V 8 – Analysis of protein-protein binding

- Construct cliques in a sparse PPI network
- Modelling by homology
- Structural properties of PP interfaces
- Predicting PP properties / affinity of interactions
- Review V1 – V7

Fri, May 11, 2018

Mesoscale properties of networks
- identify cliques and highly connected clusters

Most relevant processes in biological networks correspond to the **mesoscale** (5-25 genes or proteins), not to the entire network.

However, it is computationally enormously expensive to study mesoscale properties of biological networks.

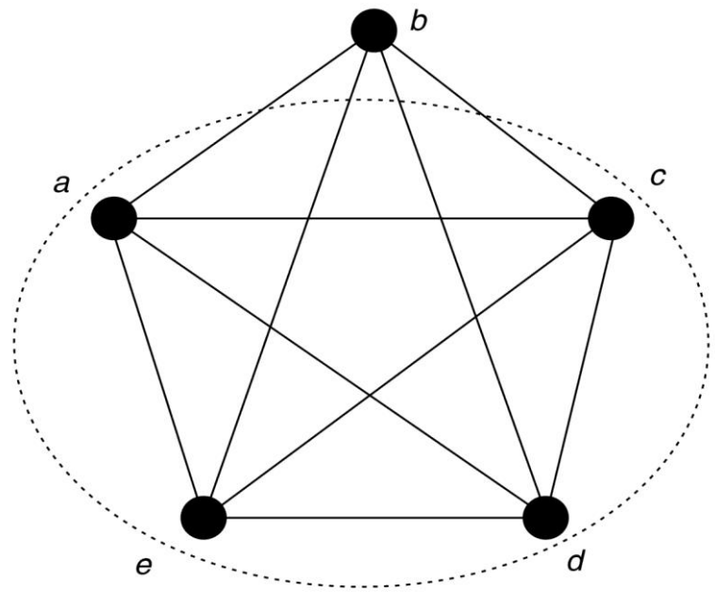
E.g. a network of 1000 nodes contains 1×10^{23} possible 10-node sets.

Spirin & Mirny analyzed combined network of protein interactions in *S. cereviseae* with data from CELLZOME, MIPS, BIND: 6500 interactions.

Identify connected subgraphs

Aim: identify **fully connected subgraphs** (cliques) in protein interaction network.

A clique is a set of nodes that are all neighbors of each other.



The „**maximum clique problem**“ – finding the largest clique in a given graph is known be NP-hard.

In this example, the whole graph is a clique and consequently any subset of it is also a clique, for example *{a,c,d,e}* or *{b,e}*.

A **maximal clique** is a clique that is not contained in any larger clique. Here only *{a,b,c,d,e}* is a maximal clique.

In general, protein complexes need not to be fully connected.

Spirin, Mirny,
PNAS 100, 12123 (2003)

Identify all fully connected subgraphs (cliques)

The general problem - finding all cliques of a graph - is very hard.

But the protein interaction graph is quite **sparse**:
interactions (edges) is similar to # proteins (nodes)).

-> the cliques can be found relatively quickly in the PPI network.

Idea:
cliques of size *n* can be found by enumerating the cliques of size *n-1* etc.

Spirin, Mirny, PNAS 100, 12123 (2003)

Identify all fully connected subgraphs (cliques)

Spirin & Mirny started their search for cliques with $n = 4$.

Consider all (known) pairs of edges (6500 × 6500 protein interactions).

For every **pair** *A-B* and *C-D* check whether there are edges

between *A* and *C, A* and *D, B* and *C,* and *B* and *D.*

If these edges are present, *ABCD* is a **clique**.

For every clique identified, *ABCD,* check all proteins in the PPI network.

For every additional protein *E*:

if all of the interactions *E-A, E-B, E-C,* and *E-D* exist,

then *ABCDE* is a clique with size 5.

Continue for *n = 6, 7, ...*

Spirin, Mirny, PNAS 100, 12123 (2003)

Identify all fully connected subgraphs (cliques)

The largest clique found in the protein-interaction network had size 14.

These results include, however, many redundant cliques.

E.g., the clique with size 14 contains 14 cliques with size 13.

To find all **nonredundant cliques**, mark all proteins in the clique of size 14.

Out of all subgraphs of size 13 pick those that have at least one protein other than marked.

After all redundant cliques of size 13 are removed,
proceed to remove redundant twelves etc.

In total, only 41 nonredundant cliques with sizes 4 - 14
were found by Spirin & Mirny.

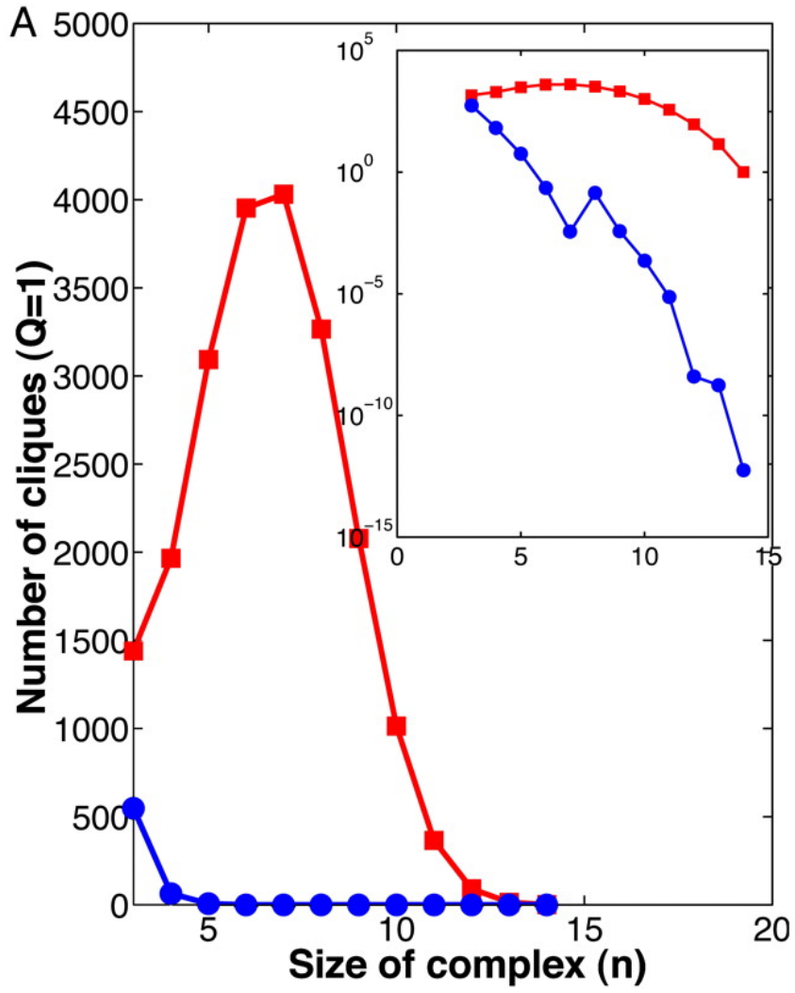
Spirin, Mirny, PNAS 100, 12123 (2003)

Statistical significance of cliques

complete cliques as a function of clique size.

Red: real network of protein interactions

Blue: > 1000 randomly rewired graphs, that have the same number of interactions for each protein.



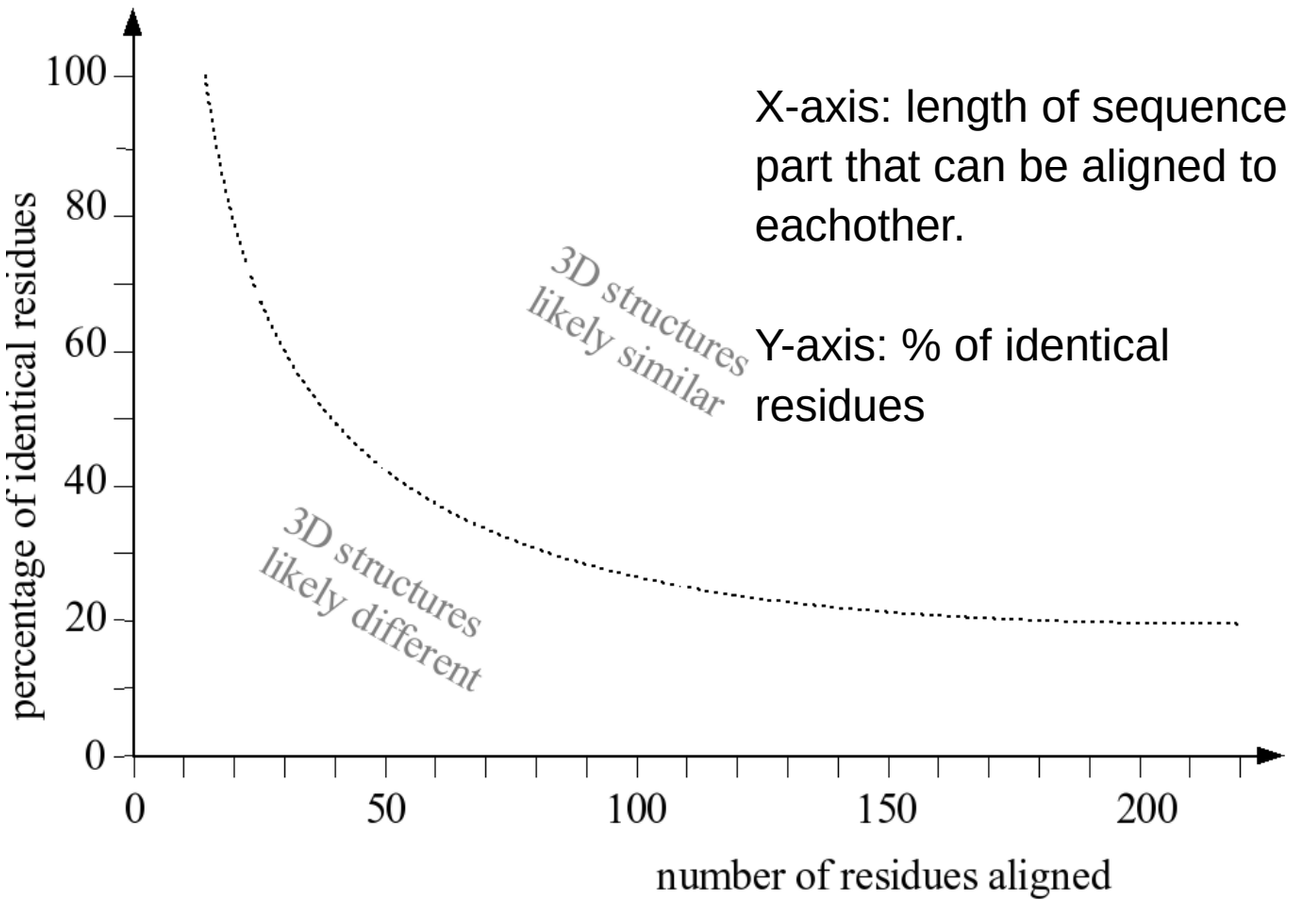
Inset shows the same plot on a log-normal scale. Note the dramatic enrichment in the number of cliques in the protein-interaction graph compared with the random graphs. Most of these cliques are parts of bigger complexes and modules.

Spirin, Mirny, PNAS 100, 12123 (2003)

3.1 Model protein structures by homology

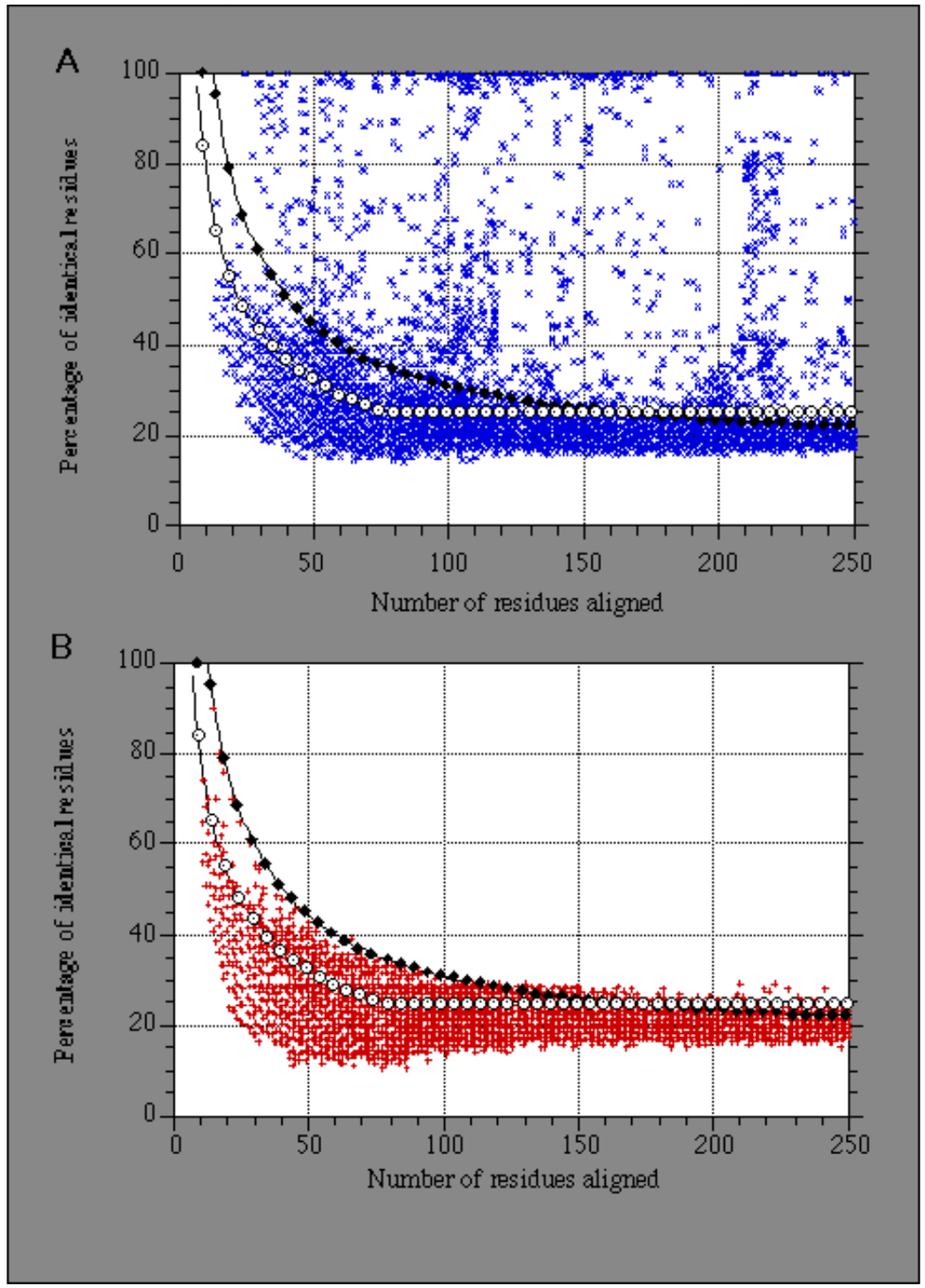
Figure shows “**twilight zone”** below the dotted line.

If two sequences A and B have a higher sequence identity than this line, their 3D structures are highly likely to be similar to eachother.



X-axis: length of sequence part that can be aligned to eachother.

Y-axis: % of identical residues



Top: sequence pairs A:B with similar structure

Bottom: pairs with different structure

Rost, Prot. Eng. 12, 85 (1999)

measure structural similarity of complexes

Critical Assessment of PRedicted Interactions (**CAPRI**) competition uses
3 criteria for ranking the protein complex predictions:

1- '**fnat**‘: the number of native residue–residue contacts in the predicted complex divided by the number of native contacts in the target.

2- **L-rms**: the backbone RMSD of the **ligands** (smaller one of both proteins) in the predicted versus the target structures.
Here, the larger proteins (**receptor**) are superimposed first.

3- **i-rms**: the RMSD of the backbone of the interface residues only, in the predicted versus the target complexes
(interface residues: here, residues with 10 Å of the other protein.
Map complementary residues in sequence alignment.)

Assessment of Blind Predictions of Protein–Protein Interactions: Current Status of Docking Methods, Mendez et. al. PROTEINS: Structure, Function, and Genetics 52:51–67 (2003)

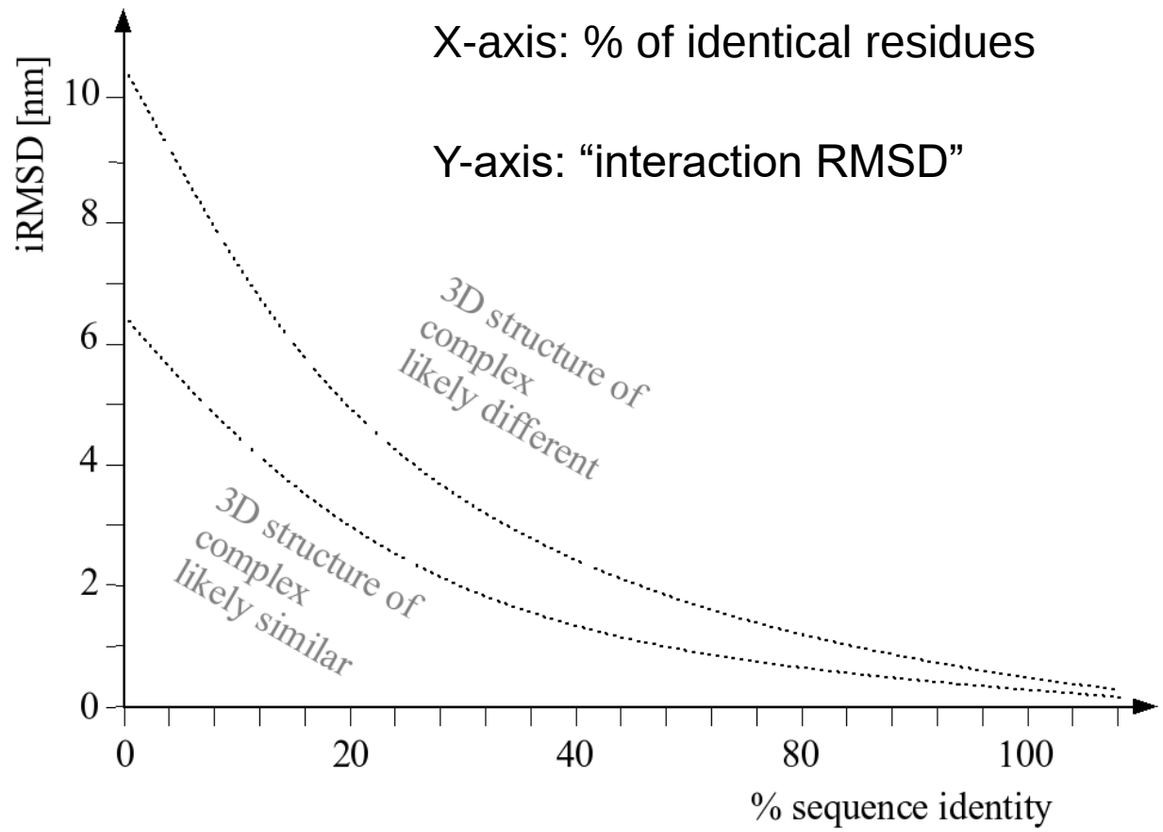
3.1 Model protein complexes by homology

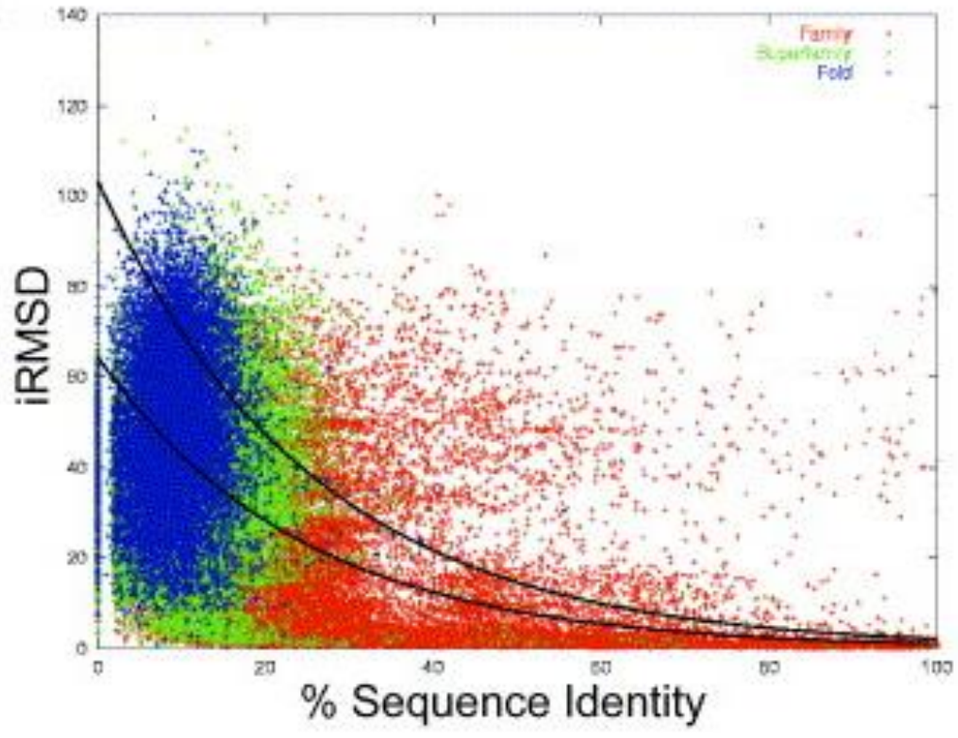
Structural similarity of protein complexes A':B' and A:B as a function of their sequence identity.

Note that x-axis and y-axis are different from previous slide.

A sequence identity level of 30-40% usually means that the binding mode of interaction is conserved (iRMSD < 3Å).

These plot show the "interaction RMSD", which is similar to L-RMSD.



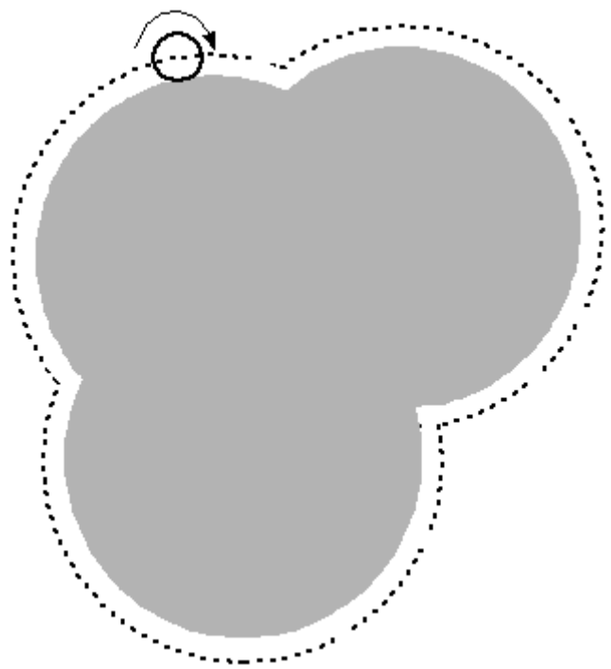


Aloy & Russell (2003) J Mol Biol 332, 989

3.2 Structural properties of PP interfaces

Size of protein-protein interface is commonly computed from **solvent-accessible surface area** (SASA) of the protein complex and of the individual proteins:

$$\Delta SASA = SASA_A + SASA_B - SASA_{AB}$$



Computation of the SASA. A small probe is rolled over the complete surface of the large molecule shown in grey. The dashed line connects the positions of the center of the probe. In three dimensions, it is a surface. Its area is the SASA.

Definition of **interface residues**:

(a) All residues that are within a cut-off distance (e.g. 5Å) to any residue of the other protein.

(b) All residues having a reduced SASA in the complex compared to the unbound state.

3.2.1 Structural properties of PP interfaces

Parameter	Protein-protein complexes	Homodimers	Weak dimers	Crystal packing
Number in dataset	70	122	19	188
Buried surface area (Å)^2	1910	3900	1620	1510
Amino acids per interface	57	104	50	48
Composition (%) Non-polar Neutral polar Charged	 58 28 14	 65 23 12	 62 25 13	 58 25 17
H-bonds per interface	10	19	7	5
Residue conservation % in core	55	60	n/a	40

Janin et al. Quart Rev Biophys 41, 133 (2008).

3.2.1 size of PP interfaces

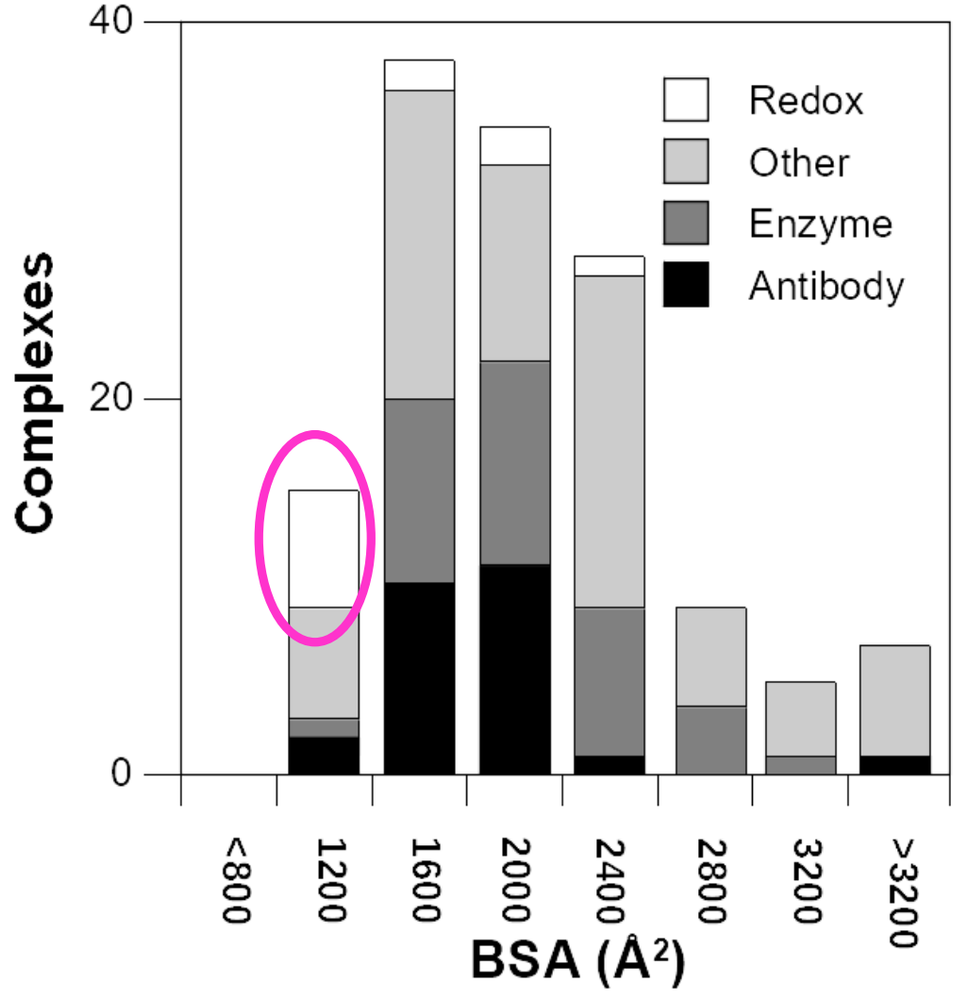
Redox complexes mediate e.g. the transfer of electrons between the binding partners.

Redox complexes possess relatively **small interfaces** -> **short life times**.

This makes biological sense. After an electron is transferred between 2 proteins, they no longer need to be bound.

In contrast, antibodies should bind their binding partners tightly so that they won't harm the organism.

The larger average interface size of antibody-antigen complexes is connected to a longer average life-time of the bound form.



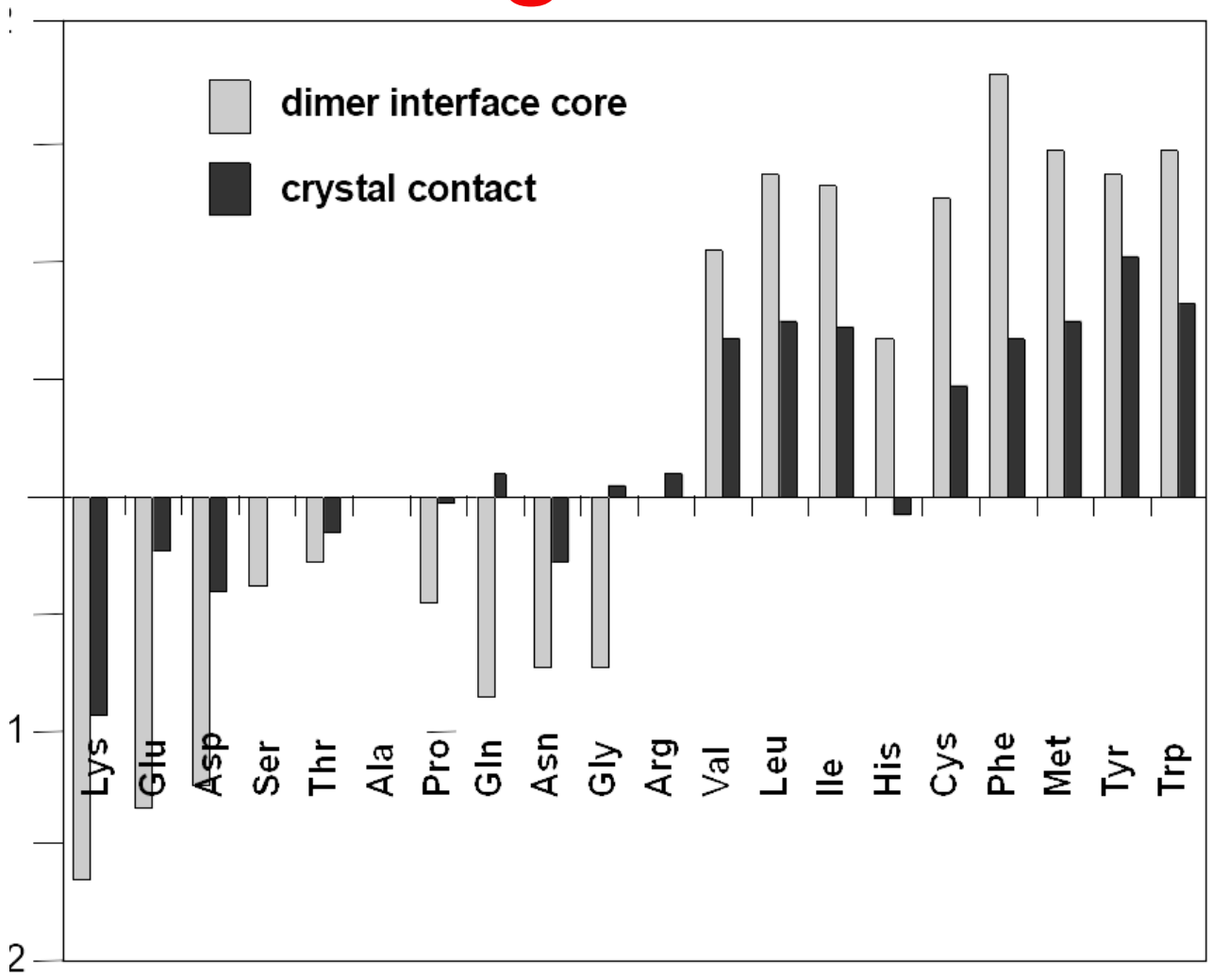
Interface size in transient protein–protein complexes. Histogram of the buried surface area (BSA) in 25 antigen–antibody complexes, 35 enzyme/ inhibitor or substrate complexes, 64 complexes of other types and in 11 redox protein complexes. The mean value of the BSA is 1290 Å^2 for the redox complexes and 1910 Å^2 for the other complexes.

Janin et al. Quart Rev Biophys 41, 133 (2008).

3.2.2 Composition of binding interfaces

Biological interfaces are enriched in aromatic (Tyr, Phe, Trp) and non-polar residues (Val, Leu, Ile, Met).
Charged side chains are often excluded from biological protein-protein interfaces except for Arg.

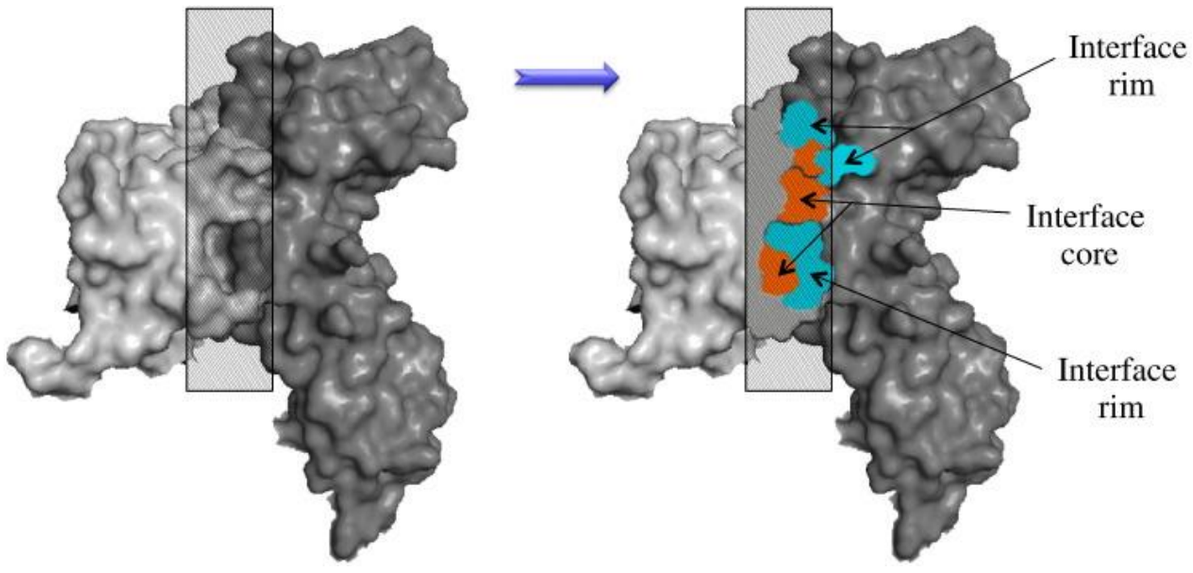
In contrast, crystal contacts contain clearly fewer hydrophobic and aromatic residues, but more charged residues than biological interfaces.
Also, the enrichment of amino acids is smaller at crystal contacts compared to biologically relevant contacts.



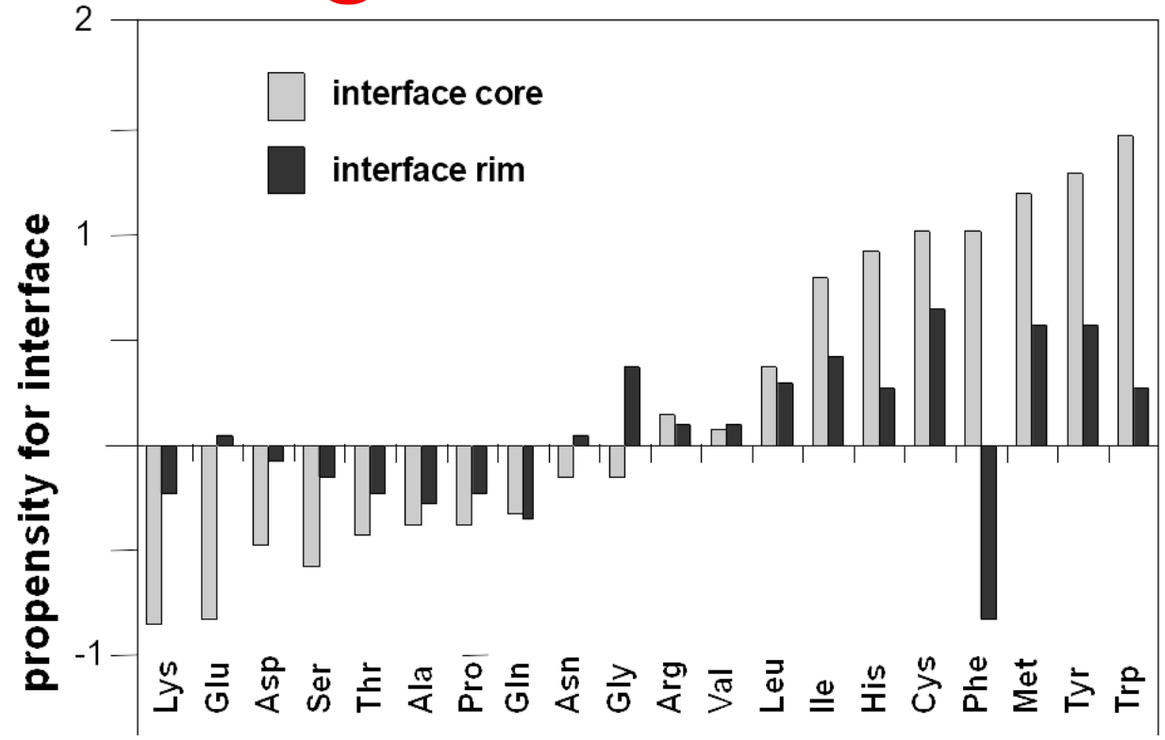
Residue propensities at protein dimer interfaces and at artificial contacts in the crystal, respectively. The propensities are derived from the relative contributions of the 20 amino acid types to the buried surface of the interfaces.

Drawn after Janin *et al.* (2008).

3.2.2 Composition of binding interfaces



David and Sternberg (2015)

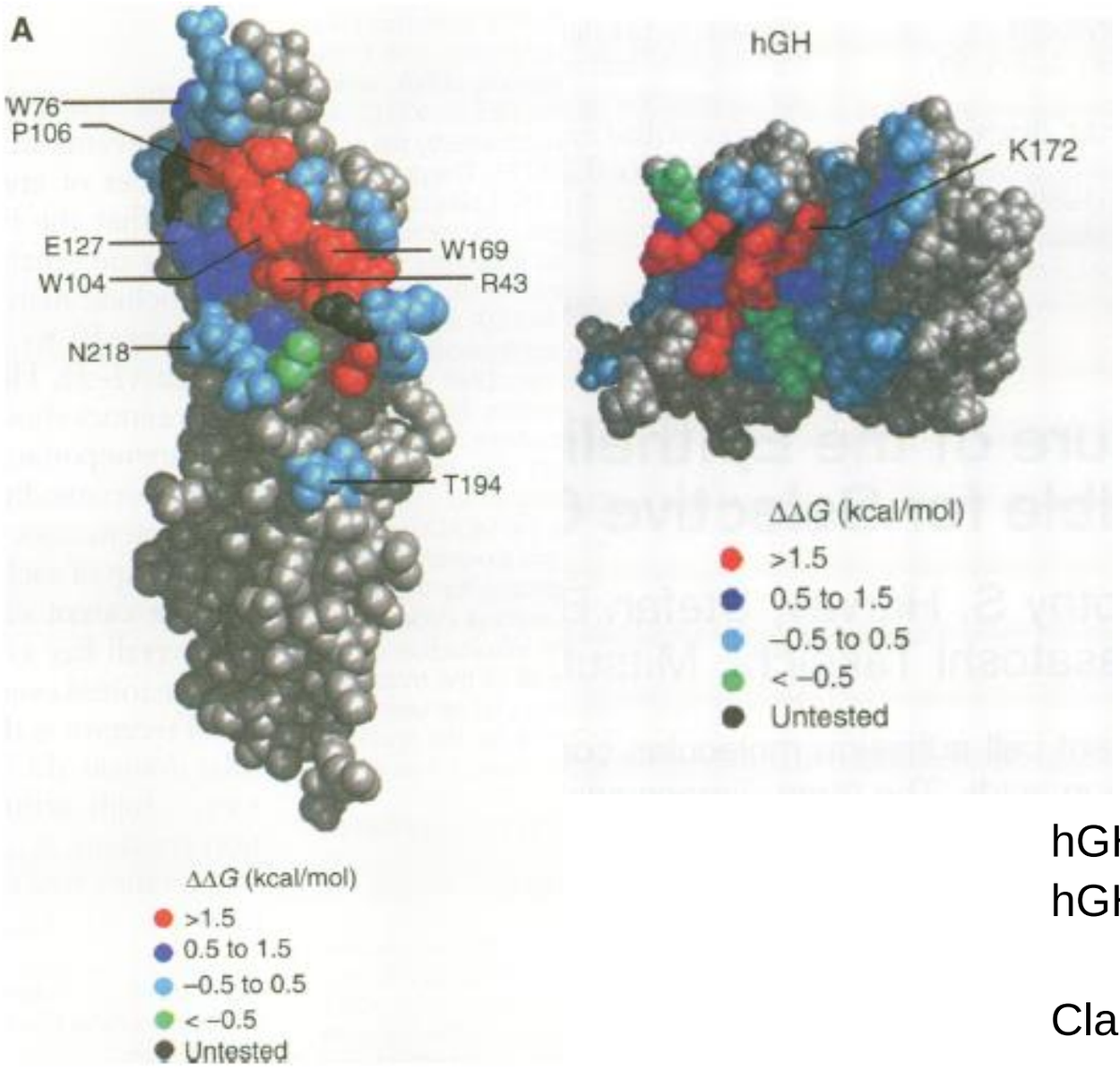


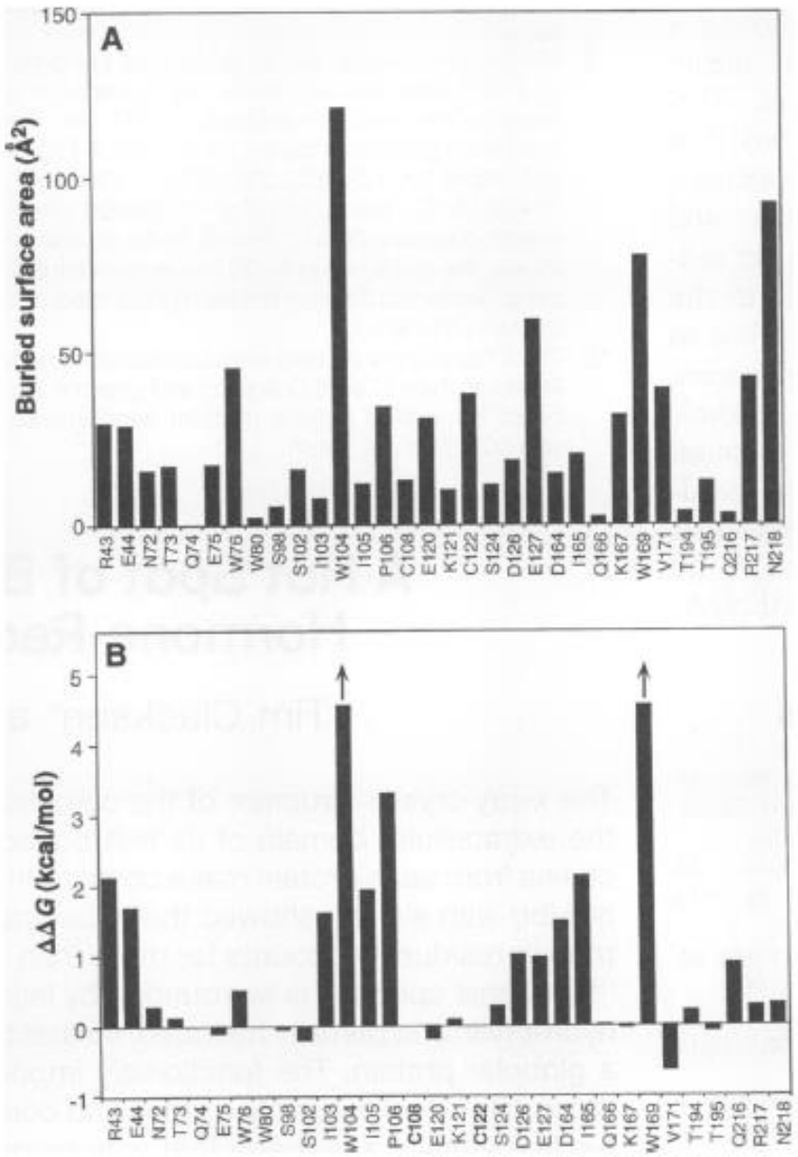
Residue propensities for core and rim regions at interfaces of protein–protein complexes. Drawn after Janin *et al.* (2008).

(Left) Residues in the center (“core”) of the roughly spherical interface are “responsible” for making tight contact and are thus mostly occluded from solvent.
(Right) the core region is strongly enriched in aromatic residues and depleted in charged residues. The surrounding ring of “rim” residues is much more similar to the remaining protein surface as these residues make partial contact to solvent molecules even in the bound state.

3.2.3 Hot spot residues

Hot spot residues at interfaces:
affinity drops by > 2 kcal/mol when such a residue is mutated to Ala.





hGH: human growth hormone
hGHR: human growth hormone receptor

Clackson, Wells, Science 267, 383 (1995)

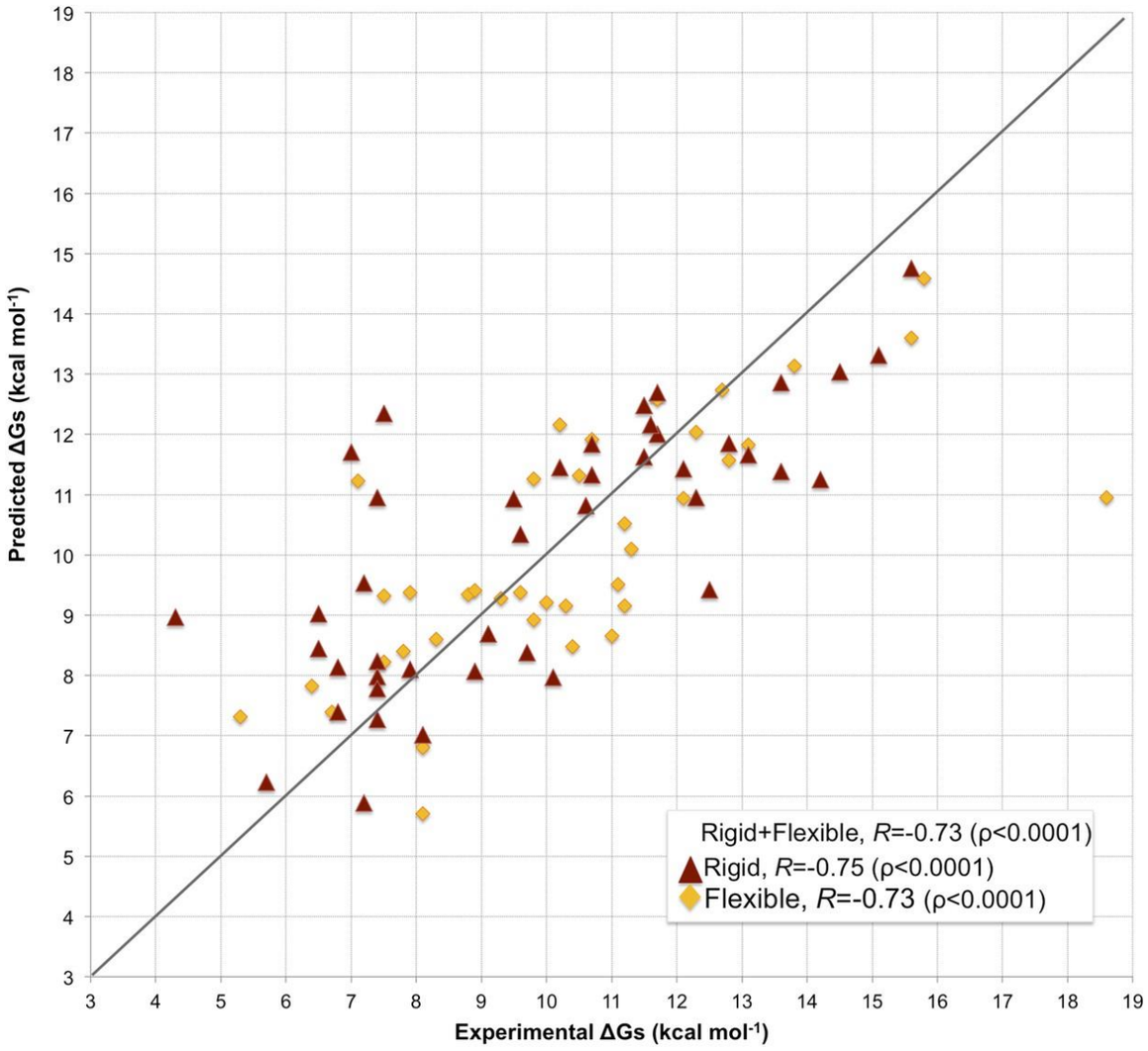
3.2.5 Predicting binding affinities

The total buried SASA has a Pearson correlation of 0.46 with experimental protein binding affinities.

Best available regression model:

$$\Delta G_{calc} = 0.09459 \times ICs_{charged/charged} + 0.10007 \times ICs_{charged\_apolar} - 0.19577 \times ICs_{polar/polar} + 0.22671 \times ICs_{polar/apolar} - 0.18681 \times \%\,NIS_{apolar} - 0.13810 \times \%\,NIS_{charged} + 15.9433\ [kcal/mol]$$

NIS: non-interacting surface
IC: # contacts between residues across the binding interface



Scatter plot of predicted vs experimental binding affinities.
The predictions were made with the above regression model for a dataset of 81 protein–protein complexes. The correlation for all 81 complexes yields an R of −0.73 ($\rho < 0.0001$) with a RMSE of 1.89 kcal mol^{-1}.

rigid cases have iRMSD between superimposed free and bound components ≤1.0 Å
flexible cases have iRMSD >1.0 Å

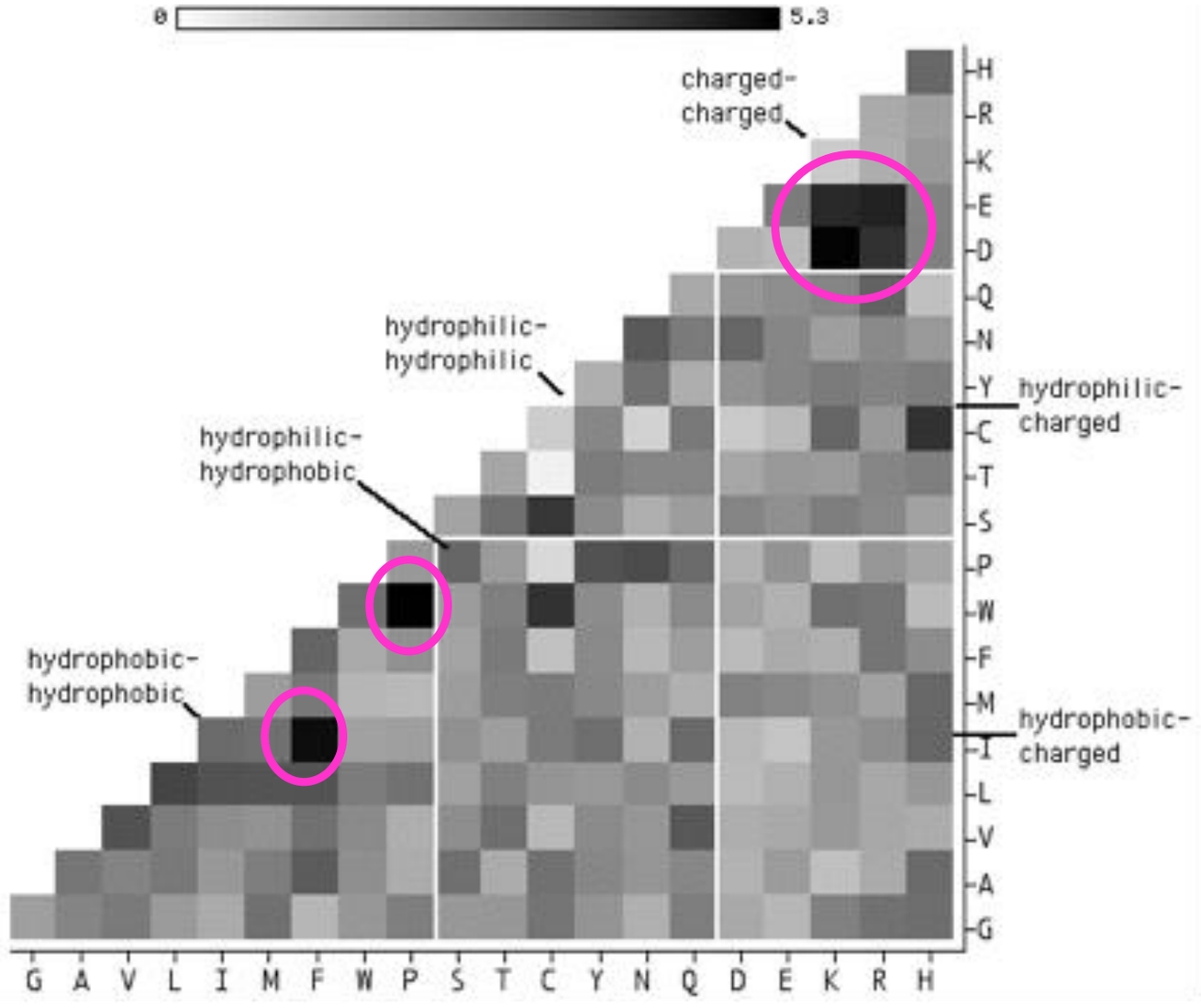
Vangone et al. Elife 4, e07454 (2015)

3.3.1 Pairing propensities

Given the set of interface residues on both proteins, one may analyze what **contacts** each of them forms with residues on the other protein.

A typical **distance threshold** for defining **contacts** is that they have pairs of atoms closer than e.g. 0.5 nm.

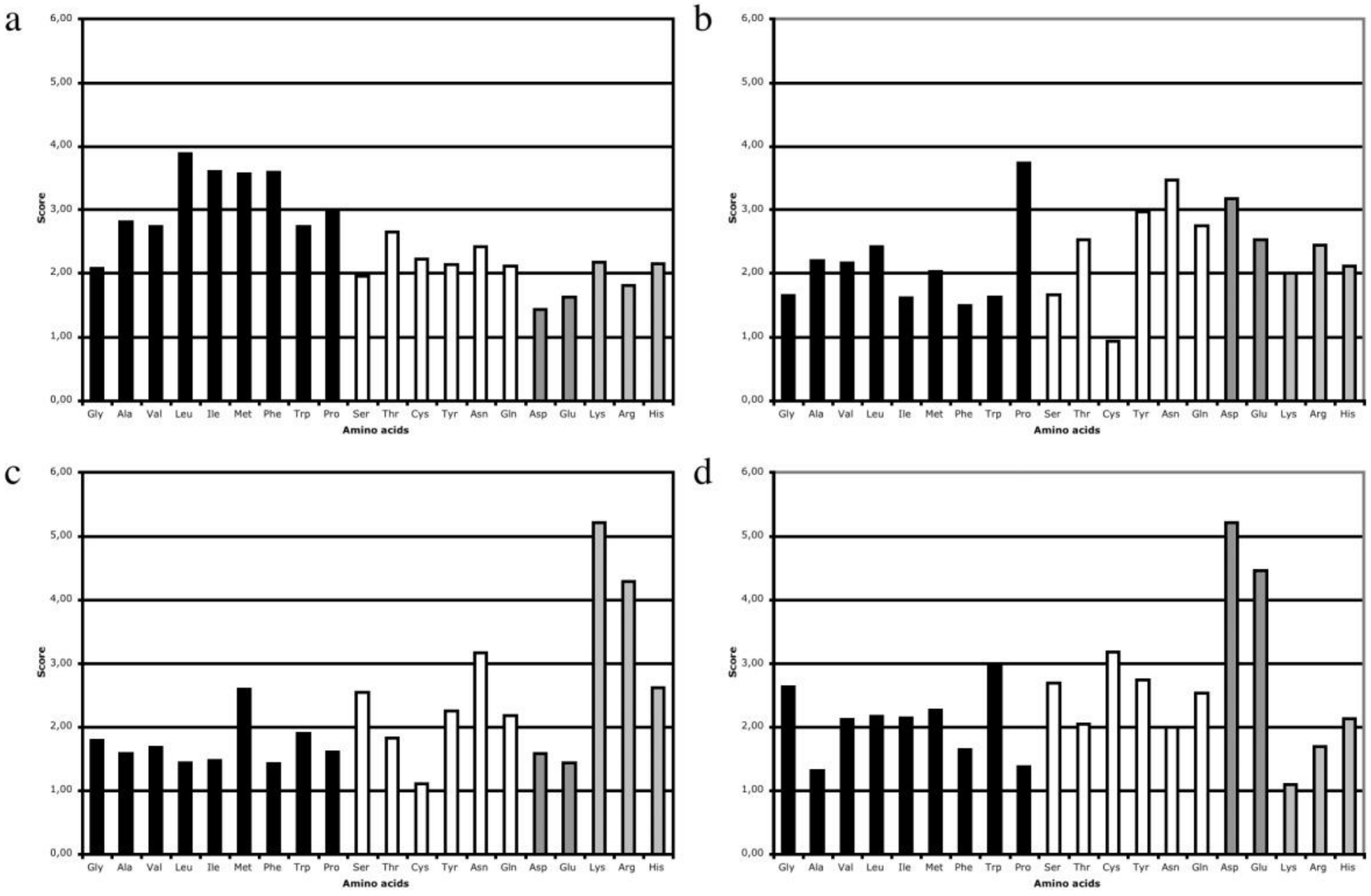
The computed statistics are conveniently represented in a 20 x 20 matrix.



Amino-acid propensity matrix of transient protein-protein interfaces. Scores are normalized pairing frequencies of two residues that occur on the protein-protein interfaces of transient complexes.

Ansari and Helms (2006).

3.3.1 Pairing propensities



Black: hydrophobic residues
White: hydrophilic residues
Grey : charged residues.

Relative occurrence for binding partners of (a) leucine, (b) asparagine, (c) aspartate, and (d) lysine.
The higher the score, the more frequently such pairs occurred in the dataset.

3.3.1 Pairing propensities

From the observed count statistics, one can compute **interfacial pair potentials** *P(i,j) (i = 1…20, j = 1 … 20).*

$$P(i,j) = -\log\left(\frac{N_{obs}(i,j)}{N_{\exp}(i,j)}\right)$$

$N_{obs(i,j)}$: observed number of contacting pairs of *i,j* between two chains,
$N_{exp}(i,j)$: expected number of contacting pairs of *i,j* between two chains.

$N_{exp}(i,j)$ is computed as

$$N_{\exp}(i,j) = X_i \times X_j \times X_{total}$$

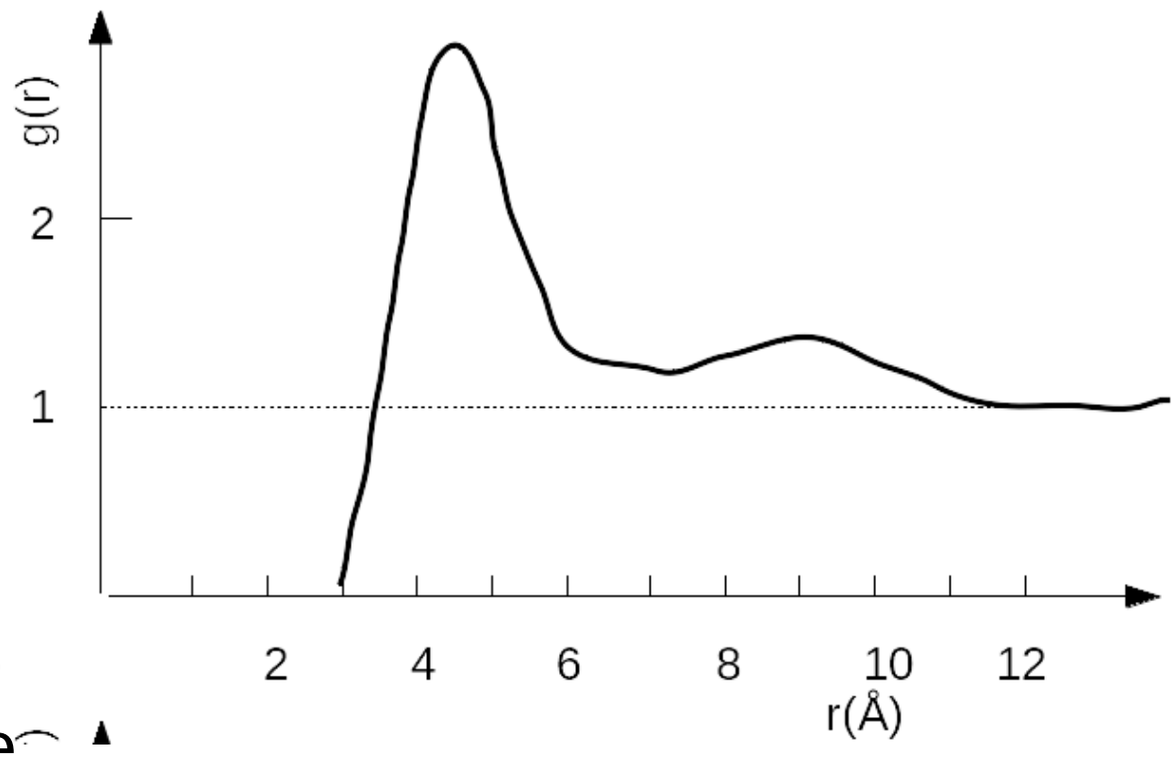
X_i : mole fraction of residue *i* among the total surface residues
X_{total} : total number of contacting pairs.

P(i,j) < 0 : observed frequency higher than expected
P(i,j) > 0 : less

3.3.2 Pair distribution function

A **radial pair distribution function** counts all pairs of amino acids at varying distance.
This distribution is then normalized with respect to an ideal gas, where particle distances are completely uncorrelated.

right: Pair distribution function of finding two alanine residues at a given distance in a protein.



Hydrophobic Ala amino acids are mostly found in the hydrophobic core of proteins. Thus, we expect to find more Ala-Ala pairs at relatively short distances than at distances spanning from one side of the protein to the other one.

3.3.2 amino acid statistical potentials

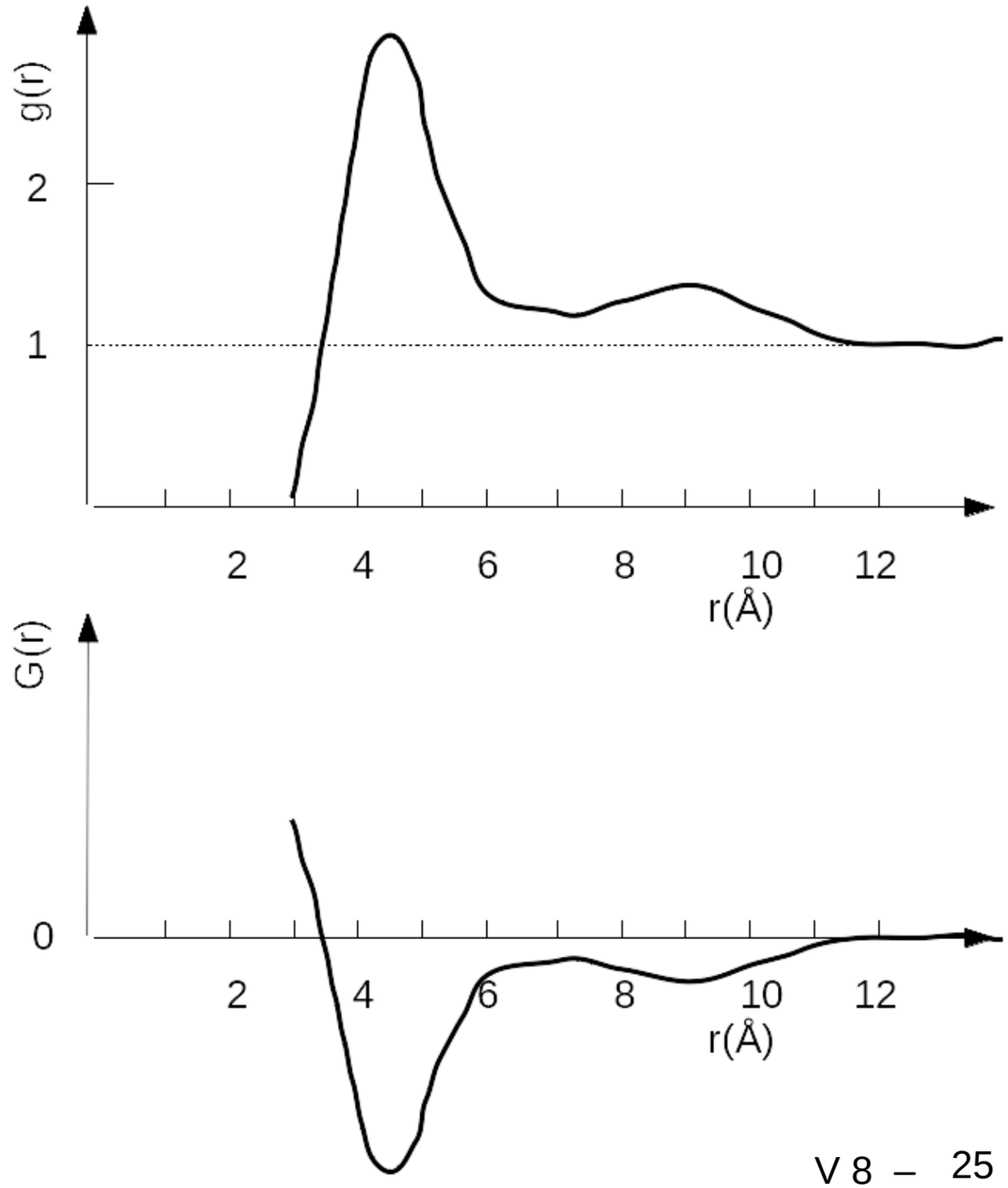
According to the **Boltzmann distribution**, the occupancy levels p_1 and p_2 of two states 1 and 2 of a system with according energies E_1 and E_2 will vary according to the exponentially weighted energy difference between them:

$$\frac{p_1}{p_2} = e^{-\frac{E_1 - E_2}{kT}}$$

If we invert this formula ("Boltzmann inversion"), we can deduce an effective (free) energy function *G(r)* for the interaction between pairs of amino acids from these radial distribution functions *p(r)*,

$$G(r) = -k_B T \ln p(r)$$

These effective potentials can be used to score candidate conformations.



3.3.3 Conservation at interfaces

Functional constraints are expected to limit the amino acid substitution rates in proteins, resulting in a **higher conservation** of **functional sites** such as binding interfaces with respect to the rest of the protein surface.

There exist various approaches for analysing **evolutionary conservation** in MSAs. One of the simplest approaches is the **variance-based method**,

$$C(i) = \sqrt{\sum_j \left(f_j(i) - f_j \right)^2}$$

$C(i)$: conservation index for sequence position i in MSA,
f_j : overall frequency of amino acid j in the alignment
$f_j(i)$: frequency of amino acid j at sequence position i.

Positions with $f_j(i)$ equal to f_j for all amino acids j are assigned $C(i) = 0$.

On the contrary, $C(i)$ takes on its maximum for the position occupied by an invariant amino acid whose overall frequency in the alignment is low.

3.3.3 Conservation at interfaces

Another way of measuring conservation is based on the **entropy** of characters at position *i*,

$$C(i) = \sum_{j=1}^{20} f_j(i) \ln f_j(i)$$

This expression takes on its maximal value for *C(i)* (with the highest entropy) when all amino acids appear with the same frequency 1/20 in position *i*.

If the position is fully conserved, so that *f(X)* = 1 for one particular amino acid X and 0 otherwise, the entropy takes on its lowest possible value.

The **rate4site** algorithm (Mayrose et al. 2004) detects conserved amino acid sites in a multiple sequence alignment (MSA) given as input.

First, the algorithms generates a phylogenetic tree that matches the available MSA (or a pre-calculated tree provided by the user). Then, the algorithm computes a relative measure of conservation for each position in the MSA.

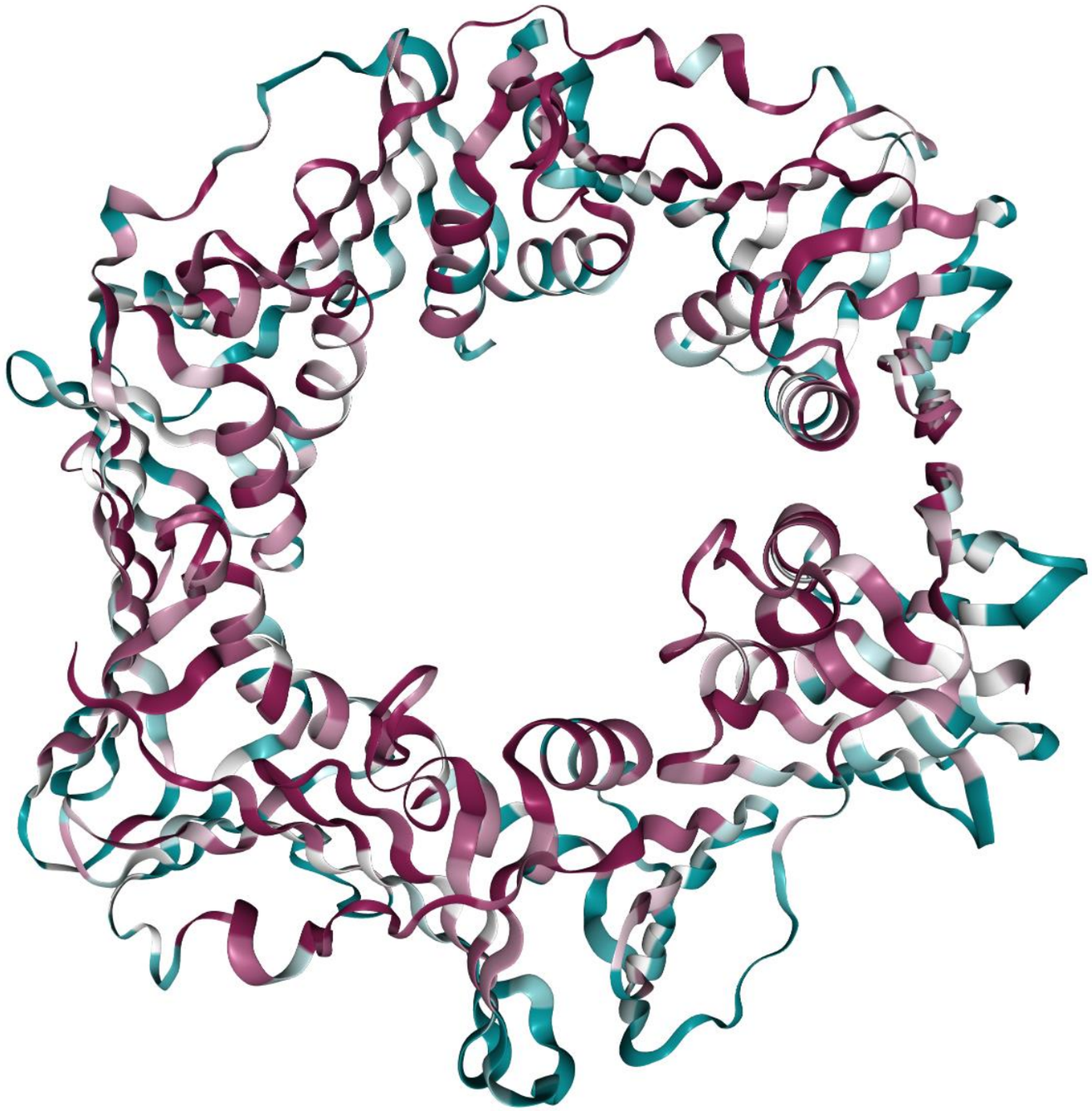
3.3.3 Visualize conservation: Consurf

The popular online-tool Consurf visualizes conservation scores computed with rate4site on 3D protein structure.

The results are color-coded by the degree of evolutionary conservation.

Red : strongly conserved,
blue : weakly conserved.

As anticipated, most of the residues at the inter-subunit interfaces are highly evolutionarily conserved.



Conservation of surface residues at the dimer interface of the homo dimer of the β subunit of DNA polymerase III from *Escherichia coli* (Ashkenazy *et al.* 2016).