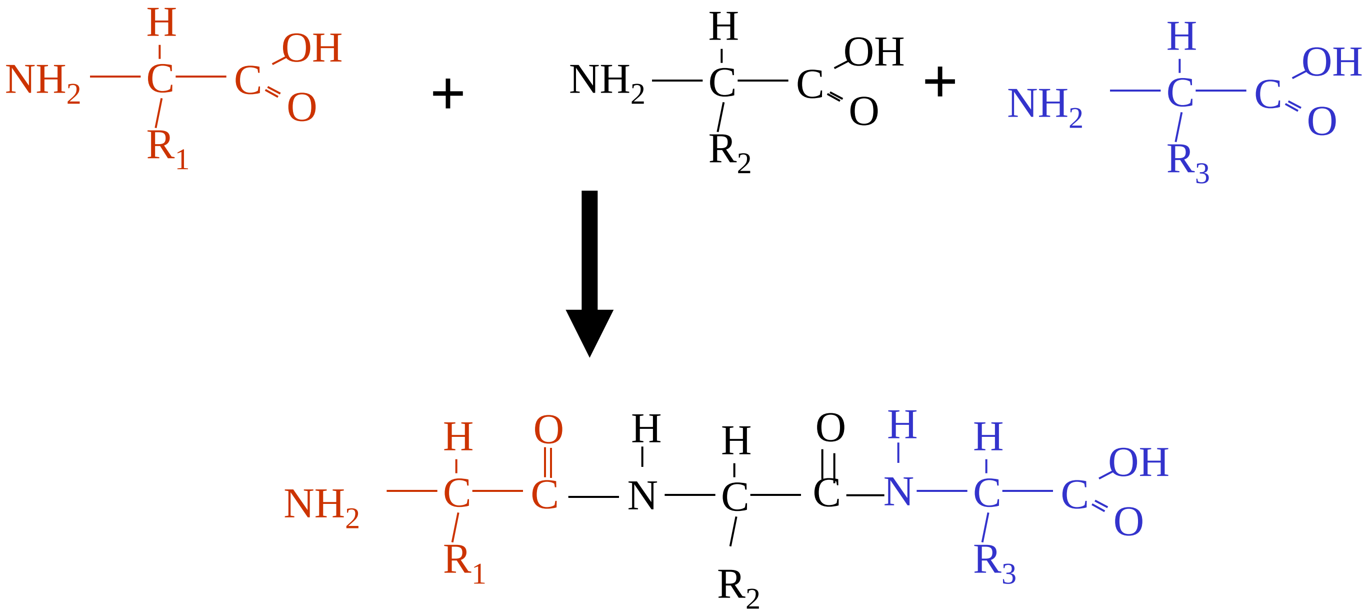
maybe



α carbon or C^{α}

- how can we construct specific structures ?
  - different kinds of "R" groups

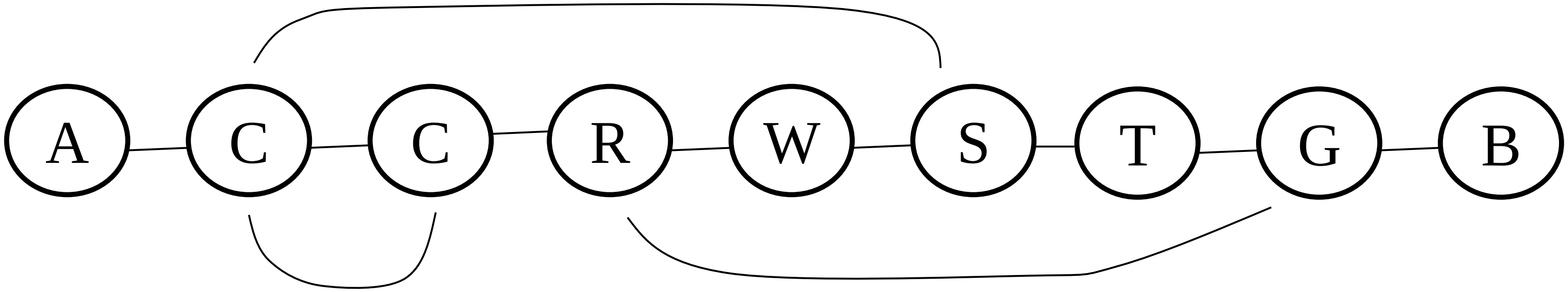
Putting monomers together



- protein synthesis story (biochemistry lectures)
- peptides and proteins
  - < 30 or 40 residues = peptide
  - > 30 or 40 residues = protein

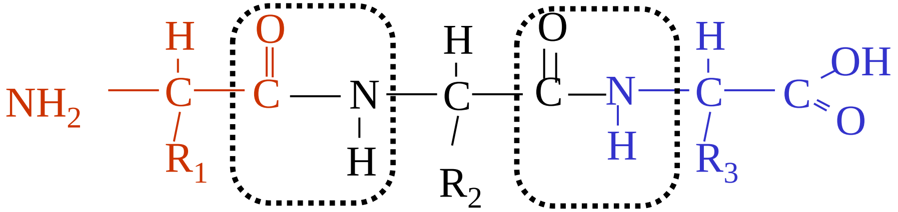
side chain possibilities

- big / small
- charged +, charged -, polar
- hydrophobic (not water soluble), polar
- interactions between sites…

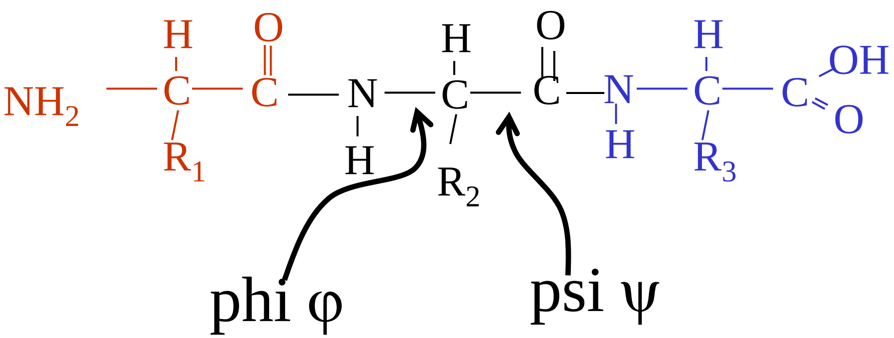


Backbone and consequences

- peptide bond is planar
  - partial double bond character (resonance forms)
  - shorter than other C-N
  - nearly always *trans*

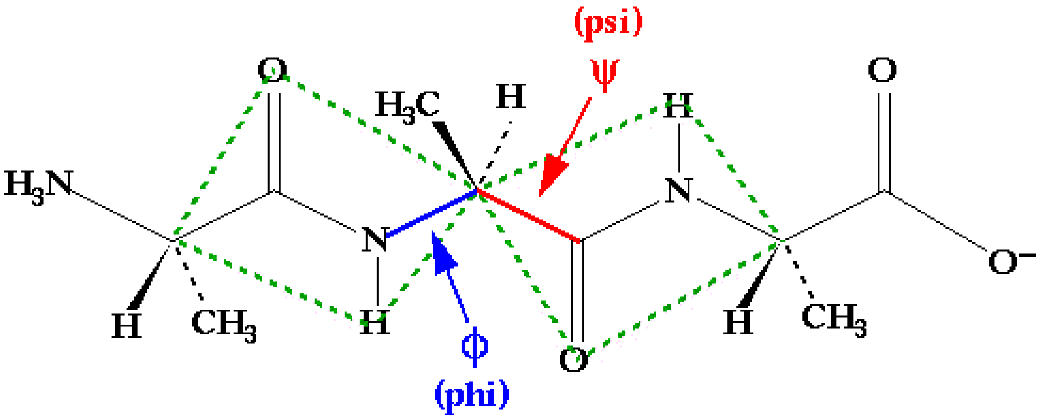


- two bonds can rotate



ramachandran plot

- can we rotate freely ?
  - no… steric hindrance



- Ramachandran plot

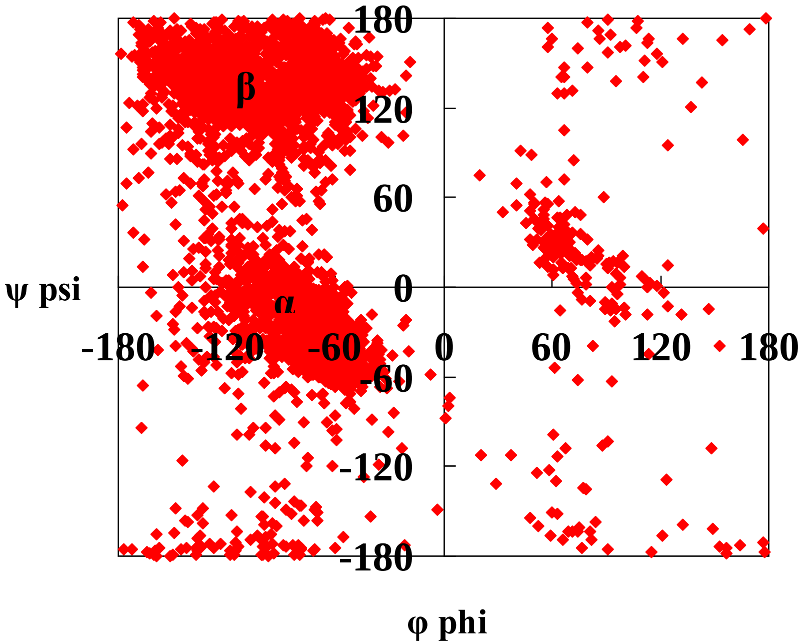
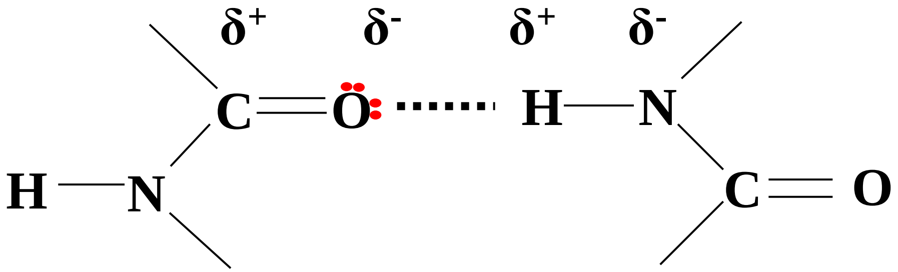


diagram from http://www.cgl.ucsf.edu/home/glasfeld/tutorial/AAA/AAA.html

Backbone H bonds

- oxygen is slightly negative
- NH bond is polar



- H-bonds
  - can be near or far in sequence
  - fairly stable at room temperature

Secondary structure

- regular structures using information so far
  - rotate phi, psi angles so as to
    - form H-bonds where possible
    - do not force side chains to hit each other (steric clash)
- two common structures
  - α-helix
  - β-strand / sheet

α helix

- each CO of residue *i* H-bonded to N of *i*+4
- 3.6 residues per turn
- 2 H-bonds per residue
- side chains well separated

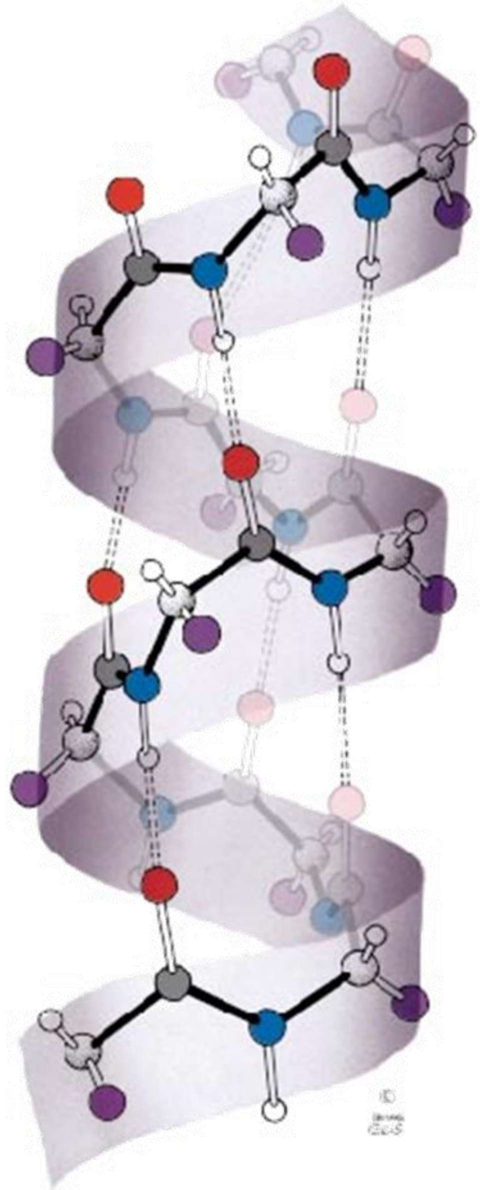


diagram from Voet, D.J. and Voet, J.G, Biochemistry, Wiley, 2004

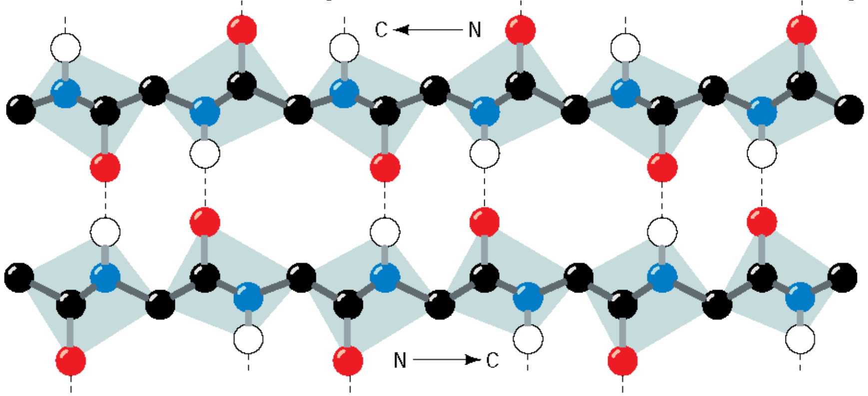
β-sheet

β-strand

- stretch out backbone and make NH and CO groups point out

β-sheet

- join these strands together with H-bonds (2 H-bonds/residue)
- anti-parallel



- or parallel

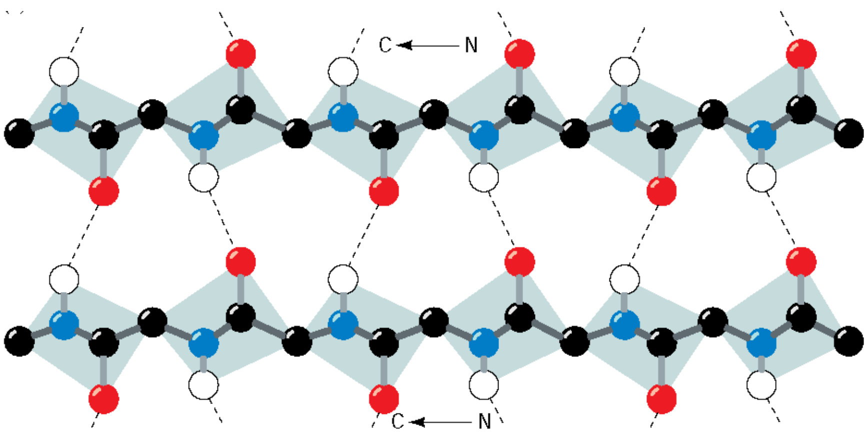
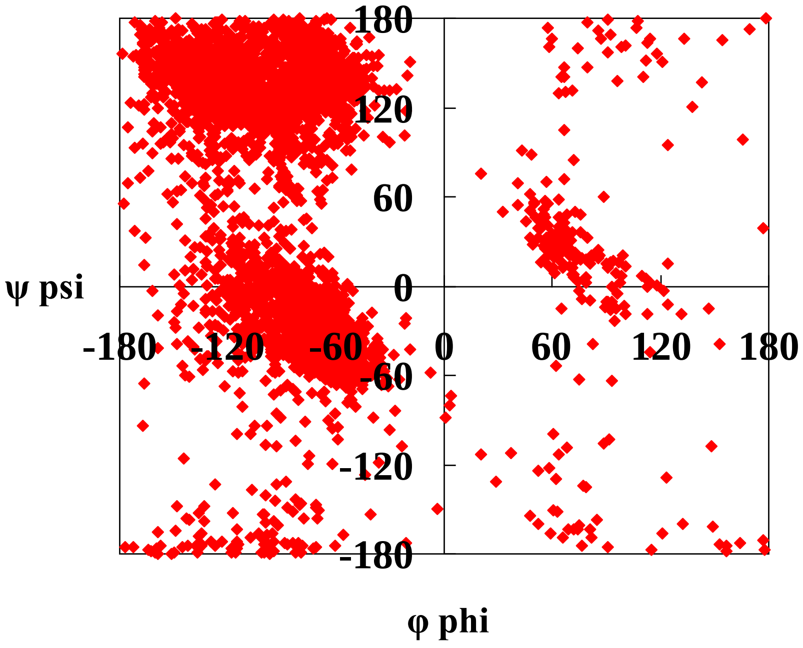


diagram from Voet, D.J. and Voet, J.G, Biochemistry, Wiley, 2004

After α-helix and β-sheet

- do helices and sheets explain everything ?
  - no
  - there is flexibility in the angles (look at plot)
    - geometry is not perfectly defined
  - there are local deviations and exceptions
  - other common structures
    - tighter helices
    - some turns
  - other structure
    - coil, random, not named



What determines secondary structure ?

So far

- secondary structure pattern of H-bonding

Almost all residues have H-bond acceptor and donor

  - all could form α-helix or β-sheet ? No

Difference ?

  - sequence of side-chains – overall folding

Why else are sidechains important

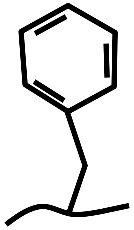
- chemistry of proteins (interactions, catalysis)

Fundamental dogma

- the sequence of sidechains determines the protein shape

Side chain properties

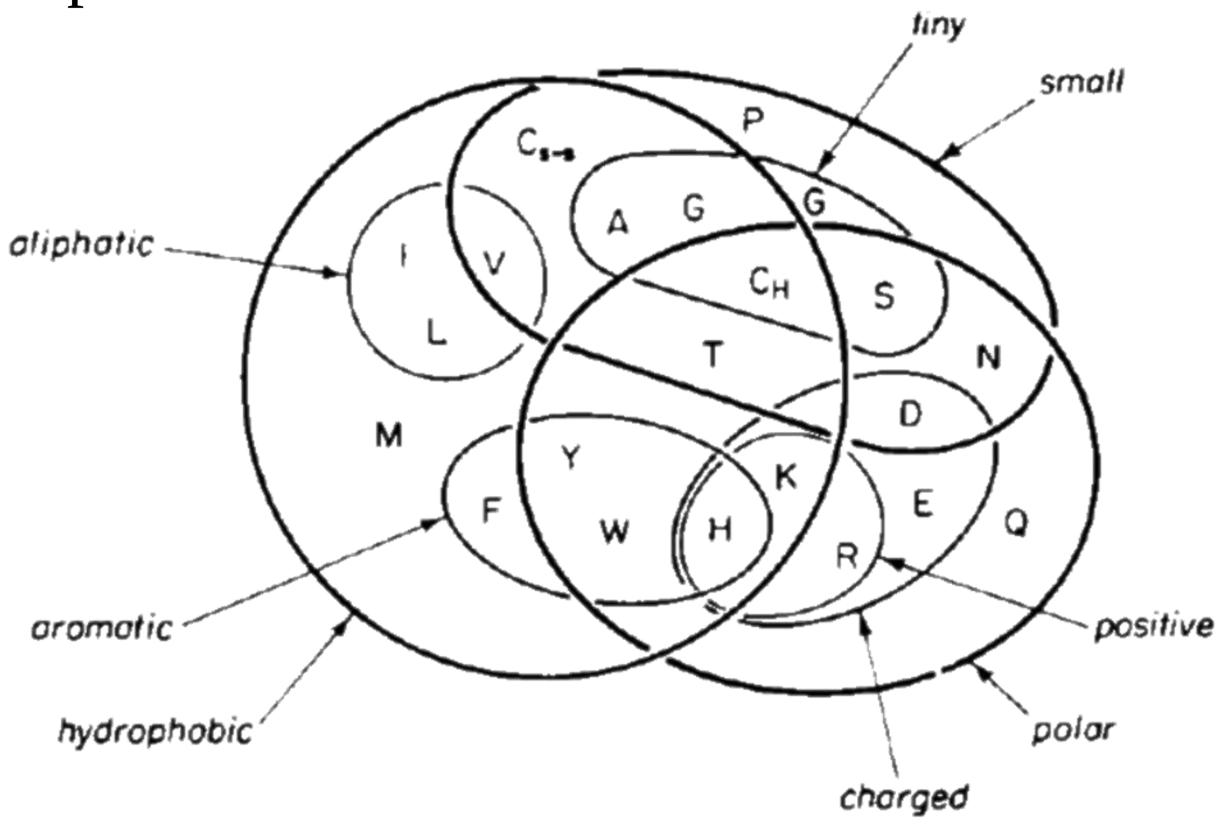
- properties
  - big / small
  - neutral / polar / charged
  - special (…)
- example
  - phenylalanine side chain looks like benzene (benzin)



    - very insoluble
    - benzene would rather interact with benzene than water
    - what if you have phe-phe-phe… poly-phe ?
      - does not happen in nature (can be made)
      - would be insoluble
      - not like a real peptide
    - phe is a constituent of real proteins – has a role

Properties are not clear cut

- You can be big / small, hydrophic / polar
  - combinations are possible



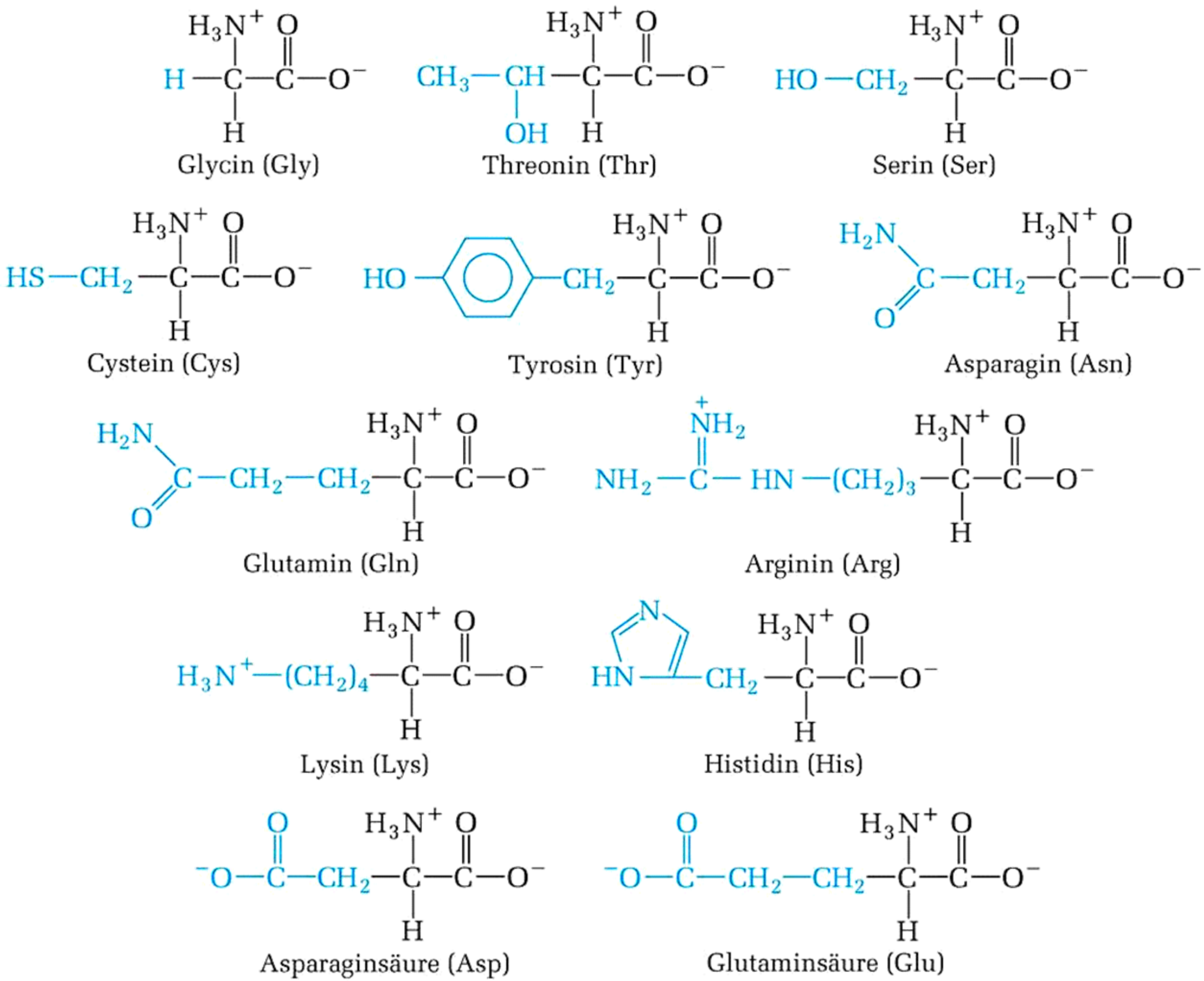
- Do not memorise this figure

Taylor, W.R. (1986) J. Theor. Biol., 215-218

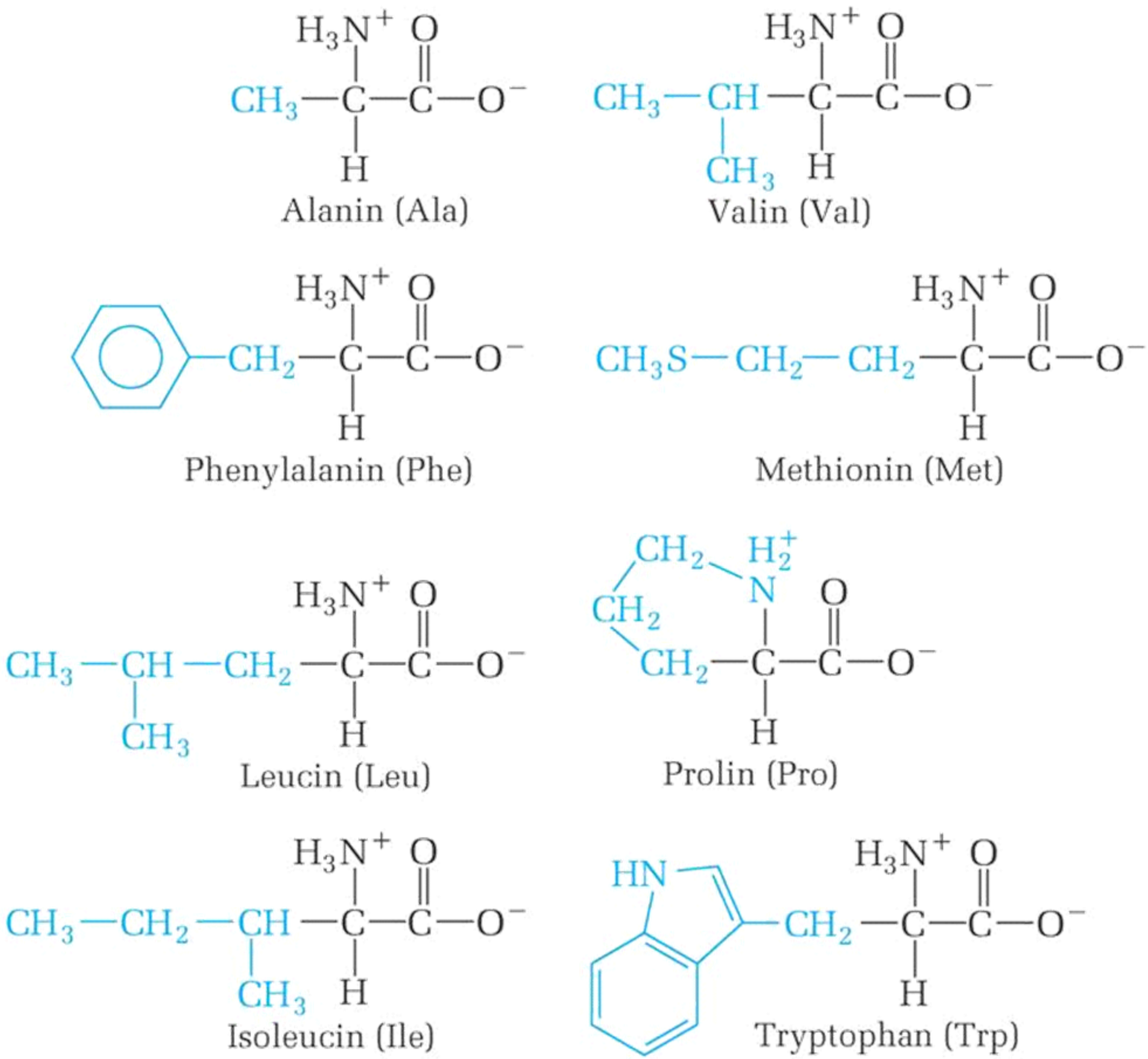
Sidechain interactions

- ionic (if the sidechains have charge)
- hydrophobic (insoluble sidechains)
- H-bonds (some donors and acceptors)
- repulsive

Summary of amino acids (first dozen)



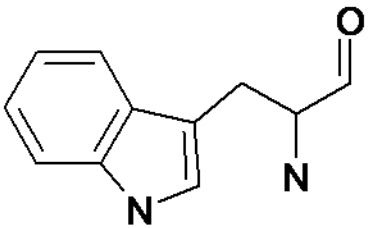
summary of amino acids (second lot)



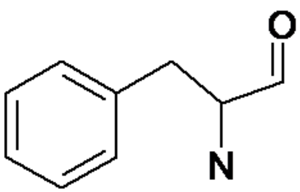
Amino Acids by property

aromatic

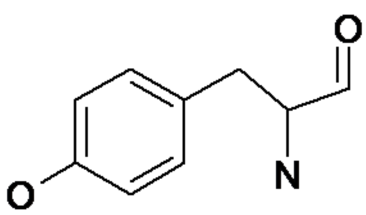
tryptophan



phenylalanine

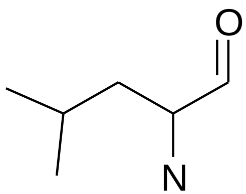


tyrosine

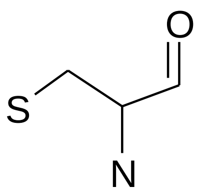


rather hydrophobic

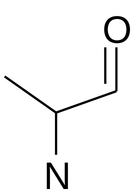
leucine



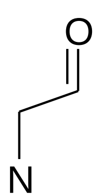
cysteine



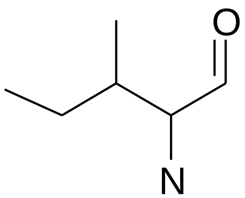
alanine



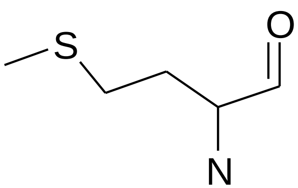
glycine



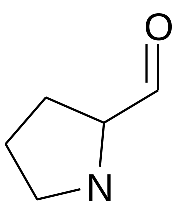
isoleucine



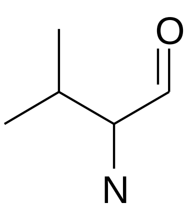
methionine



proline

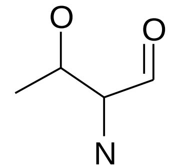


valine

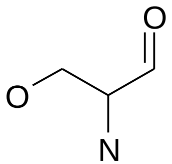


Polar

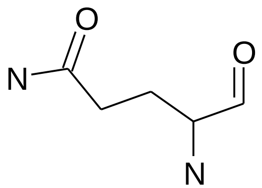
threonine



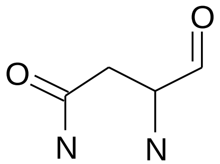
serine



glutamine

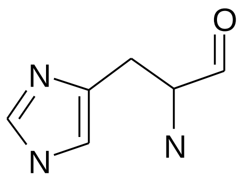


asparagine

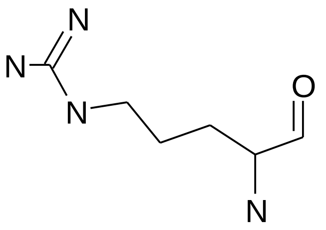


charged

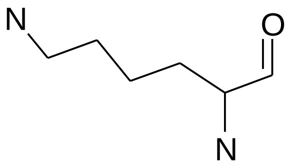
histidine



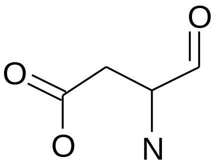
arginine



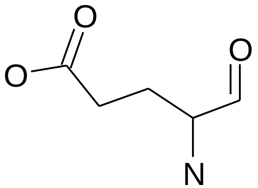
lysine



aspartate



glutamate



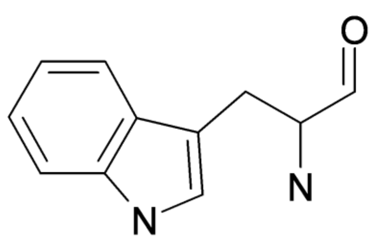
Hydrophobicity – how serious ?

- very serious, but simplified
  - the lists above are
    - pH dependent
    - difficult to measure experimentally (some aspects)
  - is hydrophobicity really defined ?

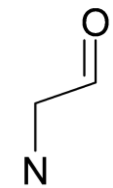
Other properties - size

big …   small →

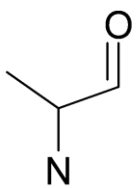
trp



gly

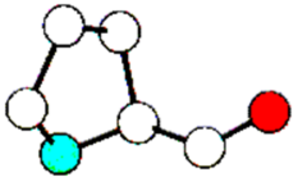


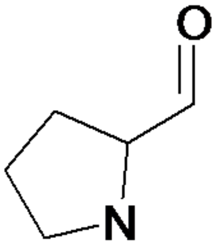
ala



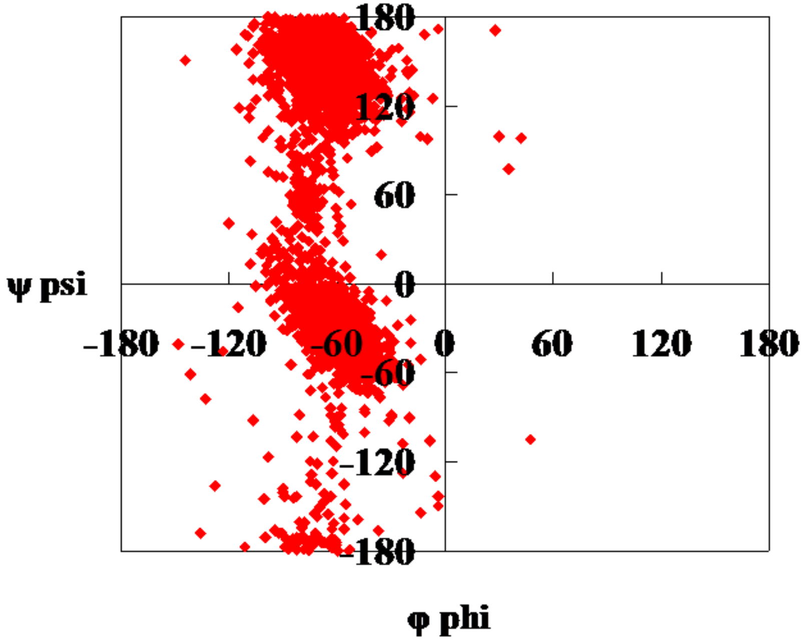
Other properties – chemistry / geometry

- proline
  - only one rotatable angle !
  - peptide bond sometimes *cis*



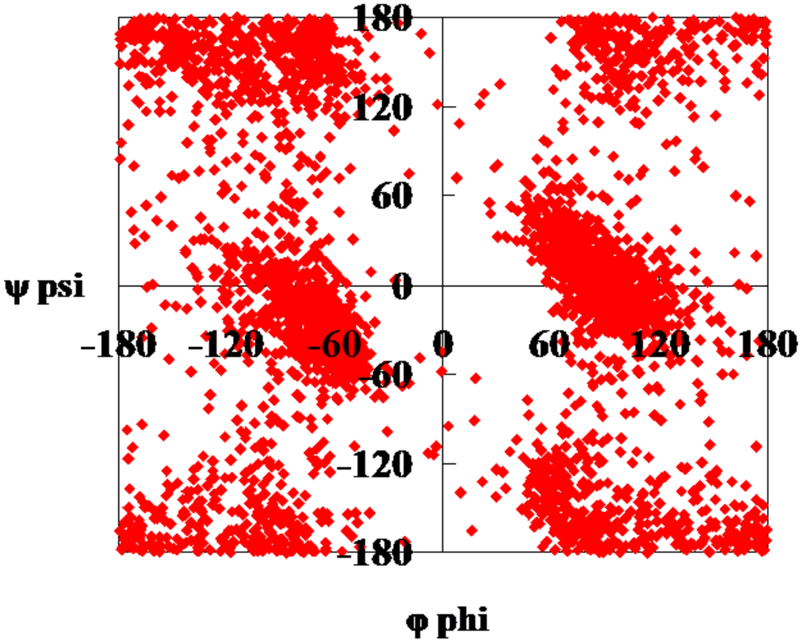


  - pro ramachandran plot

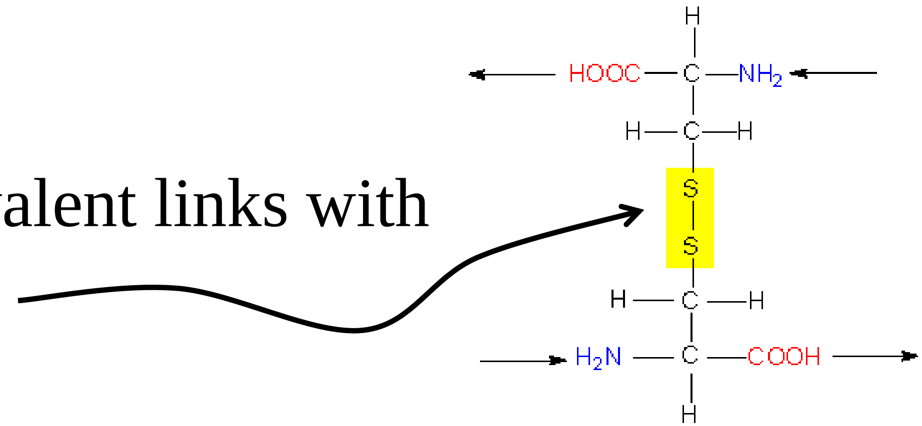


gly and cys

- glycine
  - no side chain
  - can visit forbidden parts of phi-psi map (4 000 points here)



- cysteine
  - forms covalent links with other cys



Summary so far

- proteins are heteropolymers
- backbone forms α-helices and β-strands (and more)
  - not sequence specific
- side-chains determine the
  - pattern of secondary structure
  - overall protein shape
- special amino acids
  - cys (forms disulfide bridges)
  - gly (can visit "forbidden" regions of ramachandran plot)
  - pro (no H-bond donor)
- how many sequences can one have ? $20^{n_{res}}$

Nomenclature

- some rules are unavoidable

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

- always write from N to C terminal
  - important convention

Definitions, primary, secondary …

More definitions

- primary structure
  - sequence of amino acids
    - ACDF (ala cys asp phe…)
- secondary structure
  - α-helix, β-sheet (+ few more)
    - structure defined by local backbone
- tertiary structure
  - how these units fold together
  - coordinates of a protein

Protein structure general comments

- primary, secondary, tertiary structure … how real ?
  - primary/secondary well defined
  - edges can blur
    - supersecondary struct / tertiary

Representation

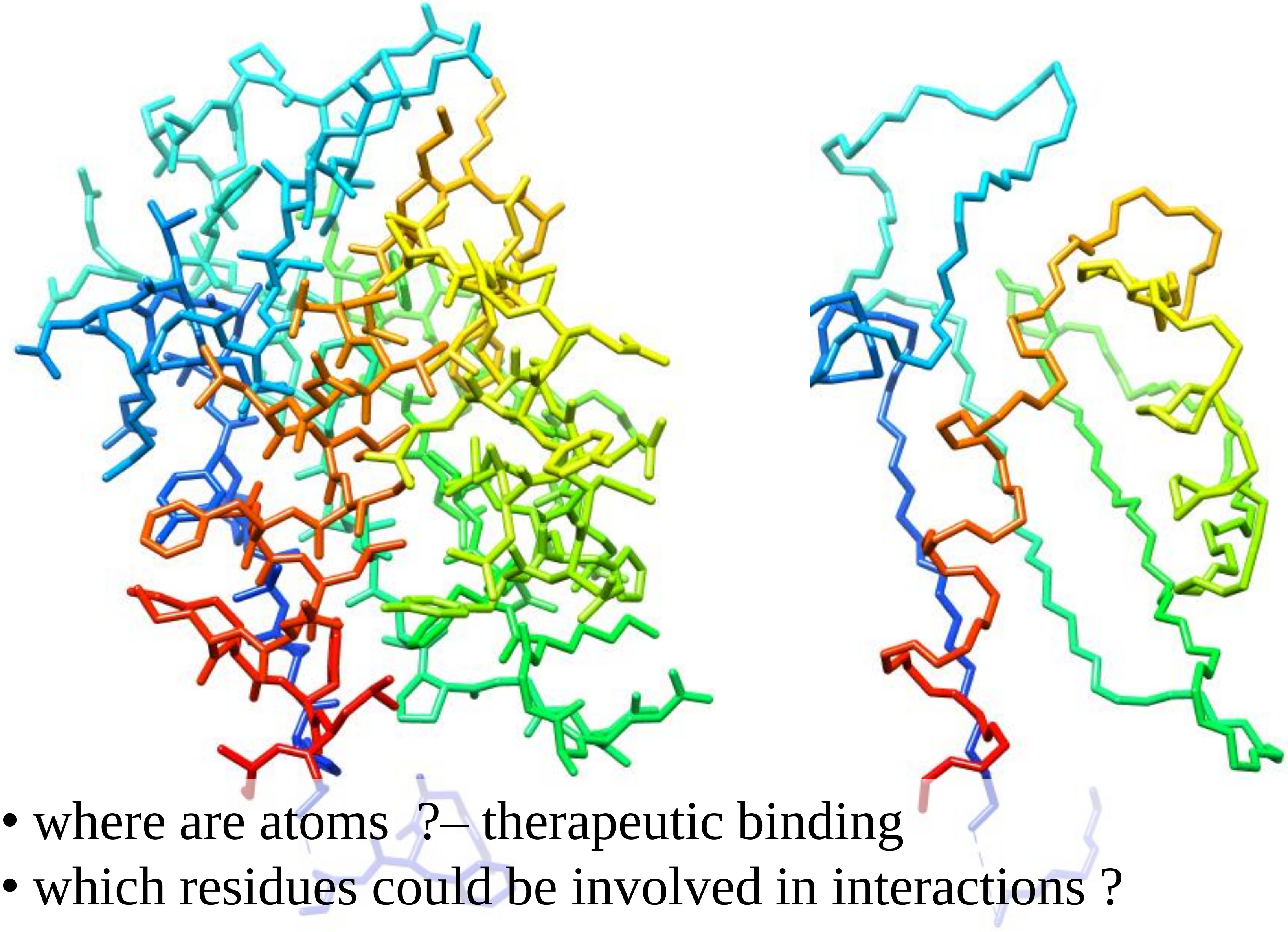
- Ultimately, our representation of a structure…

ATOM	1	N	ARG	1	31.758	13.358	-13.673	1.00	18.79	1BPI	137
ATOM	2	CA	ARG	1	31.718	13.292	-12.188	1.00	14.26	1BPI	138
ATOM	3	C	ARG	1	33.154	13.224	-11.664	1.00	18.25	1BPI	139
ATOM	4	O	ARG	1	33.996	12.441	-12.225	1.00	20.10	1BPI	140
ATOM	5	CB	ARG	1	30.886	12.103	-11.724	1.00	16.74	1BPI	141
ATOM	6	CG	ARG	1	29.594	11.968	-12.534	1.00	15.96	1BPI	142
ATOM	7	CD	ARG	1	28.700	13.182	-12.299	1.00	15.45	1BPI	143
ATOM	8	NE	ARG	1	27.267	12.895	-12.546	1.00	12.82	1BPI	144
ATOM	9	CZ	ARG	1	26.661	13.087	-13.727	1.00	17.38	1BPI	145
ATOM	10	NH1	ARG	1	27.370	13.558	-14.735	1.00	18.38	1BPI	146
ATOM	11	NH2	ARG	1	25.367	12.797	-13.838	1.00	25.73	1BPI	147
ATOM	12	N	PRO	2	33.800	13.936	-10.586	1.00	17.07	1BPI	148
ATOM	13	CA	PRO	2	34.976	13.367	-9.840	1.00	14.99	1BPI	149
ATOM	14	C	PRO	2	34.960	11.922	-9.660	1.00	13.11	1BPI	150
ATOM	15	O	PRO	2	33.962	11.306	-9.391	1.00	10.57	1BPI	151
ATOM	16	CB	PRO	2	34.922	14.145	-8.523	1.00	15.81	1BPI	152
ATOM	17	CG	PRO	2	34.058	15.391	-8.737	1.00	18.91	1BPI	153
ATOM	18	CD	PRO	2	33.371	15.273	-10.096	1.00	19.41	1BPI	154
ATOM	19	N	ASP	3	36.192	11.317	-9.707	1.00	8.73	1BPI	155

x, y, z coordinates

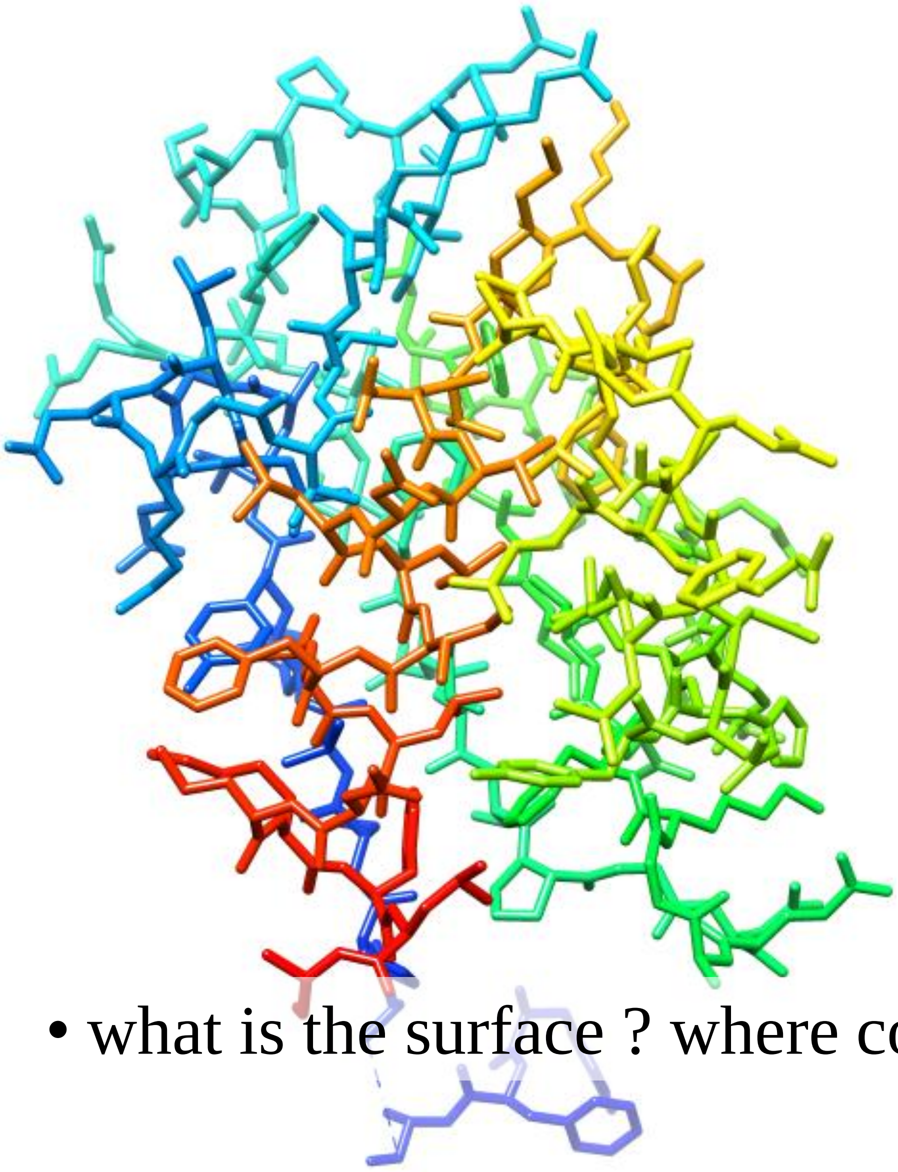
- drawing the structure ?

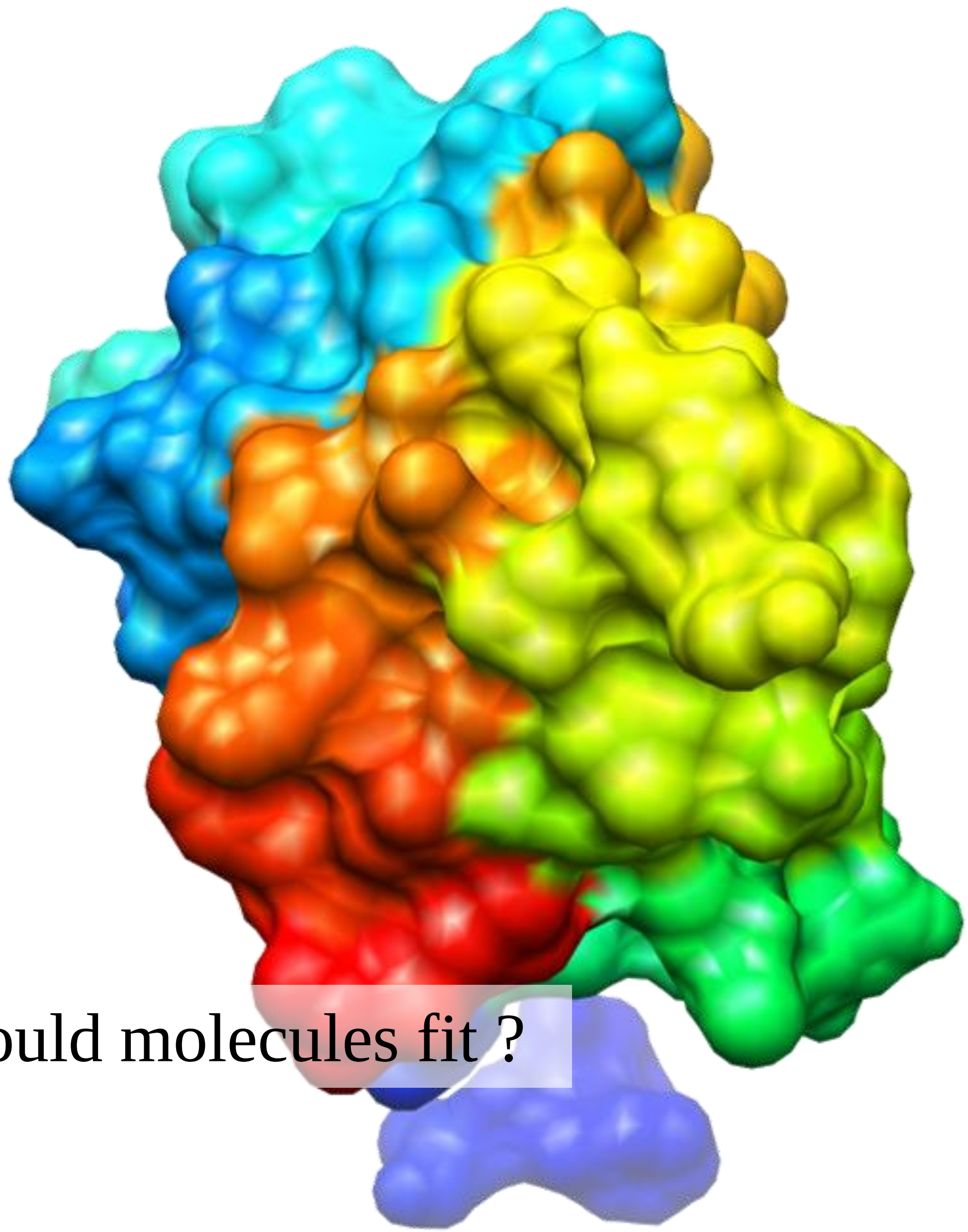
Representations



- where are atoms ?– therapeutic binding
- which residues could be involved in interactions ?

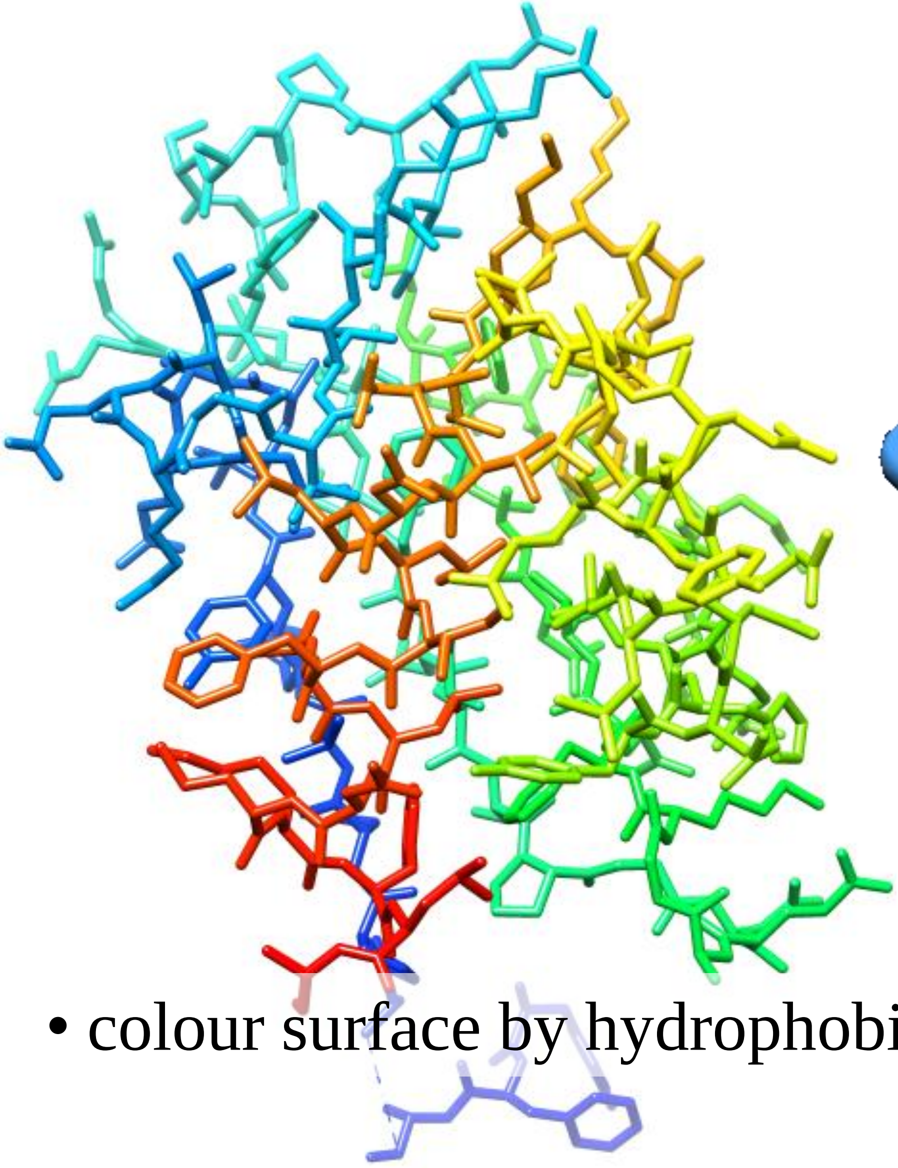
Representations

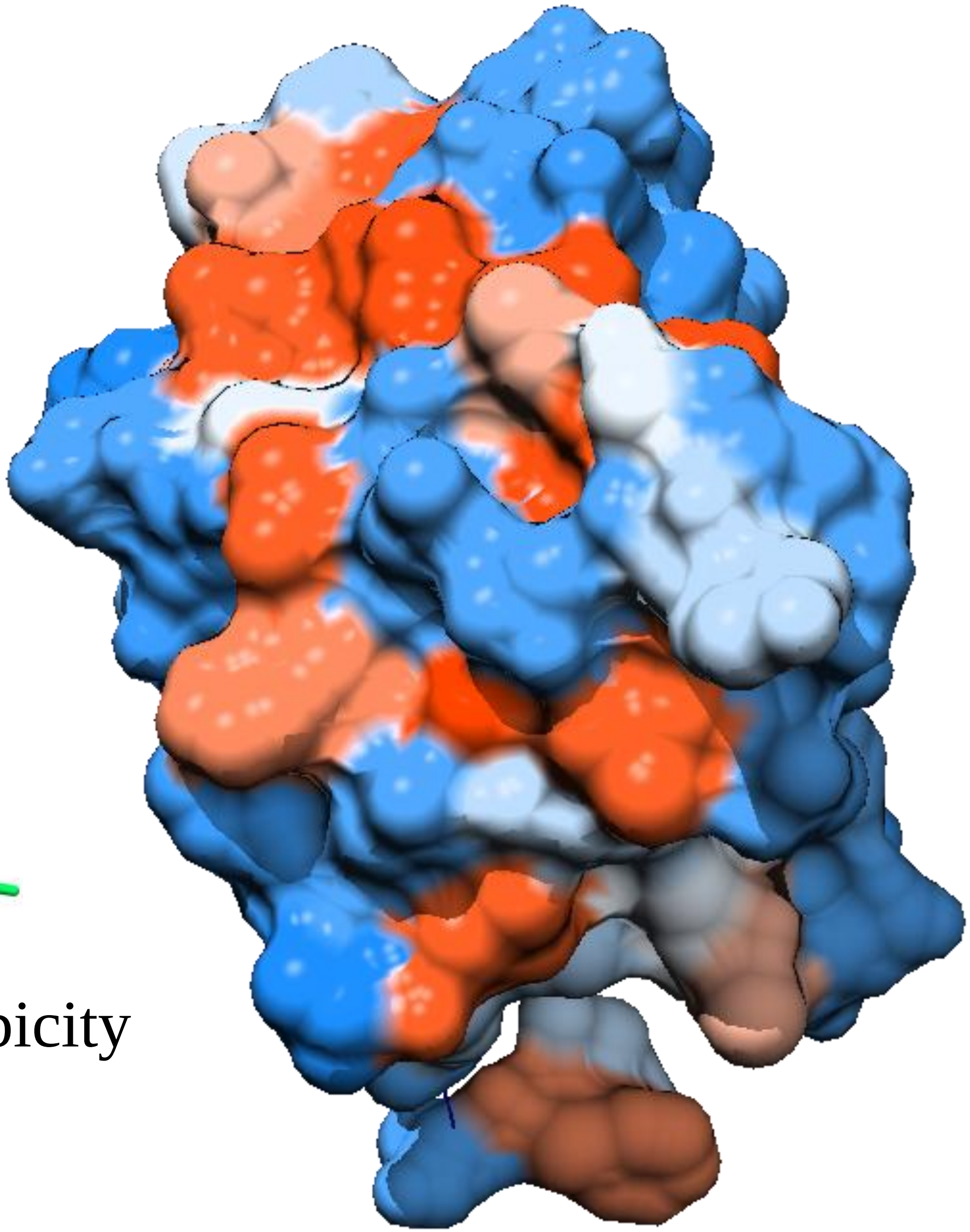




• what is the surface ? where could molecules fit ?

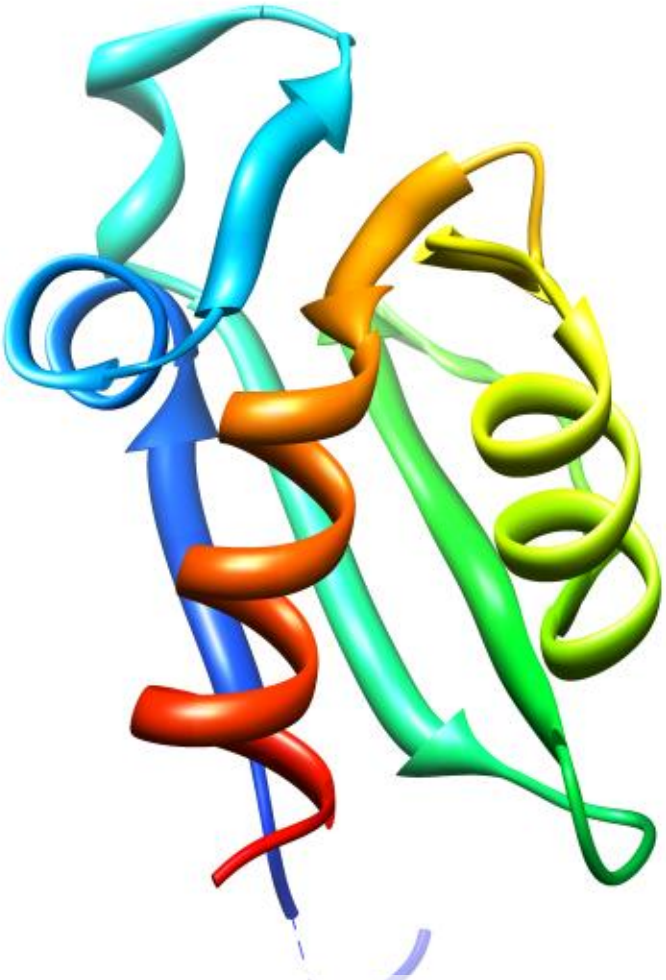
Representations

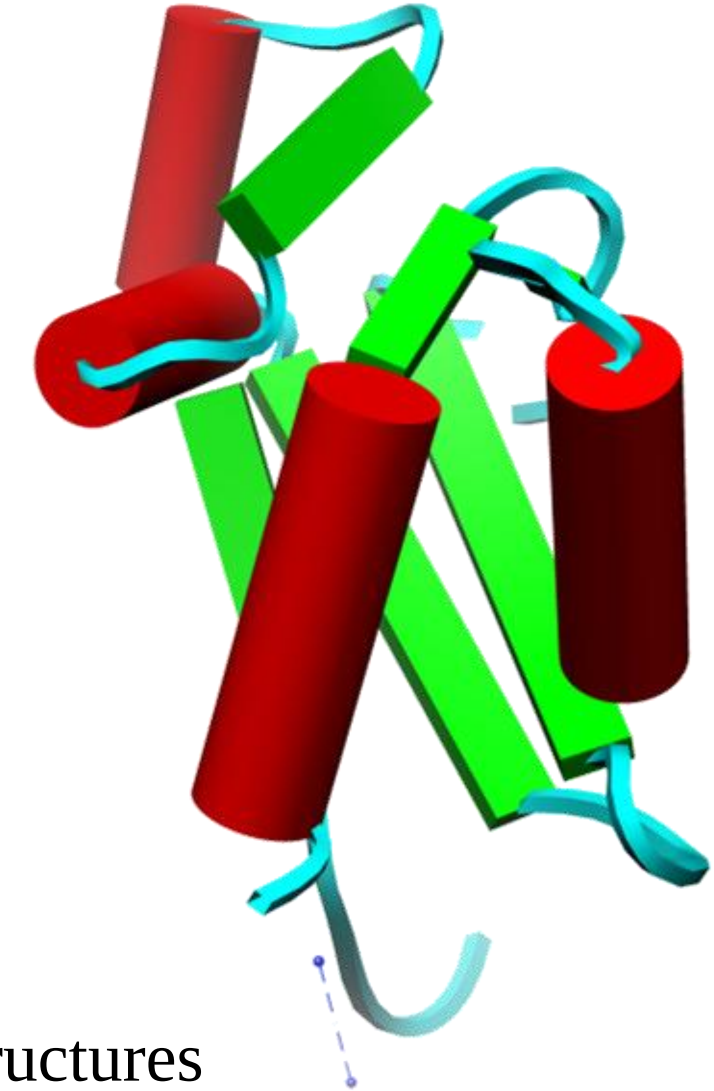




- colour surface by hydrophobicity

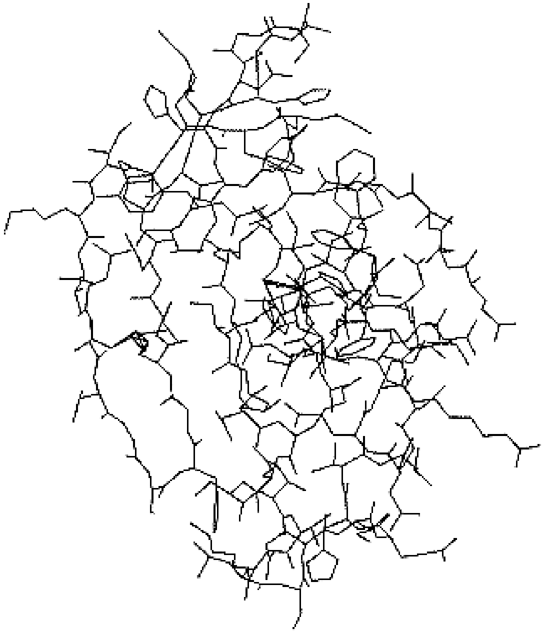
Representations

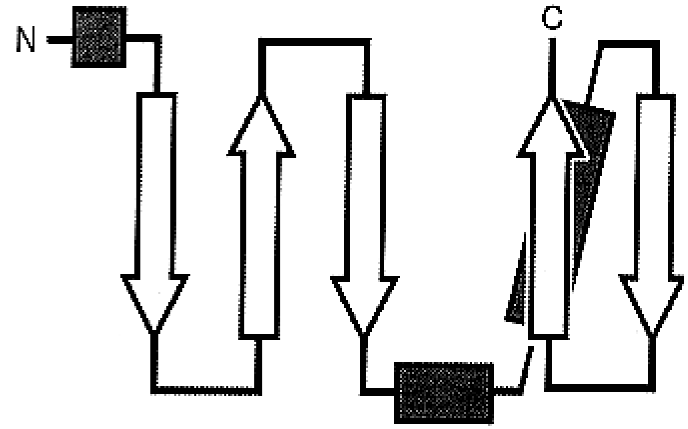


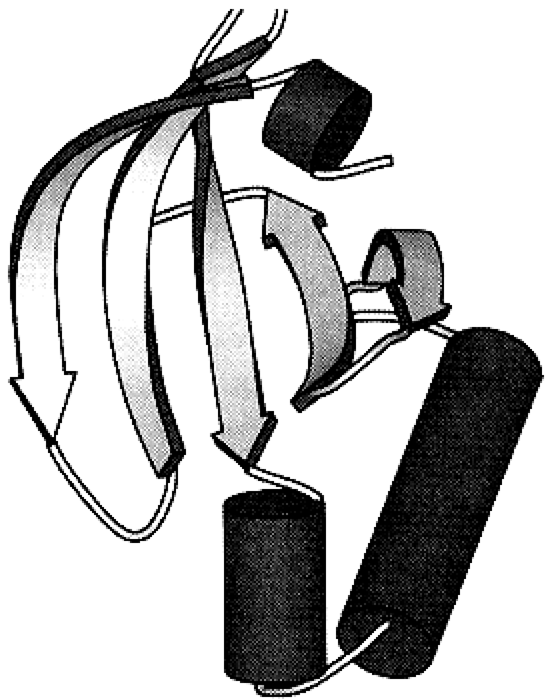


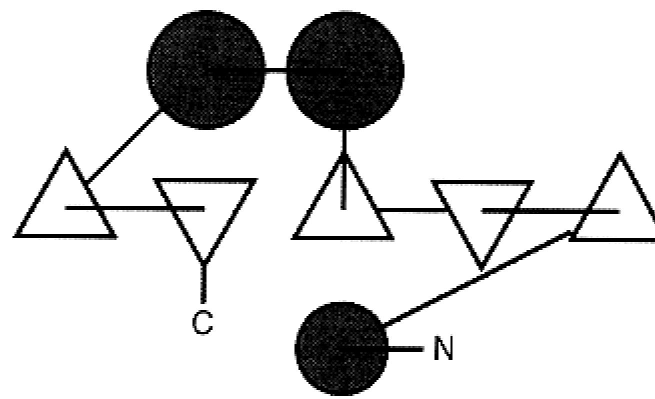
- highlight / emphasise regular structures

Different levels of abstraction







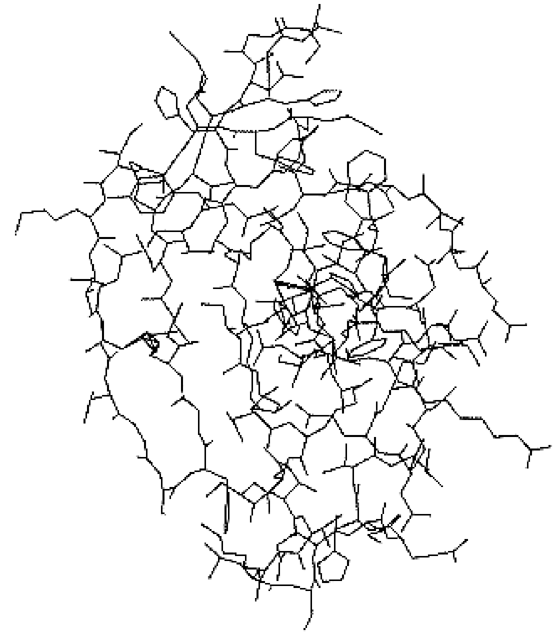


pictures from "Structural Bioinformatics", ed Bourne, PE and Weissig, H., Wiley New York (2003)

Atomistic

For details

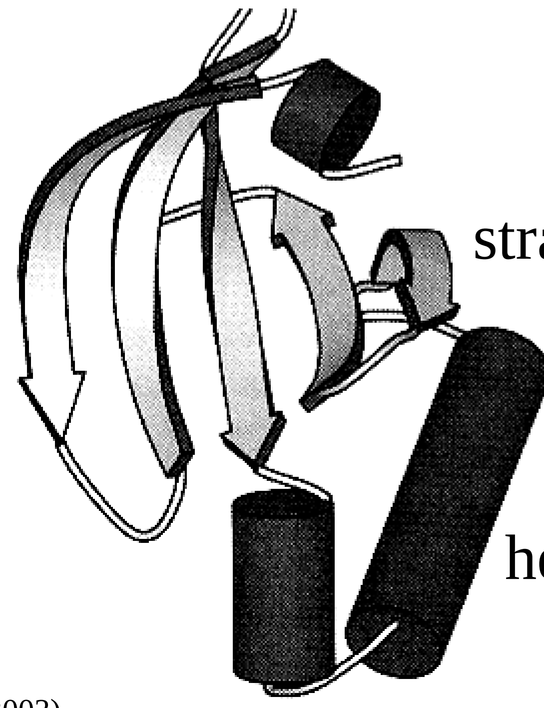
- where does a ligand bind ?
- which interactions is a residue involved in ?



Ribbons

Overview

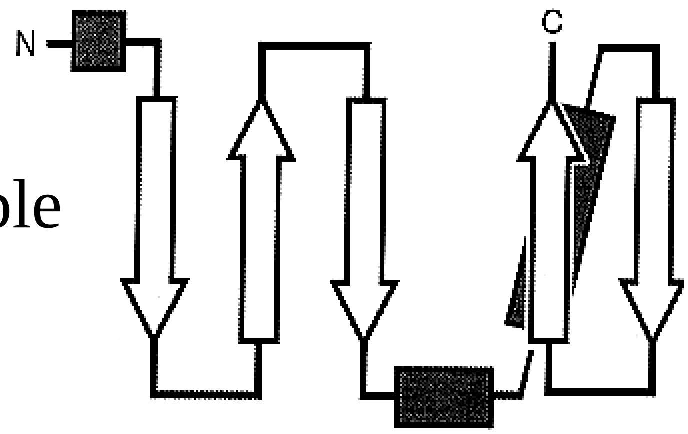
- shape
- number secondary struct elements
- symmetry

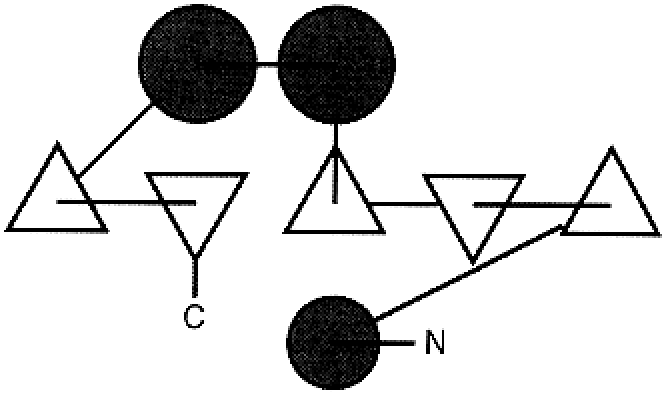


pictures from "Structural Bioinformatics", ed Bourne, PE and Weissig, H., Wiley New York (2003)

More abstract

- no idea of real shape
- very quickly classify a protein – example
  - lots of serine proteases
  - lots of different sequences
  - all very similar at this level of abstraction





pictures from "Structural Bioinformatics", ed Bourne, PE and Weissig, H., Wiley New York (2003)

Why does structure matter ?

- what residues can I change and preserve function ?
- what is the reaction mechanism of an enzyme ?
- what small molecules would bind and block the enzyme ?
- is this protein the same shape as some other of known function ?

Where do structures come from ?

- X-ray crystallography
- NMR
- + a bit of small angle X-ray scattering, electron diffraction, …

Atomic coordinates - warnings

- remember the coordinate file ?
- lots of problems
  - atoms and residues missing
  - numbering can be peculiar
- history
  - suits fortran 66 (think columns)
- non-standard amino acids
- nucleotides, ligands
- accuracy

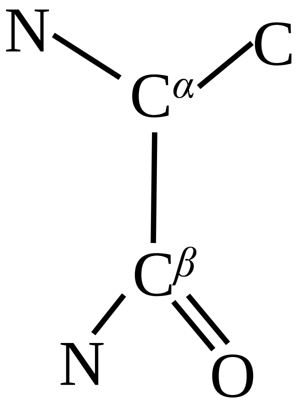
```
ATOM      1  N   ARG     1      31.758  13.358 -13.673  1.00 18.79      1BPI 137
ATOM      2  CA  ARG     1      31.718  13.292 -12.188  1.00 14.26      1BPI 138
ATOM      3  C   ARG     1      33.154  13.224 -11.664  1.00 18.25      1BPI 139
ATOM      4  O   ARG     1      33.996  12.441 -12.225  1.00 20.10      1BPI 140
ATOM      5  CB  ARG     1      30.886  12.103 -11.724  1.00 16.74      1BPI 141
ATOM      6  CG  ARG     1      29.594  11.968 -12.534  1.00 15.96      1BPI 142
ATOM      7  CD  ARG     1      28.700  13.182 -12.299  1.00 15.45      1BPI 143
ATOM      8  NE  ARG     1      27.267  12.895 -12.546  1.00 12.82      1BPI 144
ATOM      9  CZ  ARG     1      26.661  13.087 -13.727  1.00 17.38      1BPI 145
ATOM     10  NH1 ARG     1      27.370  13.558 -14.735  1.00 18.38      1BPI 146
ATOM     11  NH2 ARG     1      25.367  12.797 -13.838  1.00 25.73      1BPI 147
ATOM     12  N   PRO     2      33.800  13.936 -10.586  1.00 17.07      1BPI 148
ATOM     13  CA  PRO     2      34.976  13.367  -9.840  1.00 14.99      1BPI 149
ATOM     14  C   PRO     2      34.960  11.922  -9.660  1.00 13.11      1BPI 150
ATOM     15  O   PRO     2      33.962  11.306  -9.391  1.00 10.57      1BPI 151
ATOM     16  CB  PRO     2      34.922  14.145  -8.523  1.00 15.81      1BPI 152
ATOM     17  CG  PRO     2      34.058  15.391  -8.737  1.00 18.91      1BPI 153
ATOM     18  CD  PRO     2      33.371  15.273 -10.096  1.00 19.41      1BPI 154
ATOM     19  N   ASP     3      36.192  11.317  -9.707  1.00  8.73      1BPI 155
```

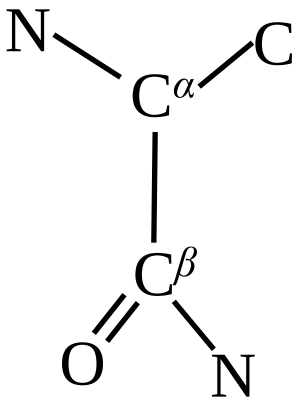
resolution, precision, accuracy

- coordinates 1.1 1.0 8.5
    - what do they mean ?
- random errors
    - non-systematic / noise / uncertainty
    - should be scattered around correct point
- from any measurement there are errors **±** *x*
- x-ray crystallography has model for data
    - uncertainty (probability)
    - resolution (experimental)
        - < 1 Å (good)
        - > 5 Å (bad, but excusable – monster structures)

X-ray crystallography

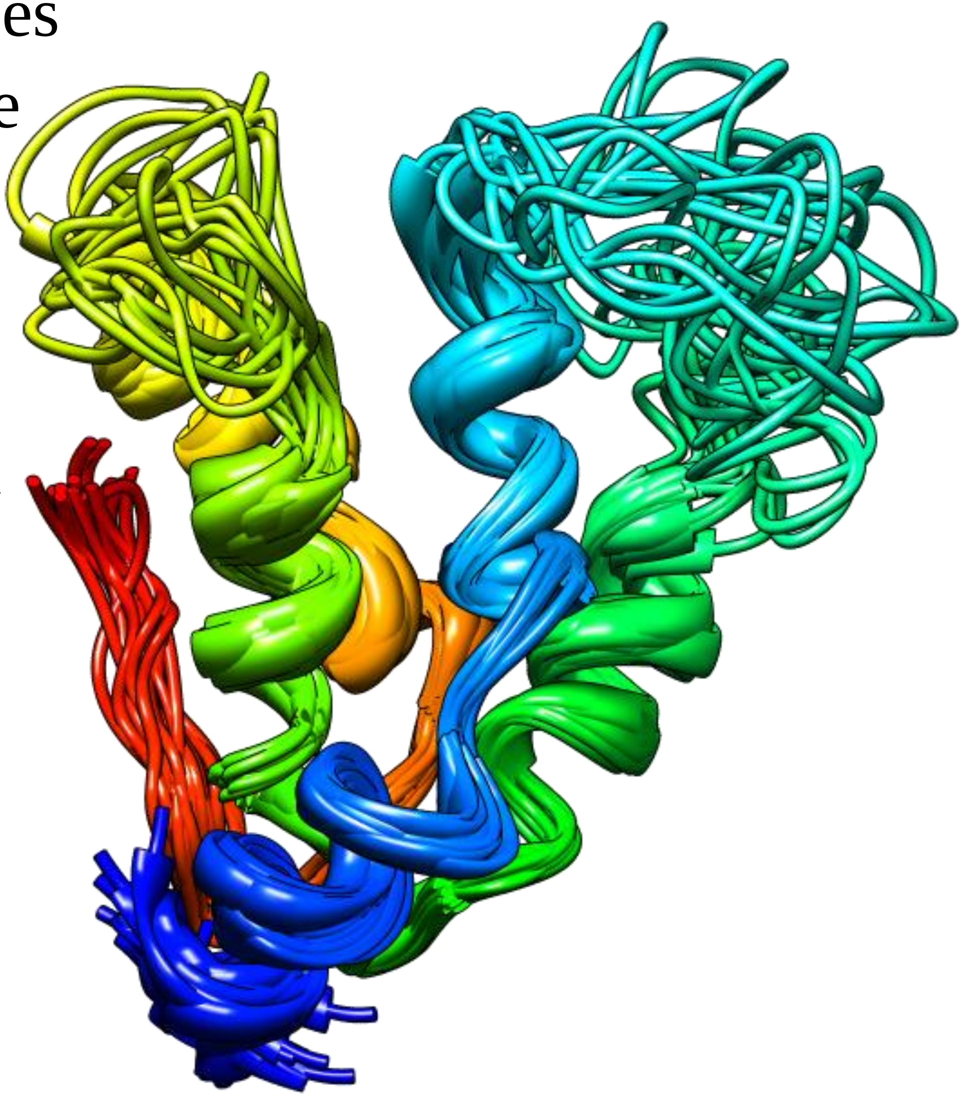
- non-systematic errors
  - small problems: (O and N look the same)
  - few huge problems
  - newer structures are better
- proteins are not static
  - overall motion
  - local motion





NMR structures

- different philosophy to X-ray
  - lots of little internal distances
  - do not quite define structure
- generate 50 or 10^2 solutions
  - look at scatter of solutions
- as with X-ray
  - some parts are well defined
  - some not



structure 1sm7

Summarise and stop

- roles of proteins
- heteropolymers – 20 types of amino acid / residue
- geometry – avoiding atomic clashes, forming H bonds
  - leads to regular secondary structure
- chemistry of amino acids very different to another
- unique structure for a sequence reflects these differences
- representations of structures
- structures in PDB are experimental – have errors