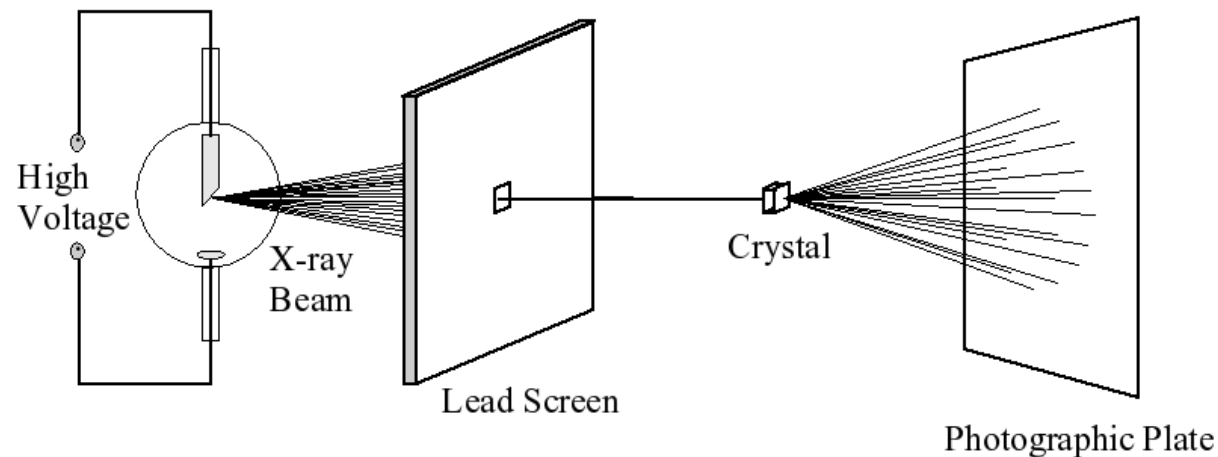


V10 – Protein structures

Program for today:

- Structures from protein X-ray crystallography
- Statistics of protein structures
- Statistical potentials

X-ray crystallography



X-rays are electromagnetic waves in the ultra short (“hard”) regime with wavelengths on the order of 0.1 nm.

When X-rays hit a sample, they interact weakly with the electron clouds around the atomic nuclei.

This leads to partial **diffraction** of the incoming beam into different angles.

As the interaction is quite weak, a noticeable diffraction intensity can only be detected in orientations where the diffracted beams from many molecules sum up in a **constructive way**.

X-ray crystallography

Electromagnetic waves are sinusoidal waves that may be described by an amplitude and phase.

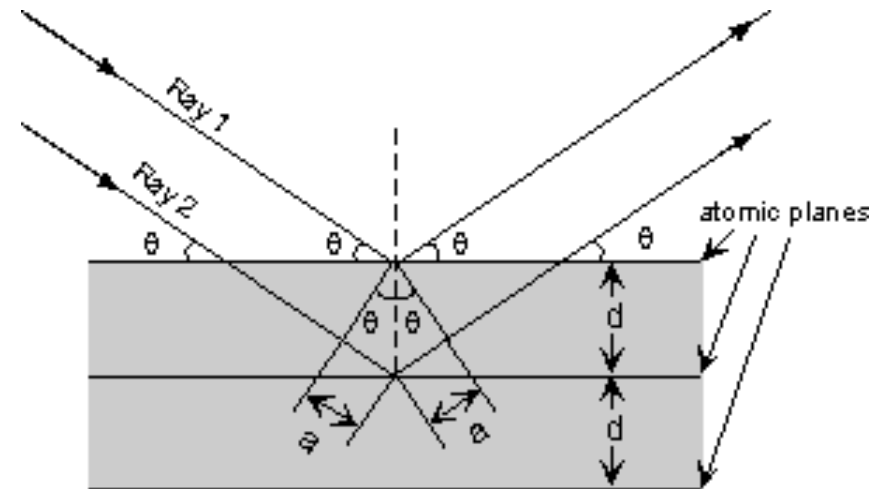
Intensities are only detected in those orientations where the path difference of waves ($2 \times a$) originating from different molecules equals integer multiples of their wavelength:

$$n\lambda = 2a$$

$$n\lambda = 2d \sin \theta$$

This is known as **Bragg's Law** for X-ray diffraction.

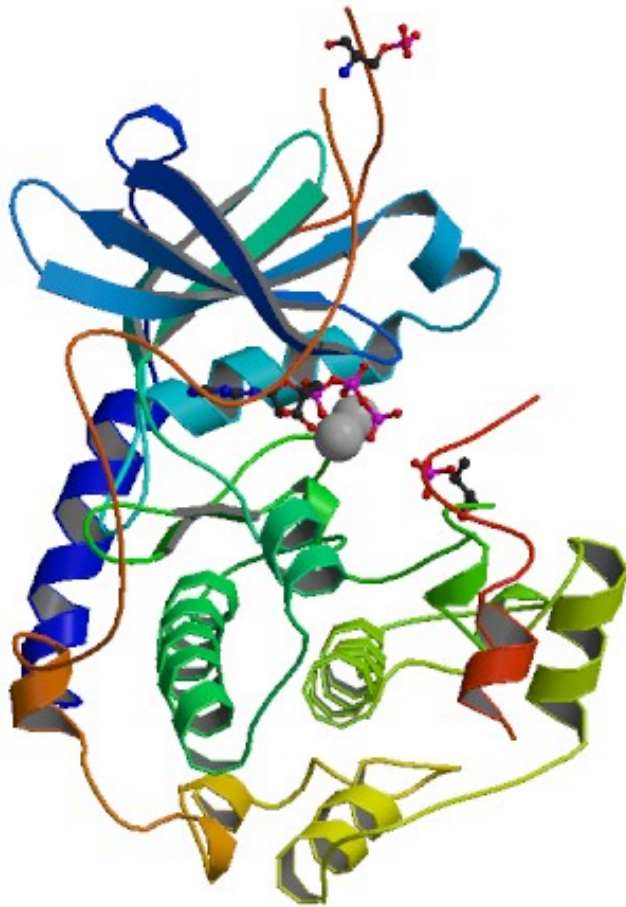
This condition requires a very ordered orientation of all molecules like in a 3D crystal. Still, in almost all orientations, the overlap of various waves will not be constructive.



Images on the photographic plate (or charge coupled display detector) are recorded for various rotational orientations of the crystal.

Structure determination involves reconstruction of the molecular structure of the target molecule that will give rise to the observed reflections.

PDB files



X-ray structure 1atp of
the cAMP-dependent
protein kinase

www.rcsb.org

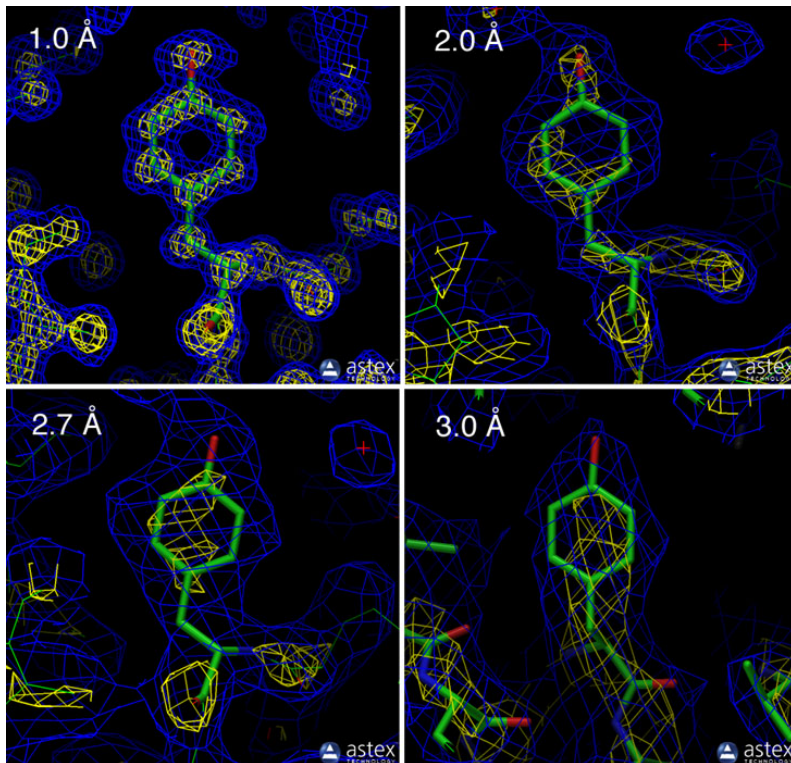
	Atom-number	Atom type	Residue type	Chain ID	Residue number	X-coordinate	Y-coordinate	Z-coordinate	Occupancy	B-factor
ATOM	1	N	VAL	E	15	-6.512	-12.177	-13.595	1.00	64.39
ATOM	2	CA	VAL	E	15	-5.276	-11.431	-13.476	1.00	47.83
ATOM	3	C	VAL	E	15	-4.815	-10.815	-14.785	1.00	35.56
ATOM	4	O	VAL	E	15	-4.806	-9.592	-14.904	1.00	99.02
ATOM	5	CB	VAL	E	15	-4.193	-12.092	-12.629	1.00	100.00
ATOM	6	CG1	VAL	E	15	-2.823	-11.529	-12.987	1.00	50.97
ATOM	7	CG2	VAL	E	15	-4.494	-11.830	-11.149	1.00	35.72
ATOM	8	N	LYS	E	16	-4.475	-11.641	-15.778	1.00	35.94
ATOM	9	CA	LYS	E	16	-4.060	-11.108	-17.074	1.00	55.13
ATOM	10	C	LYS	E	16	-5.100	-10.105	-17.531	1.00	59.23
ATOM	11	O	LYS	E	16	-4.877	-9.036	-18.103	1.00	35.80
ATOM	12	CB	LYS	E	16	-3.916	-12.209	-18.110	1.00	47.57
ATOM	13	CG	LYS	E	16	-2.850	-11.886	-19.158	1.00	100.00
ATOM	14	CD	LYS	E	16	-1.491	-12.525	-18.888	1.00	94.01
ATOM	15	CE	LYS	E	16	-0.665	-11.794	-17.836	1.00	100.00
ATOM	16	NZ	LYS	E	16	-0.505	-12.557	-16.586	1.00	89.11

In high-resolution X-ray structures, one can sometimes resolve different side chain orientations („occupancies“)

Resolution

Resolution : measure of the quality of the data that has been collected on the crystal containing the protein or nucleic acid.

It is defined as the **minimum plane spacing** d given by Bragg's law for a particular set of X-ray diffraction intensities.



Electron density maps for structures with different resolutions.

The first 3 show Tyr103 from myoglobin:

1a6m (1.0 Å resolution),

106m (2.0 Å resolution), and

108m (2.7 Å resolution).

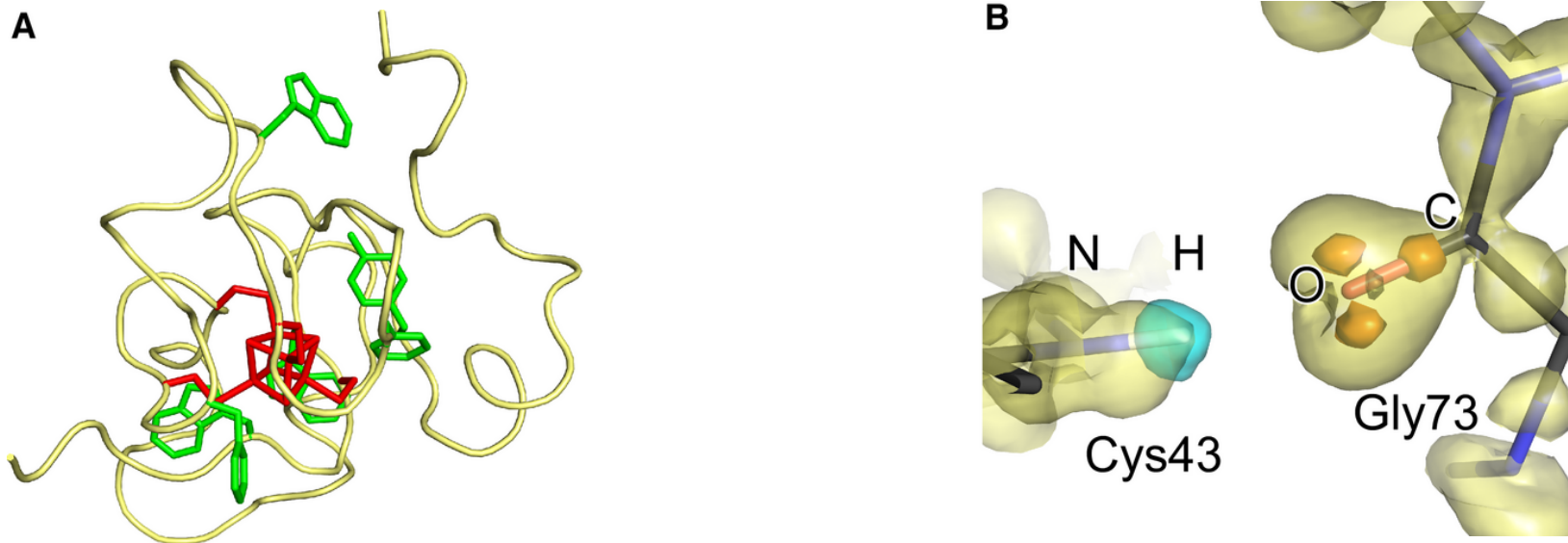
Bottom right: Tyr130 from hemoglobin,

1s0h (3.0 Å resolution).

Blue and yellow contours surround regions of high electron density.

The atomic model is shown with sticks.

Ultra high resolution structure (0.48 Å) of HiPIP



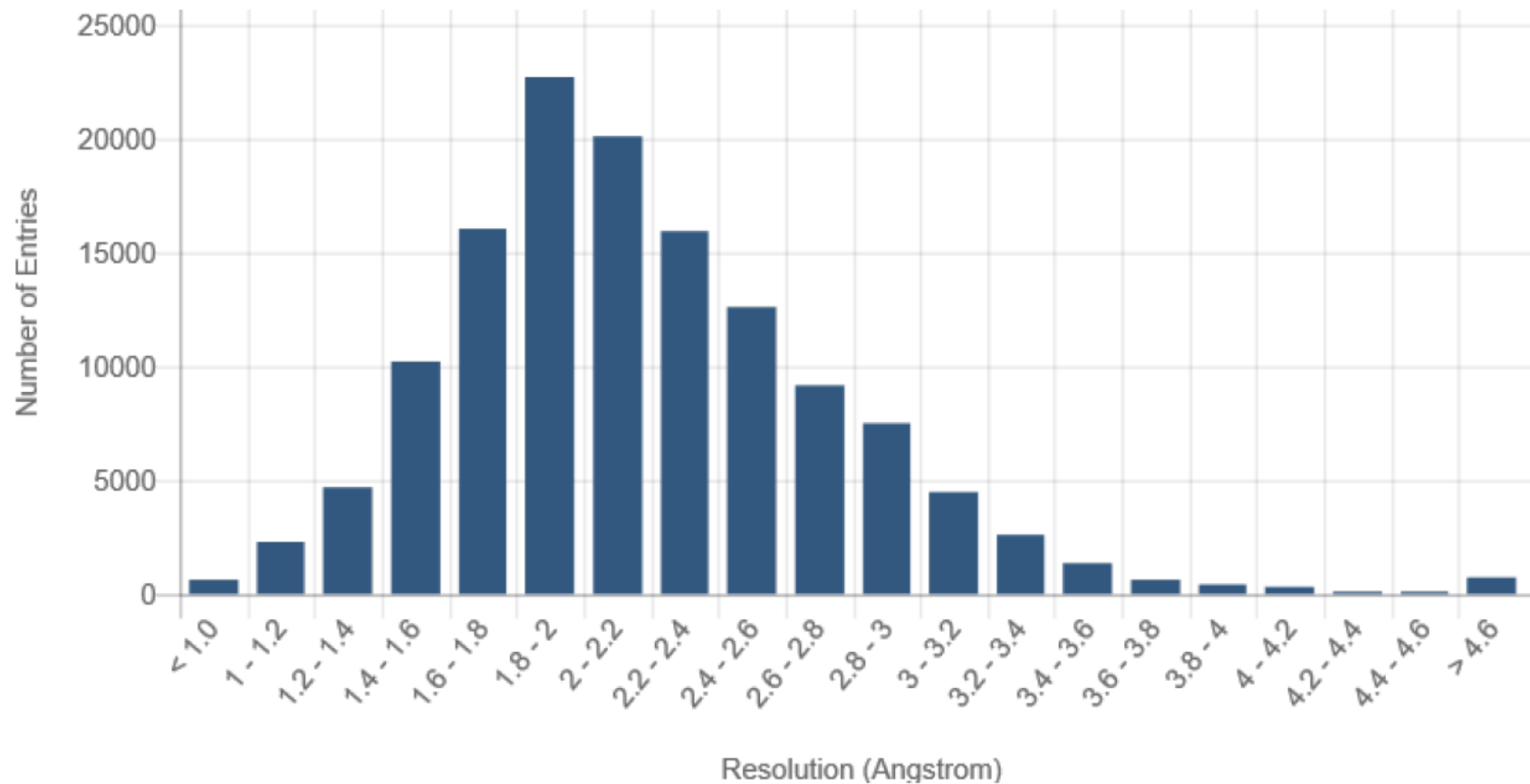
$R_{\text{free}} = 0.078$! At this resolution, enormous levels of detail can be detected.

(Left) The overall structure of HiPIP is shown as a tube model, where aromatic residues and the iron–sulfur cluster are represented as green and red sticks.

(Right) Hydrogen bonding formed between lone pair electrons of the carbonyl O of Gly73 and the amide H atoms of Cys43.

Takeda, Miki, FEBS
J. (2017)

Statistics of PDB



The resolution of protein structures has steadily improved in recent years.

R-value

R-value is the measure of the quality of the atomic model obtained from the crystallographic data.

$$R = \frac{\sum |F_{obs} - F_{calc}|}{\sum |F_{obs}|}$$

When solving the structure of a protein, the researcher first builds an atomic model and then back-calculates a simulated diffraction pattern based on that model.

The R-value measures how well the simulated diffraction pattern matches the experimentally-observed diffraction pattern.

F stands for the „structure factor“ describing the amplitude and phase of a wave diffracted from crystal lattice planes.

A totally random set of atoms will give an R-value of about 0.63, whereas a perfect fit would have a value of 0.

Typical values of “well refined” protein structures are about 0.20.

B-factor

The "temperature-factor" or "Debye-Waller factor" describes the degree to which the **electron density** of an atom **is spread out**.

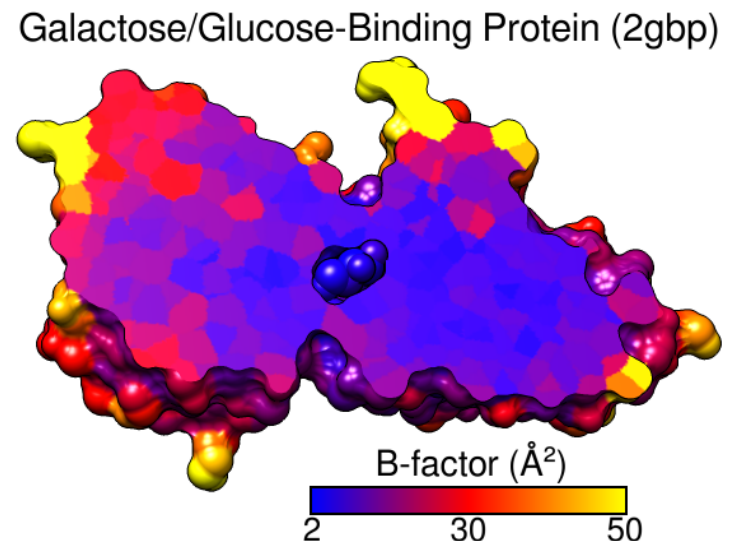
In theory, the B-factor indicates the true static or dynamic **mobility** of an atom. However, it can also indicate where there are errors in model building.

The B-factor of atom i is related to its mean square displacement U_i :

$$B = 8 \pi^2 U_i^2$$

In general, protein structures (should) have larger B-factors in loop regions and on the protein surface and low B-factors in the protein core.

<http://pldserver1.biochem.queensu.ca/~rlc/work/teaching/definitions.shtml>
<https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/tutorials/bfactor.html>



Water in protein structures

Proteins also contain **internal water molecules**!

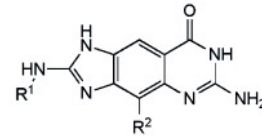
-> waters fill spaces

-> waters are required to aid the protein folding process

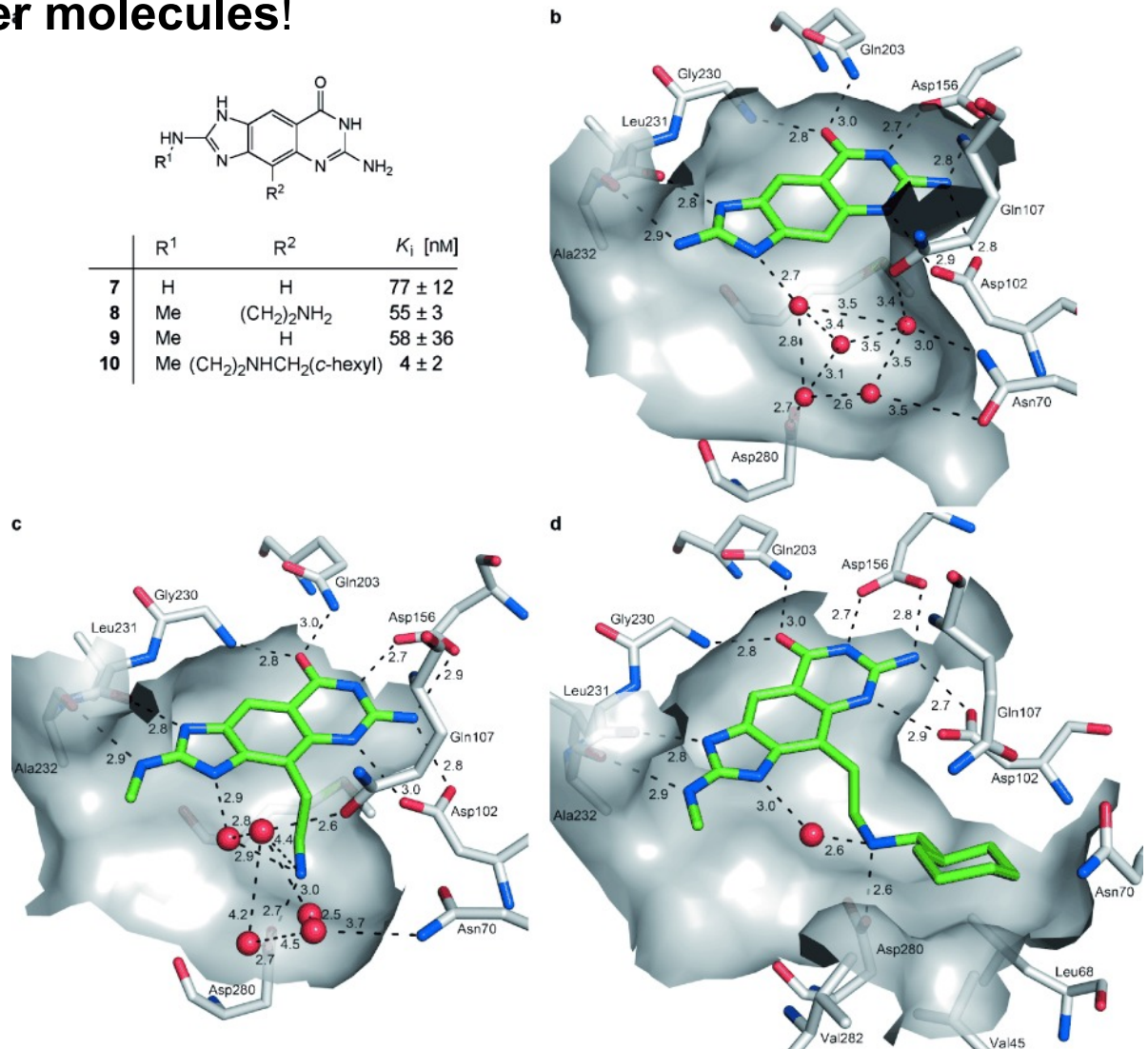
-> waters support conformational flexibility of the protein

-> waters are important for biomolecular recognition

-> waters can only be detected at resolution $< 2.3 \text{ \AA}$ or so.



	R ¹	R ²	K _i [nM]
7	H	H	77 ± 12
8	Me	(CH ₂) ₂ NH ₂	55 ± 3
9	Me	H	58 ± 36
10	Me	(CH ₂) ₂ NHCH ₂ (c-hexyl)	4 ± 2



Persch et al. Angew. Chemie 54, 3290-3327 (2015)

Occupancy

1.1 Å structure of heterogeneous nuclear ribonucleoprotein A1:

6 amino acids have **alternative side chain conformations**

- Glu24, Gln36, and Lys78 are located in loop regions and are exposed to the solvent.

- Phe17 on $\beta 1$, Val44 on $\beta 2$ and Phe59 on $\beta 3$ are located at the RNA-binding surface.

Phe17 side chain occupancies: 0.65 and 0.35.

Phe59 side chain occupancies: 0.57 and 0.43

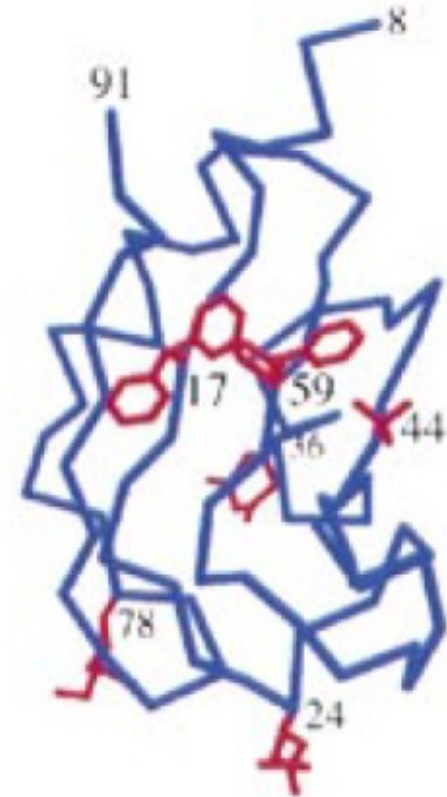
Not all of these conformations can be independently adopted by these residues because of potential steric clashes.

Permissible combinations are:

(i) Phe17A/Phe59A/Val44A,B,C – occupancy 0.57

(ii) Phe17A/Phe59B/Val44A – occupancy 0.08

(iii) Phe17B/Phe59B/Val44A - occupancy 0.35.



Head of PDB file 1L3K

REMARK 3 OTHER REFINEMENT REMARKS: RESIDUES PHE 17, VAL 44 AND PHE 59
REMARK 3 SHOW CORRELATED DISORDER IN THE SIDE CHAIN CONFORMATIONS AND
REMARK 3 THIS BEHAVIOR WAS TAKEN INTO CONSIDERATION IN REFINEMENT. THE
REMARK 3 RESIDUES WERE SPLIT IN FIVE PARTS -- B, C, D, K, L,
REMARK 3 CORRESPONDING TO THE FIVE PERMISSIBLE COMBINATIONS OF
REMARK 3 CONFORMATIONS OF PHE 17, PHE 59, AND VAL 44 ...

Alternative conformations are only detected in high-resolution data.

Vitali et al. *Nucl Ac
Res* (2002) 30,
1531–1538

PDB file 1L3K

ATOM 338 CB	BVAL A 44	-23.016 -1.594 -1.744	0.19	17.60 C
ATOM 339 CB	CVAL A 44	-23.016 -1.594 -1.744	0.20	17.60 C
ATOM 340 CB	DVAL A 44	-23.016 -1.594 -1.744	0.18	17.60 C
ATOM 341 CB	KVAL A 44	-23.016 -1.594 -1.744	0.35	17.60 C
ATOM 342 CB	LVAL A 44	-23.016 -1.594 -1.744	0.08	17.60 C

CB has the same position in the 5 conformers

ATOM 343 CG1	BVAL A 44	-22.101 -2.293 -0.750	0.19	21.01 C
ATOM 344 CG1	CVAL A 44	-22.465 -1.845 -3.138	0.20	21.66 C
ATOM 345 CG1	DVAL A 44	-24.405 -2.206 -1.621	0.18	25.18 C
ATOM 346 CG1	KVAL A 44	-24.405 -2.206 -1.621	0.35	25.18 C
ATOM 347 CG1	LVAL A 44	-24.405 -2.206 -1.621	0.08	25.18 C

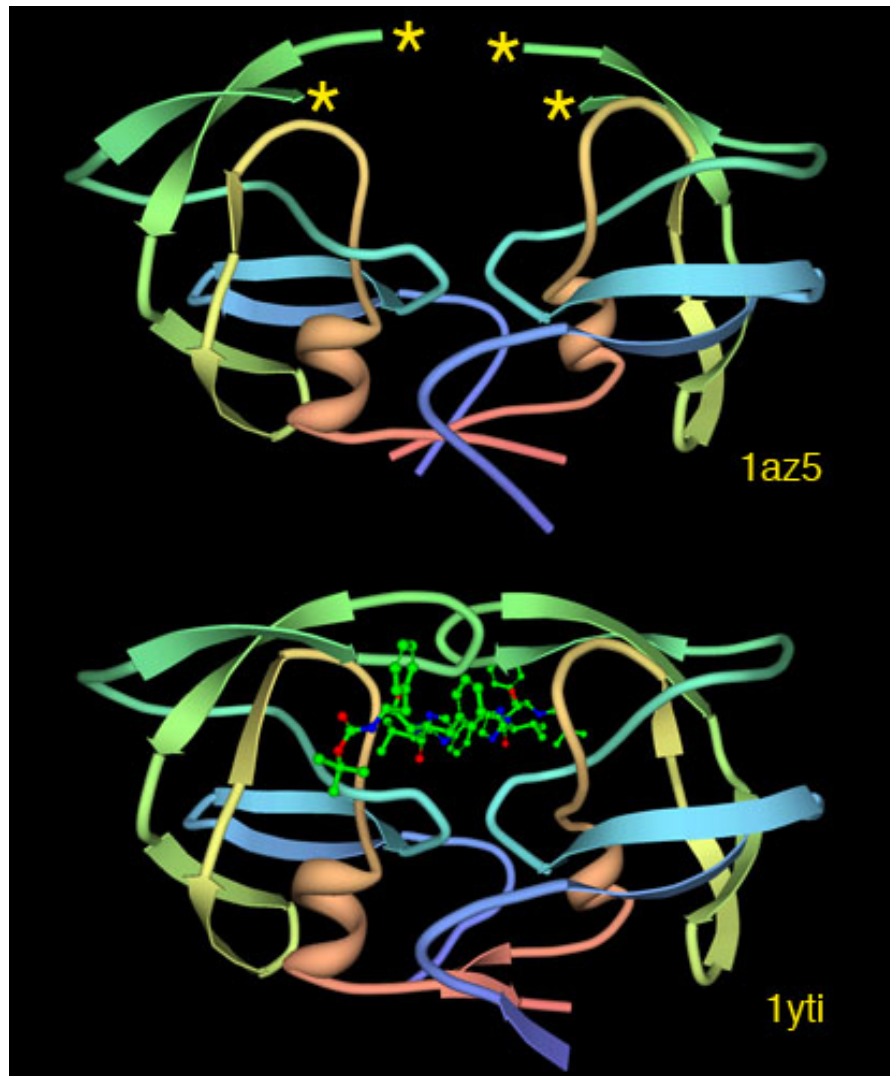
3 alternative conformations: B, C, D/K/L
D, K, L conformers have the same position, but different occupancies

ATOM 348 CG2	BVAL A 44	-24.405 -2.206 -1.621	0.19	25.18 C
ATOM 349 CG2	CVAL A 44	-22.101 -2.293 -0.750	0.20	21.01 C
ATOM 350 CG2	DVAL A 44	-22.465 -1.845 -3.138	0.18	21.66 C
ATOM 351 CG2	KVAL A 44	-22.465 -1.845 -3.138	0.35	21.66 C
ATOM 352 CG2	LVAL A 44	-22.465 -1.845 -3.138	0.08	21.66 C

3 alternative conformations: B, C, D/K/L
D, K, L conformers have the same position

Vitali et al. *Nucl Ac Res* (2002) 30,
1531–1538

Missing loops and tails



www.rcsb.org

Top: X-ray structure of SIV protease solved without its active site (PDB entry 1az5). The protein contains 2 loops (“flaps”) that were too flexible to be detected in the experiment (shown with stars).

Q: are the loops missing from the protein?

Bottom: when the protein was crystallized with inhibitors, however, the loops adopted a stable structure that may be detected (PDB entry 1yti).

Alternative conformations compatible with data

Are X-ray structures of proteins uniquely defined by the data?

Answer: only in the case of ultra-high-resolution data.

As a test, 10 and 20 independent conformers of 3 proteins were generated with a discrete restraint-based modeling algorithm, called RAPPER, based on propensity-weighted ϕ/ψ and χ angle sampling of the protein backbone.

The PDB structure was used to restrain conformational sampling to conformations whose C_{α} coordinates were within 2 Å of the C_{α} atoms of the original PDB structures.

Further, all atoms were restrained to lie in regions of positive electron density in a $2F_{\text{obs}} - F_{\text{calc}}$ map phased with the PDB structure.

De Pisto, de Bakker,
Blundell, Structure 12
(2004) 831–838

Quality of alternative conformations

Table 1. Crystallographic Quality Metrics

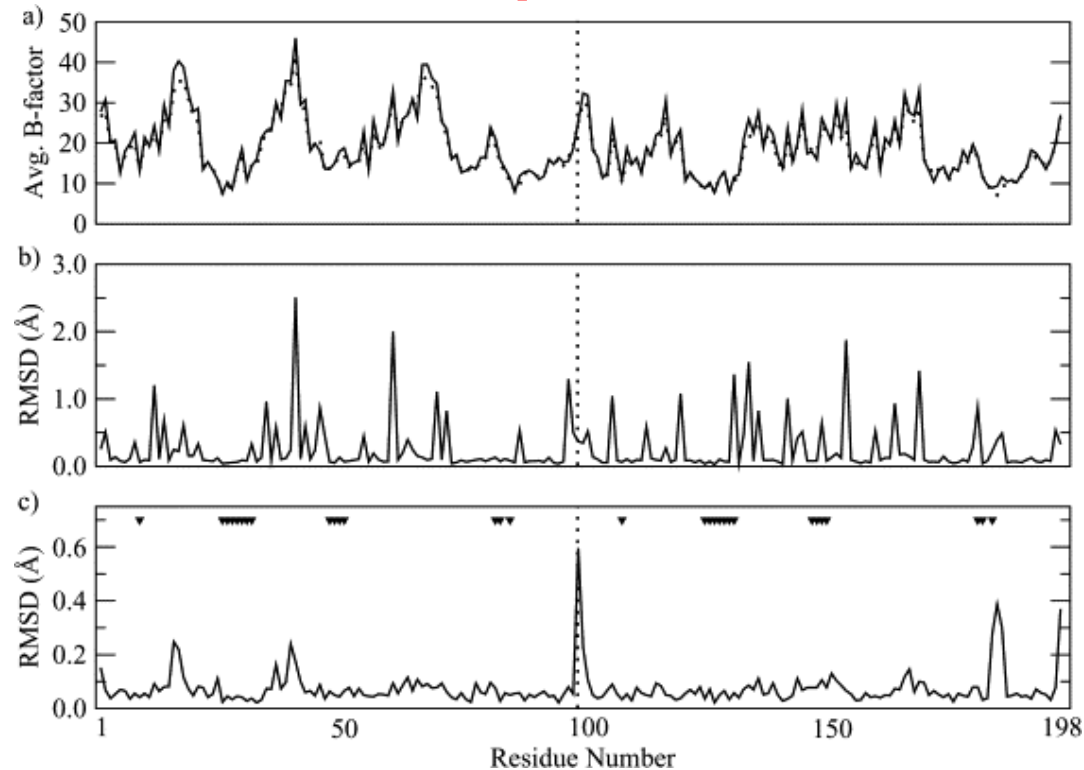
	Amicyanin		HIV Protease		Interleukin-1 β	
	PDB	Models ^a	PDB	Models ^a	PDB	Models ^a
Resolution ($\text{\AA}$)	8–1.3	20–1.3	50–1.8		15–2.3	55–2.3
R_{work} (%)						
Published ^b	15.5		19.5		15.7	
Refined ^{c,d}	15.0	14.3–14.8	19.4	17.5–18.3	16.0	15.7–16.2
R_{free} (%)						
Published ^b			23.0		21.0	
Refined ^{c,d}		16.8–17.1	22.6	20.9–21.9		20.3–21.7
Real-space R	0.981	0.980	0.964	0.967	0.850	0.846
Rms bond ^e ($\text{\AA}$)	0.014	0.012	0.014	0.011	0.018	0.017
Rms angles ^e ($^{\circ}$)	2.36	1.48	1.84	1.51	2.36	1.70
Allowed ϕ/ψ (%)	100.0	99.6	100.0	99.8	98.0	98.6
Bad rotamers ^f	0/85	0.2/85	2/166	3.3/166	12/129	12.8/140
Esu from R factor ^g ($\text{\AA}$)	0.05		0.15		0.23	

Alternative conformations have equal or better R_{free} values than the PDB structure and lower RMS deviations of bond lengths and bond angles from the ideal values.

→ they look like “better” structures

De Pisto, de Bakker,
Blundell, Structure 12
(2004) 831–838

B-factors and RMSD per Residue for HIV Protease



De Pisto, de
Bakker, Blundell,
Structure 12 (2004)
831–838

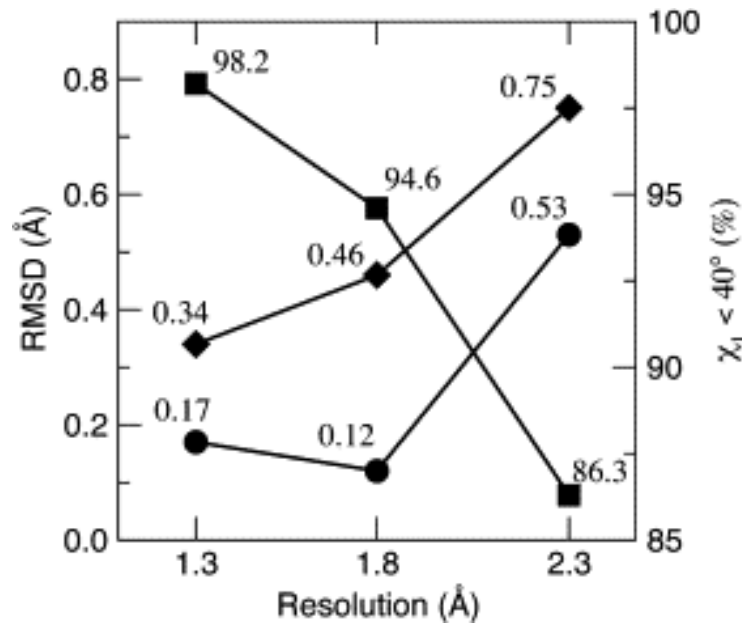
Averaged B factor (A) of the PDB structure (dots) and the 5 alternate models (line).
Note the similarity of the average B factors between the PDB and RAPPER models.

All-atom (B) and main chain (C) RMSD for each residue of the alternate models
compared to the PDB structure.

Triangles: residues in contact with the inhibitor molecule.

Vertical dotted line: break between the 2 chains of the protease dimer.

Difference between models and PDB structure



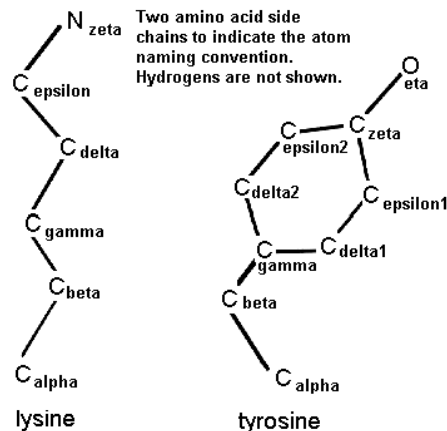
Amicyanin (1.3 Å resolution),
HIV protease (1.8 Å)
h-IL1 β (2.3 Å).

Pairwise differences among the PDB
and alternate models increase with
lowered resolution

Circles: main chain RMSD

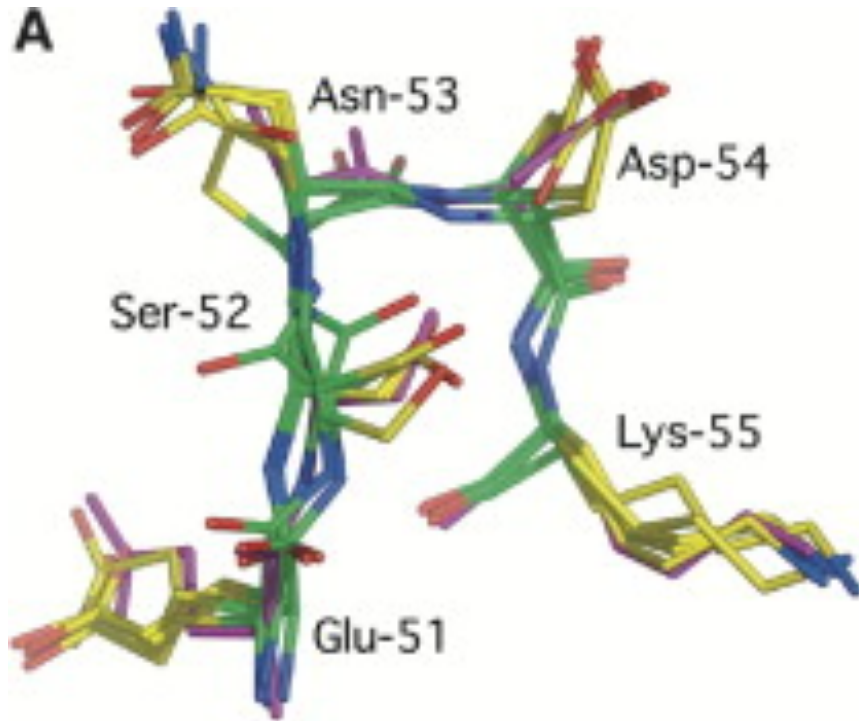
Diamonds: all-atom RMSD

Squares: rotamer state conservation :
fraction of residues with side chain χ_1
angle within 40° of the PDB structure.



De Pisto, de Bakker,
Blundell, Structure 12
(2004) 831–838

Main Chain and Side Chain Heterogeneity in Human Interleukin-1 β (2.3 Å)



Shown are residues 51–55 from h-IL1 β .

The PDB structure is in magenta. The 5 alternate models generated with RAPPER are colored according to: **nitrogen**; **oxygen**; **main chain carbon**; **side chain carbon**.

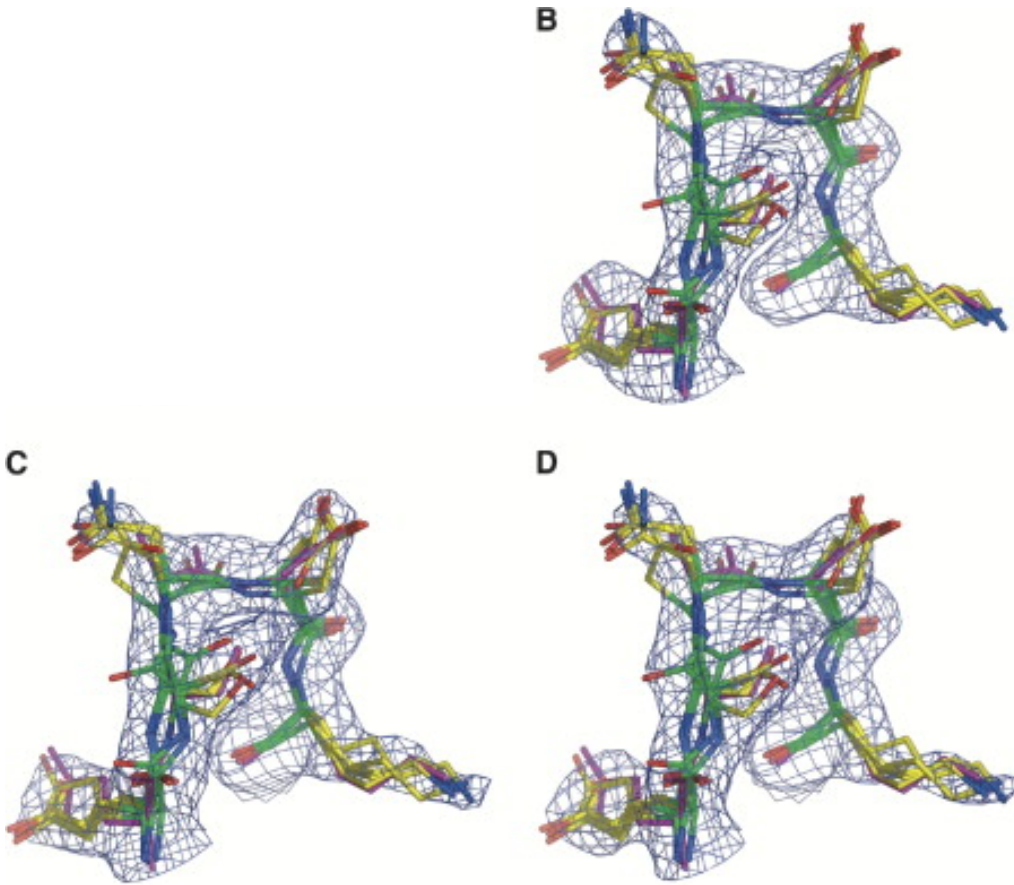
Note the pronounced backbone variability and side chains with anisotropic motion (Ser52, Asn53) and multiple discrete conformations (Glu51, Asp54, Lys55).

De Pisto, de
Bakker, Blundell,
Structure 12 (2004)
831–838

Main Chain and Side Chain Heterogeneity in Human Interleukin-1 β

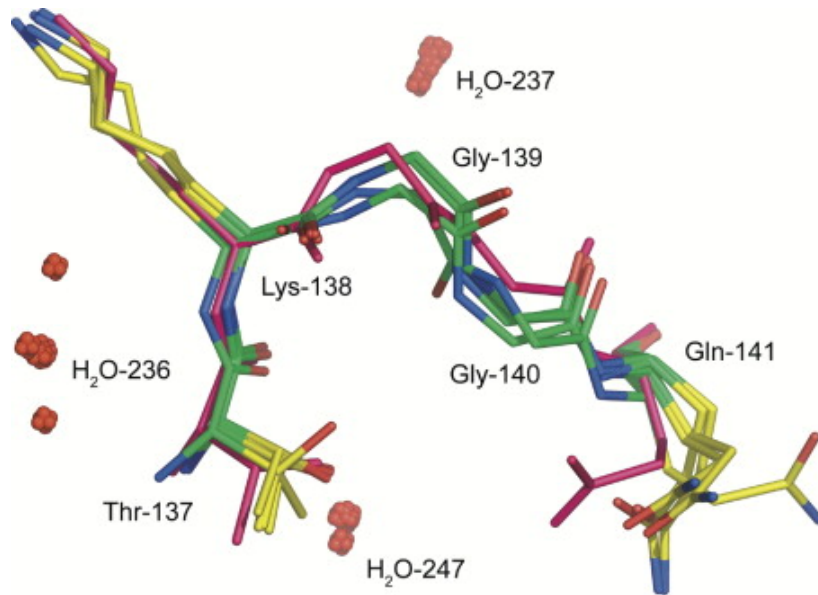
(B)–(D) show simulated-annealing omit maps contoured at 1 σ , for the original PDB structure (B) and alternate models 2 (C) and 3 (D).

→ Maps are practically indistinguishable.



De Pisto, de
Bakker, Blundell,
Structure 12 (2004)
831–838

Main Chain and Water Heterogeneity in Human IL-1 β



Residues 137–141 from h-IL1 β are shown, highlighting backbone variability and disordered side chains and waters.

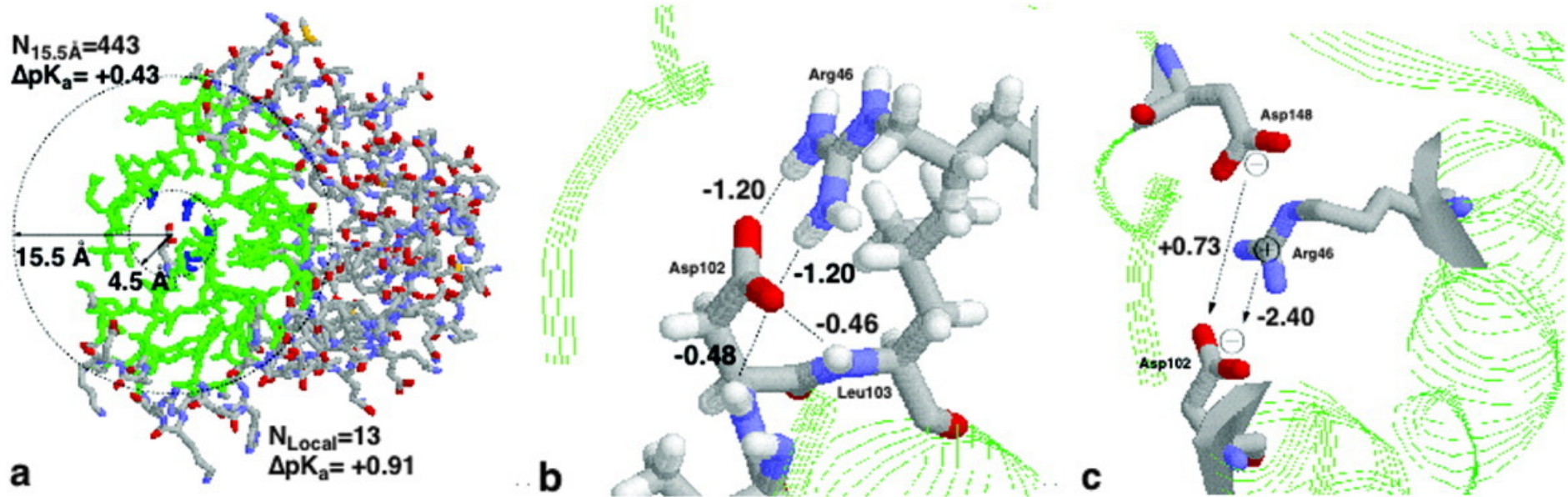
Note the significant variability in the main chain (Gly139 and Gly140) and side chain (Thr137 and Lys138) conformations, while Gln141 appears to be total disordered.

Waters H₂O-237 and H₂O-247 are well ordered, whereas H₂O-236 has a mean square displacement of 3.5 Å.

→ Mid-range resolution structures do **not** provide unique information about atomic positions and relative orientations.

Assigning macromolecular protonation states

At pH = 7, Asp and Glu residues are in principle negatively charged, Arg and Lys residues are in principle positively charged. His can be neutral or positive ($pK_a = 6.5$)



The determinants of the pK_a value of Asp102 in RNase H (2RN2): (a) desolvation effects, (b) hydrogen bonding, and (c) Coulomb interactions.

Whether Asp148 titrates before or after Asp102 depends on the relative pK_a values.

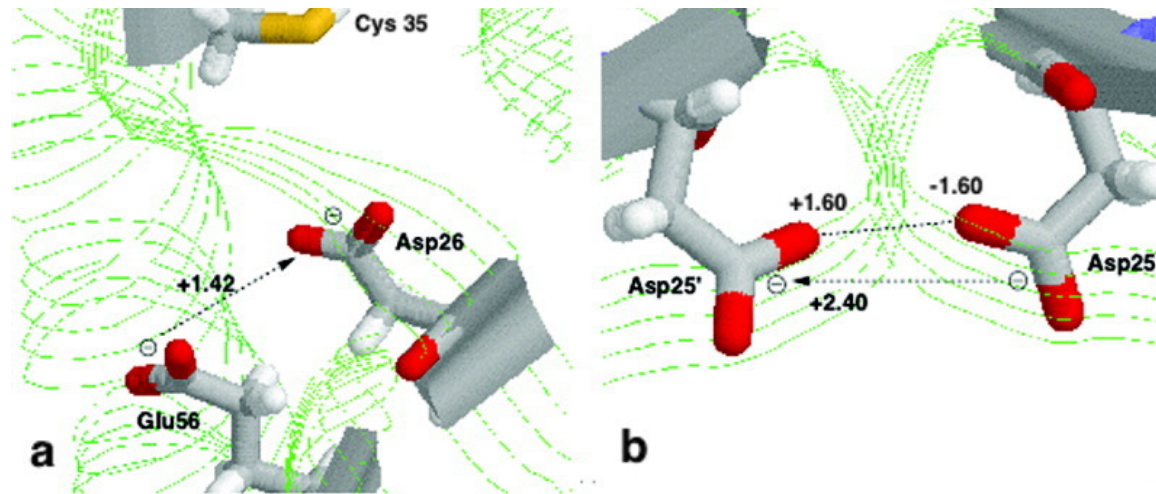
Li et al. Proteins, 61,
704-721 (2005)

Assigning macromolecular protonation states

Sites with unusual pK_a values:

(a) Asp26 of human thioredoxin, 1ERT;

(b) Asp25 of chain A and Asp25 of chain B in HIV-1 protease dimer, 1HPX;



The experimental pK_a values of **Asp26** in oxidized and reduced **thioredoxin** are 8.1 and 9.9 pH units. These are among the highest carboxyl pK_a values observed in a protein.

Asp25 of HIV-1 protease: PROPKA predicts pK_a values of 3.8 and 9.3. Exp values are unknown → One Asp will be protonated, one Asp deprotonated.

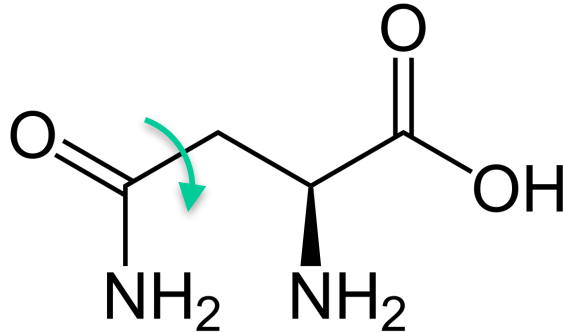
pK_a values can be accurately computed by solving the Poisson-Boltzmann equation.

Orienting Asn / Gln side chains

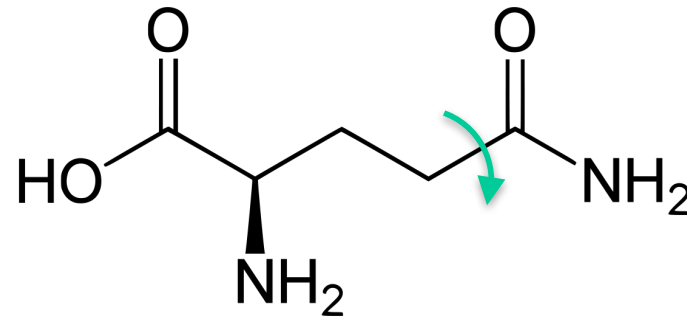
The electron density near the nitrogen and oxygen atoms of Asn and Gln amide groups is compatible with 2 rotamers that can be interconverted by a 180° flip.

Therefore, electron density maps obtained from X-ray diffraction experiments of protein crystals yield the **positions** of the oxygen and nitrogen atoms with high precision but not their **identity**.

This sometimes results in the assignment of wrong rotamers.



asparagine

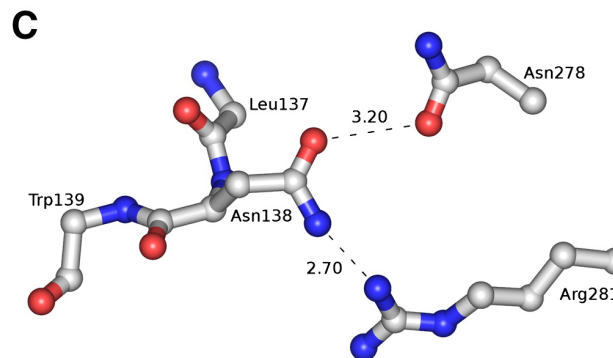
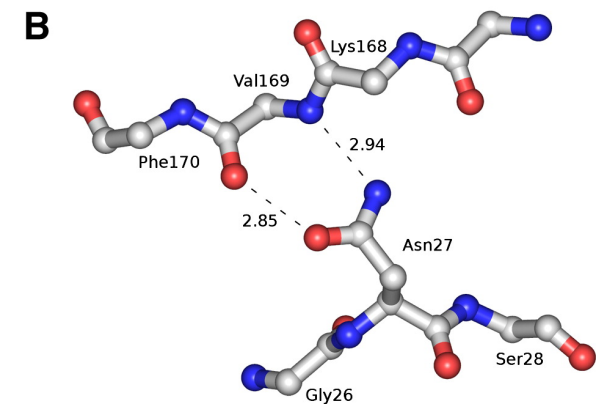
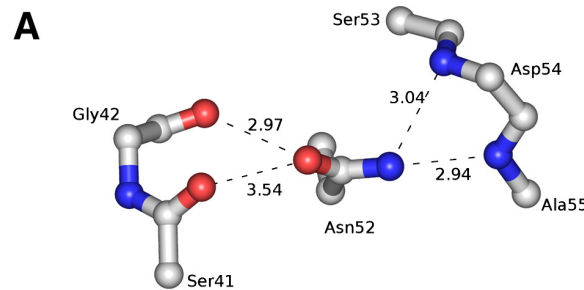


glutamine

Examples of wrong Asn side chain orientations

(A) Asn-52 of dethiobiotin synthetase, 1dad (resolution 1.6 Å):

its amide oxygen and nitrogen atoms have unfavorable interactions with the main chain oxygen atoms of Ser-41 and Gly-42, and the backbone nitrogen atoms of Asp-54 and Ala-55, respectively.



(B) Asn-27 of cutinase, 1cus (resolution 1.25 Å)

(C) Asn-138 of concanavalin B, 1cnv (resolution 1.65 Å).

In all 3 cases, the Asn side chain should be **flipped** by 180° .

Such cases can be determined by algorithms that optimize the hydrogen bond network.

Statistics on protein structures: derive understanding from statistical enrichment

Idea: some positions in/on protein structures are energetically more favorable for certain amino acids → these amino acids should be **enriched** in these regions.

The energetics is difficult to estimate.

BUT the frequency of amino acids can be easily computed as a statistical average over all known protein structures.

Hayat et al. Comput
Biol Chem (2011)
35, 96–107

Statistics on protein structures

Q: how does the amino acid composition of trans-membrane barrels (TMB) differ in the membrane from that in the cytosol?

Split the membrane into the non-polar membrane-core (aliphatic lipid tails) and the medium-polarity membrane interface region (phospholipid head-groups).

We compiled a non-redundant data set of known TMB structures by removing those protein sequences for which less than 20 homologous sequences were found or where the pair-wise sequence identity of the aligned retrieved sequences was greater than 80%.

The final data set for TMBs comprises of 20 protein chains with 1725 and 572 TM residues in the hydrophobic core and interface regions, respectively

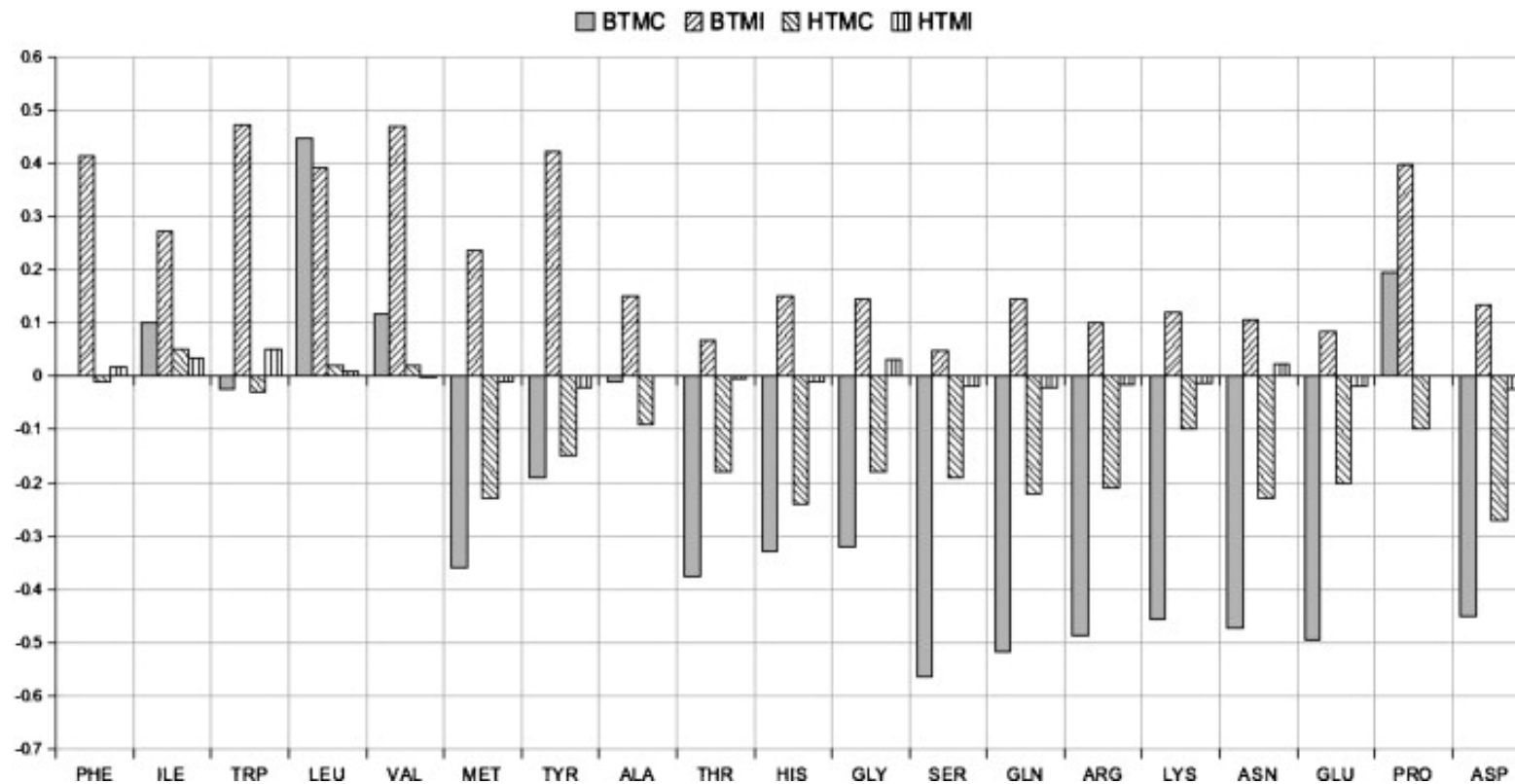
Hayat et al. Comput
Biol Chem (2011)
35, 96–107

Propensity scale: over / under-representation

Enrichment / depletion of amino acids in

- **membrane core** of beta-barrels (BTMC) and helical membrane proteins (HTMC) or
- **interface region** of membrane (BTMI and HTMI).

In each case, values are log-ratios of this region vs. the full protein sequences.



Composition of protein interfaces

Q: Are protein-protein interfaces comparable to protein-ligand interfaces?

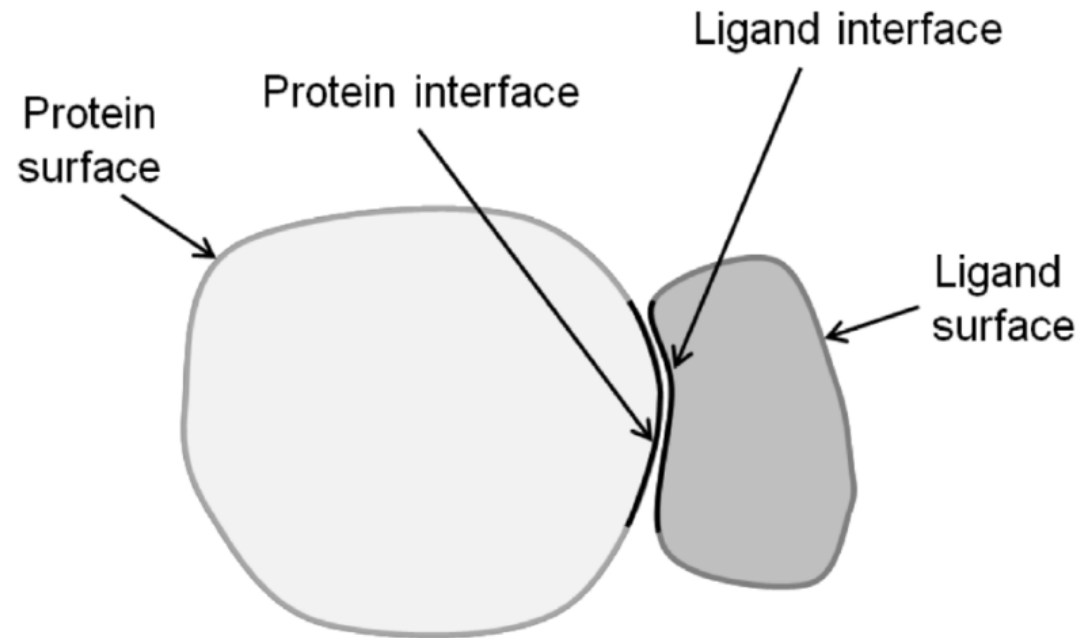
Dataset : 174 protein-protein complexes and 161 protein-ligand complexes.

These complementary PP and PL datasets fulfill the following criteria:

- (i) PP: PL pairs represent pairs of complexes, where one protein may bind either a second protein or a small molecule ligand at the same interface,
- (ii) every pair of the dataset is represented as $(P_{i1}, P_{i2}): (P_{i3}, L_j)$, where P_{i1} , P_{i2} and P_{i3} are three proteins and L_j is a small molecule ligand,
- (iii) P_{i1} and P_{i3} share at least 40% sequence identity, and
- (iv) the aligned positions in the binding interfaces of $P_{i1}-P_{i2}$ and $P_{i3}-L_j$ have at least 2 residues in common.

Mohamed et al.
PLoS ONE (2015)
10, e0140965

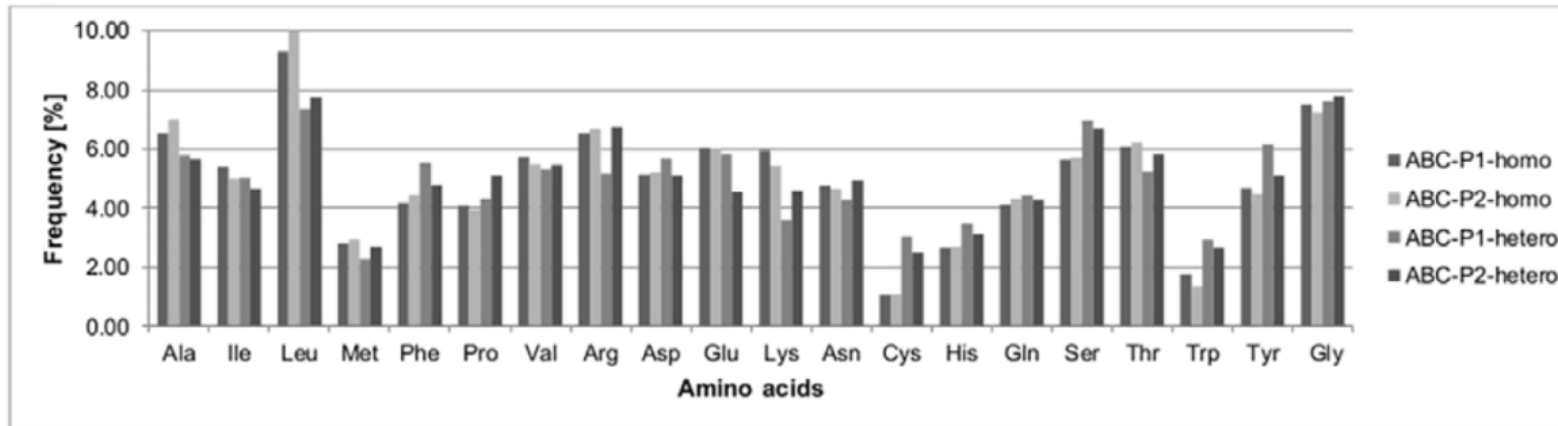
Protein:ligand interface



$$\text{Interface residue propensity } AA_j = \left(\frac{\sum \text{interface residues of type } j}{\sum \text{all interface residues}} \right) / \left(\frac{\sum \text{surface residues of type } j}{\sum \text{all surface residues}} \right)$$

An interface residue propensity of > 1.0 indicates that a residue type occurs more frequently in interfaces than on the protein surface in general.

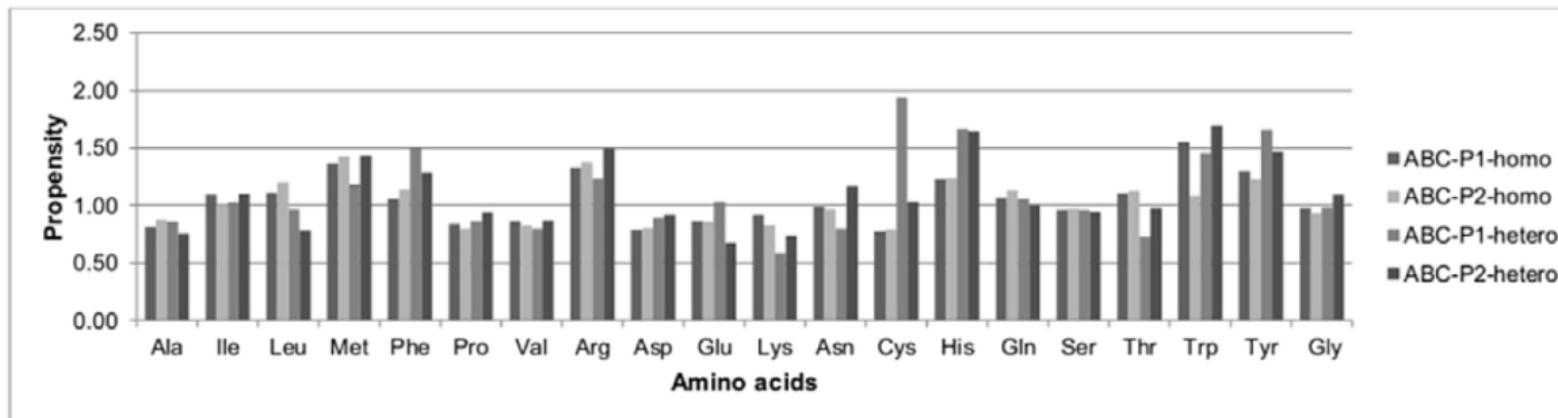
Frequencies vs. propensities



Frequencies are raw counts.

Propensities are normalized by the proportion of the amino acids.

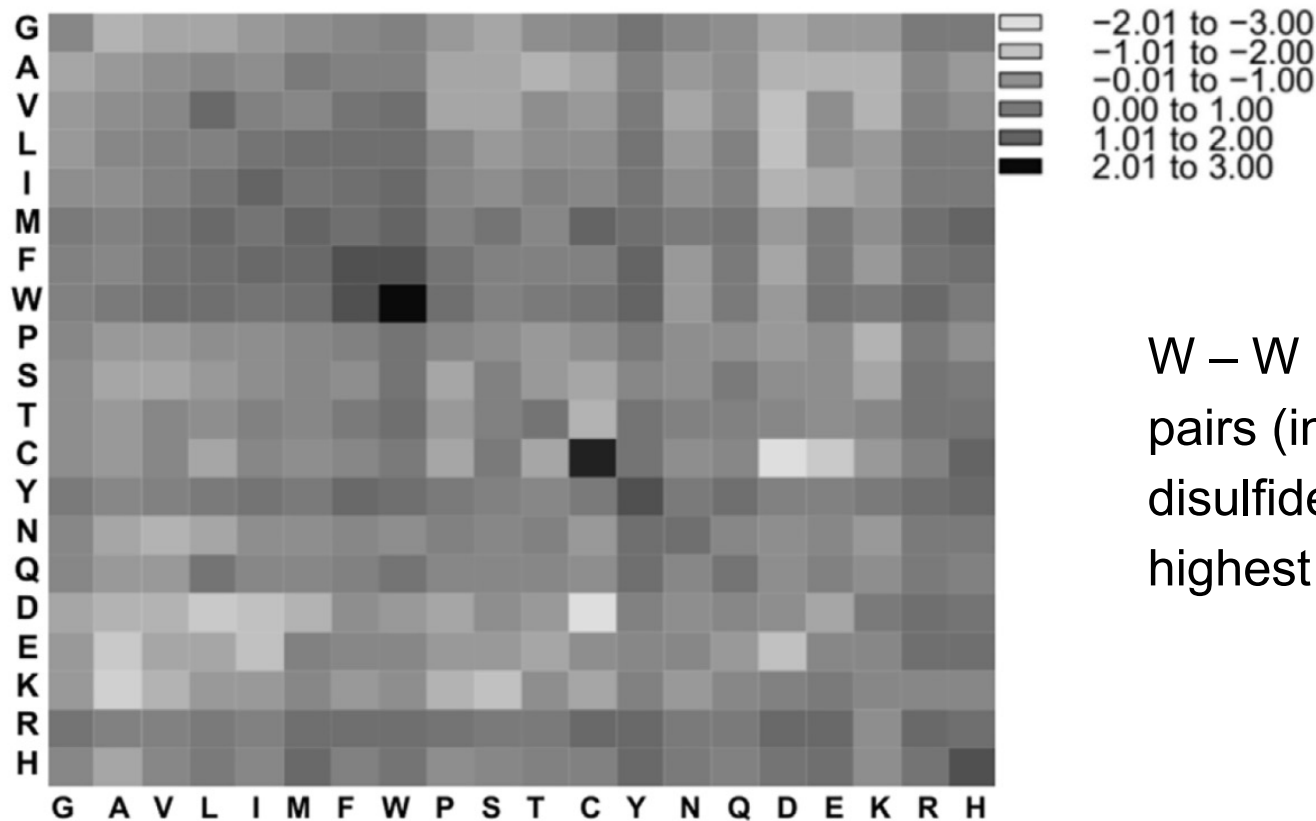
B



Trp has overall a low frequency, but is frequently found at interfaces.

Mohamed et al.
PLoS ONE (2015)
10, e0140965

Amino acid pairing propensity at interface



W – W pairs and C – C pairs (inter-protein disulfide bridges) have highest propensities.

$$\text{Normalization} = \frac{\left(\frac{\sum \text{contacts of residue pair } XY}{\sum \text{all residue contacts}} \right)}{\left(\frac{\sum \text{observed } X \text{ on surface}}{\sum \text{all surface residues of first protein}} \right) \left(\frac{\sum \text{observed } Y \text{ on surface}}{\sum \text{all surface residues of second protein}} \right)}$$

Mohamed et al.
PLoS ONE (2015)
10, e0140965

Statistical potential: Boltzmann inversion

$$P(r) = \frac{1}{Z} e^{-\frac{F(r)}{kT}}$$

Probability $P(r)$ at position r according to Boltzmann distribution as a function of the free energy $F(r)$ at this position.

k is the Boltzmann constant, T is the temperature.

This can be re-arranged into

$$F(r) = -kT \ln P(r) - kT \ln Z$$

and taken with respect to a reference state with distribution $Q_R(r)$.

This is called a statistical potential,

$$\Delta F(r) = -kT \ln \frac{P(r)}{Q_R(r)}$$

e.g. from the probability to find two amino acids at a certain distance r from each other one can derive their effective interaction free energy.

Sippl MJ (1990). J Mol Biol. **213**: 859–883.
www.wikipedia.org

Rosetta energy function

David Baker and co-workers justified PMFs from a Bayesian point of view and used these in the construction of the coarse grained ROSETTA energy function.

According to Bayesian probability calculus, the conditional probability $P(X | A)$ of a structure X , given the amino acid sequence A , can be written as:

$$P(X | A) = \frac{P(A | X) P(X)}{P(A)} \propto P(A | X) P(X)$$

$P(X | A)$ is proportional to the product of the likelihood $P(A | X)$ times the prior $P(X)$

Sippl MJ (1990). J Mol Biol. **213**: 859–883.
www.wikipedia.org

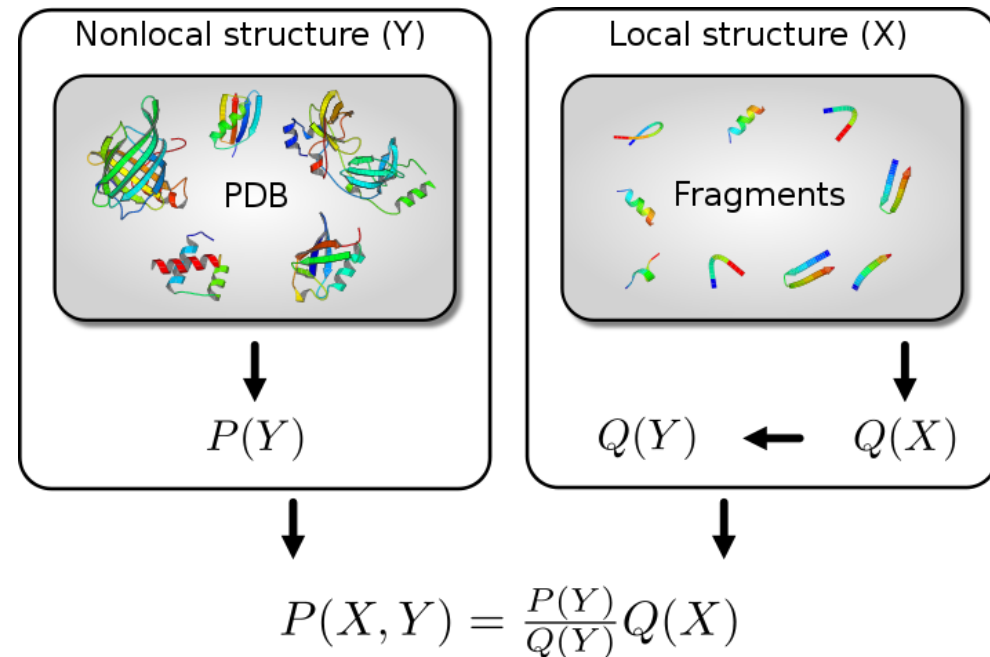
Rosetta energy function

By assuming that the likelihood can be approximated as a product of pairwise probabilities, and applying Bayes' theorem, the likelihood can be written as:

$$P(A | X) \approx \prod_{i < j} P(a_i, a_j | r_{ij}) \propto \prod_{i < j} \frac{P(r_{ij} | a_i, a_j)}{P(r_{ij})}$$

where the product runs over all amino acid pairs a_i, a_j (with $i < j$), and r_{ij} is the distance between amino acids i and j .

The assumption that the likelihood can be expressed as a product of pairwise probabilities is questionable.



Sippl MJ (1990). J Mol Biol. **213**: 859–883.
www.wikipedia.org

Science of HIV project

http://scienceofhiv.org/wp/?page_id=20