Macromolecular complexes

and interactions

Outline

❑ Macromolecular complexes

❑ Structure of complexes

❑ Prediction of 3D structures of complexes

❑ Analysis of macromolecular complexes

Macromolecular complexes

- **Structural**: many proteins are formed by two or more polypeptide chains interacting with each other

- **Function – molecular recognition**: protein-protein and protein-nucleic acid interactions have central importance for virtually every process in a living cell

  - regulation

  - transport

  - signal transduction

  - genetic activity (transcription, translation, replication, repair, …)

  - …

Macromolecular complexes

❑ **Types of complexes**

- protein – small molecule ☑

- protein – protein

- protein – nucleic acids

- nucleic acids – small molecule

Protein-protein complexes

❑ **Oligomerization**

- ▪ native interactions between proteins in native conditions
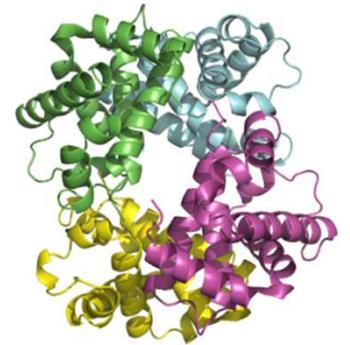
❑ **Aggregation**

- ▪ interactions between native proteins at extreme conditions

- ▪ interactions between misfolded/partially folded proteins -> disease
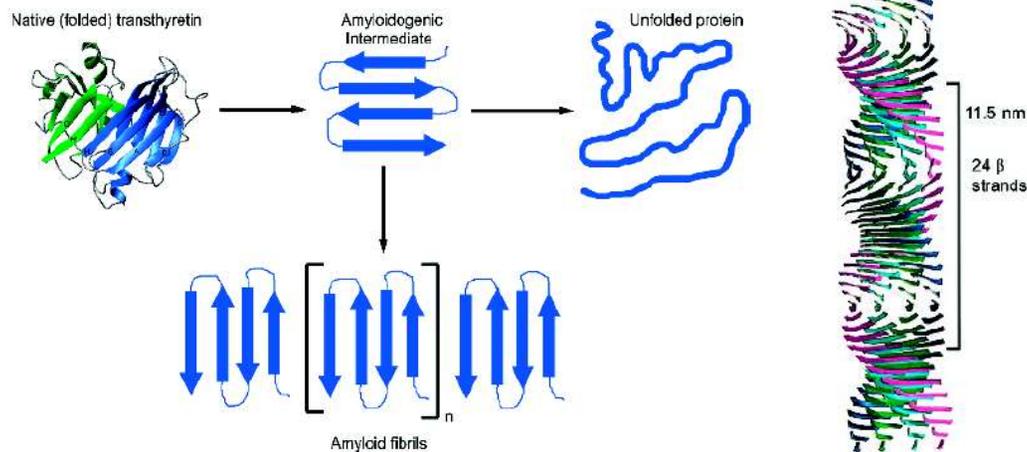
Protein oligomerization

- ❑ Oligomerization is common

  - ▪ more than 35 % of proteins in a cell are oligomers

  - ▪ average oligomeric state of cellular proteins ->

    -> tetramer

  - ▪ homo-oligomers – the most common oligomeric state

  - ▪ some proteins exists solely in the oligomeric state

- ❑ Oligomerization is favored by evolution

- ❑ Oligomerization interfaces are complementary

- ❑ Oligomers are often symmetric

homodimer: a2

heterodimer: ab

heterotetramer: a2b2
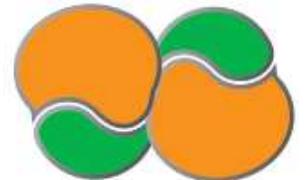
heteropentamer a2bcd

Advantages of oligomerization

- **Morphological function**

    - more complex structure often required for many functions

- **Cooperative function**

    - allostery

    - multivalent binding

- **Enhanced stability**

    - smaller surface

    - more interactions

Oligomerization interface

❑ **Characteristics of oligomeric interface**

- buried surface area > 1400 $Å^2$

- tendency to circular shape

- residues protrude from the surface

- more non-polar residues (about 2/3) than in other parts of surface

- more polar residues (about 1/5) than in protein cores

- about 1 H-bond per 200 $Å^2$

❑ **Hot-spot residues**

- responsible for most of the oligomeric interactions

- often evolutionary conserved, polar residues

- frequently located about the center of the interface

Protein-nucleic acids complexes

❑ **Protein-nucleic acid interactions**

- non-specific – electrostatic interactions with negative charge on the backbone of nucleic acid -> Lys and Arg residues

- specific – recognition of particular nucleic acid sequence

  - major groove – B-DNA

  - minor groove – A-DNA or A-RNA

  - single strand RNA

❑ **Typical interfaces/motifs**

- DNA binding proteins

- RNA binding proteins

Protein-nucleic acids complexes

❑ **DNA binding proteins**

- helix-turn-helix
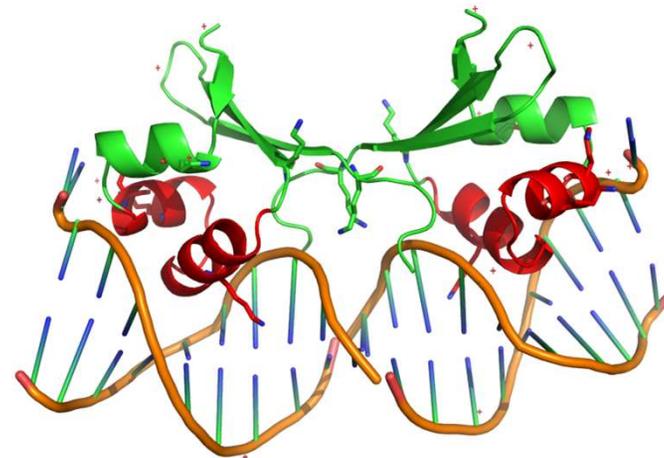
- zinc finger
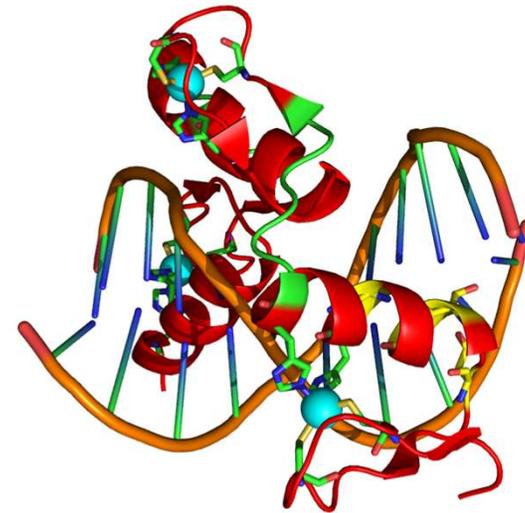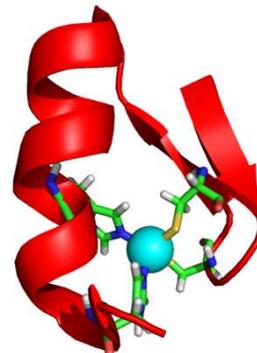
Protein-nucleic acids complexes

❑ **RNA binding proteins**

- recognition is often also governed by the particular structure of RNA

- many employed motifs
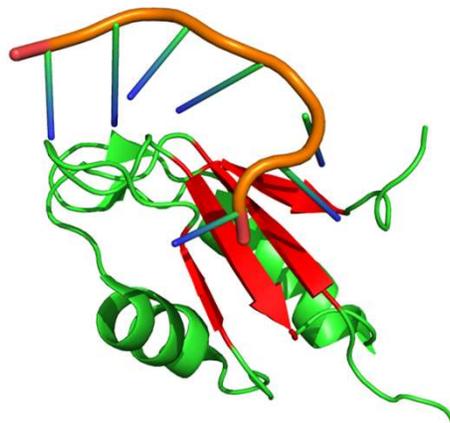
RNA recognition motif   **K-homology domain**   **Pumilio repeat domain**
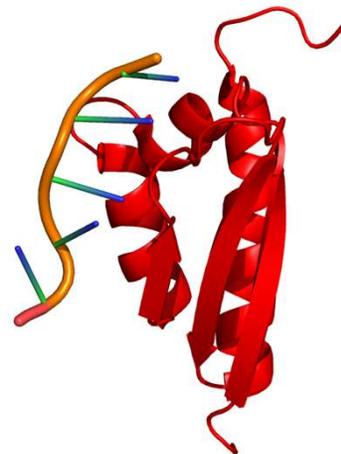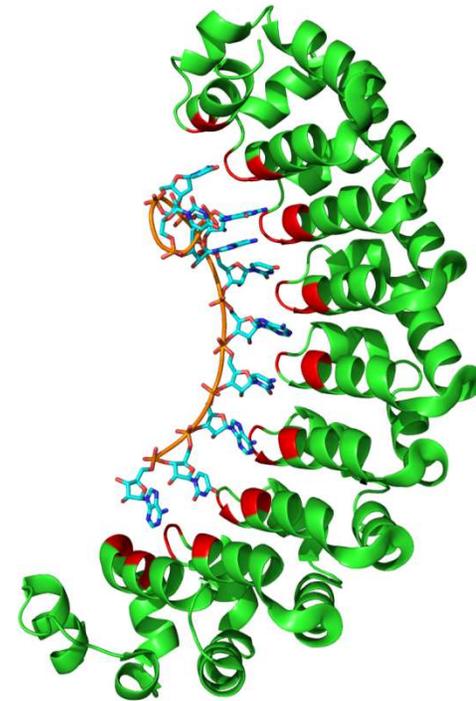
Structure of complexes

- Quaternary structure in PDB database

- Complex or artifact?

Quaternary structure in PDB database

❑ **Asymmetric unit (ASU)**

- macromolecular structures from X-ray crystallography deposited to PDB as a single asymmetric unit

- the smallest portion of a crystal structure to which symmetry operations can be applied in order to generate the unit cell

❑ **Unit cell (crystal unit)**

- the smallest portion of a crystal that, when duplicated and translated, can generate the entire crystal

Quaternary structure in PDB database

Asymmetric Unit → Unit Cell → Entire Crystal
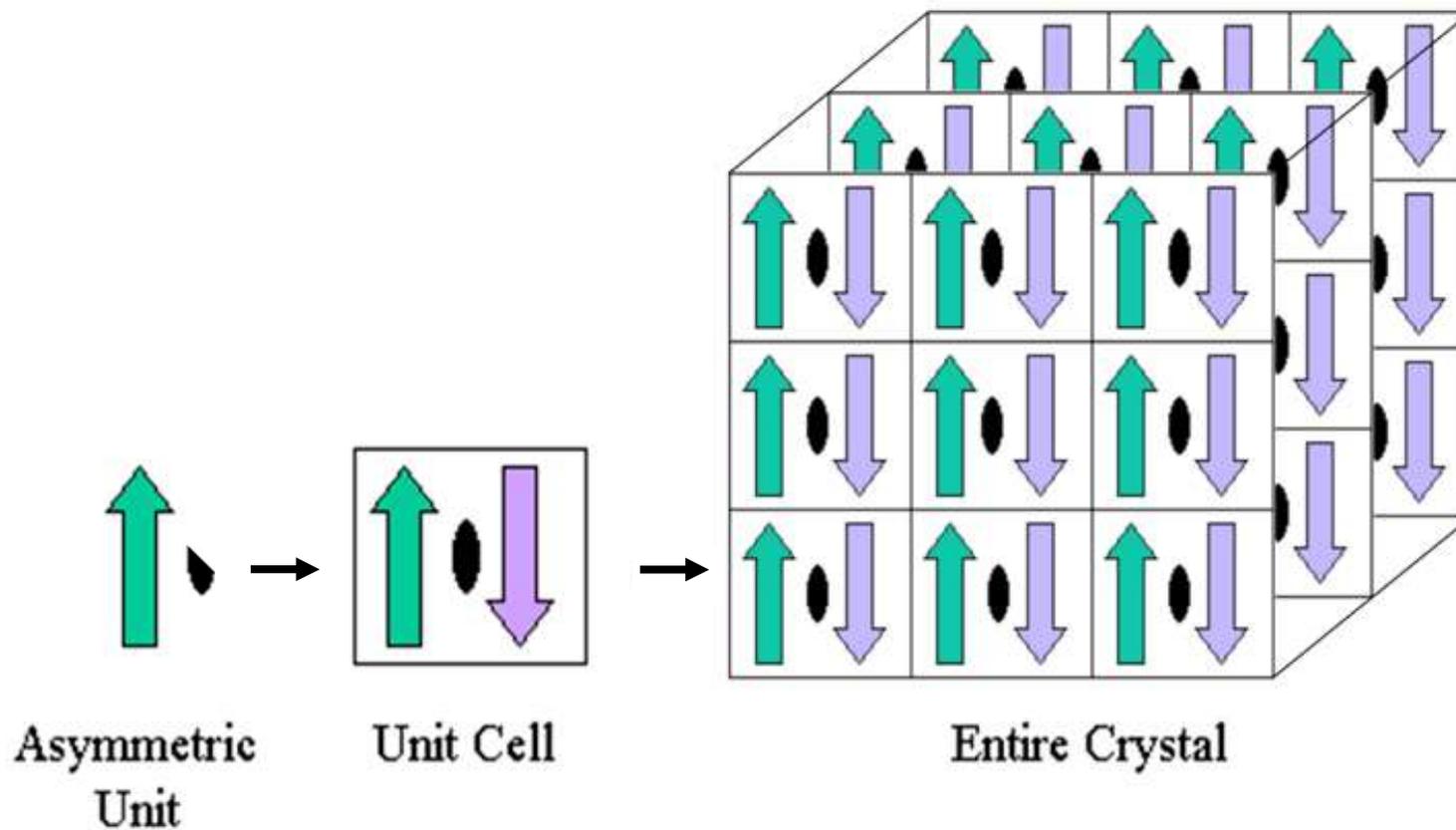
Crystalline environment

- ❑ **Crystal contacts**

  - ▪ intermolecular contacts solely due to protein crystallization

    - ▪ causes artifacts of crystallization

   - ▪ crystal packing - complicates identification of native quaternary structure

Crystal U

ASU
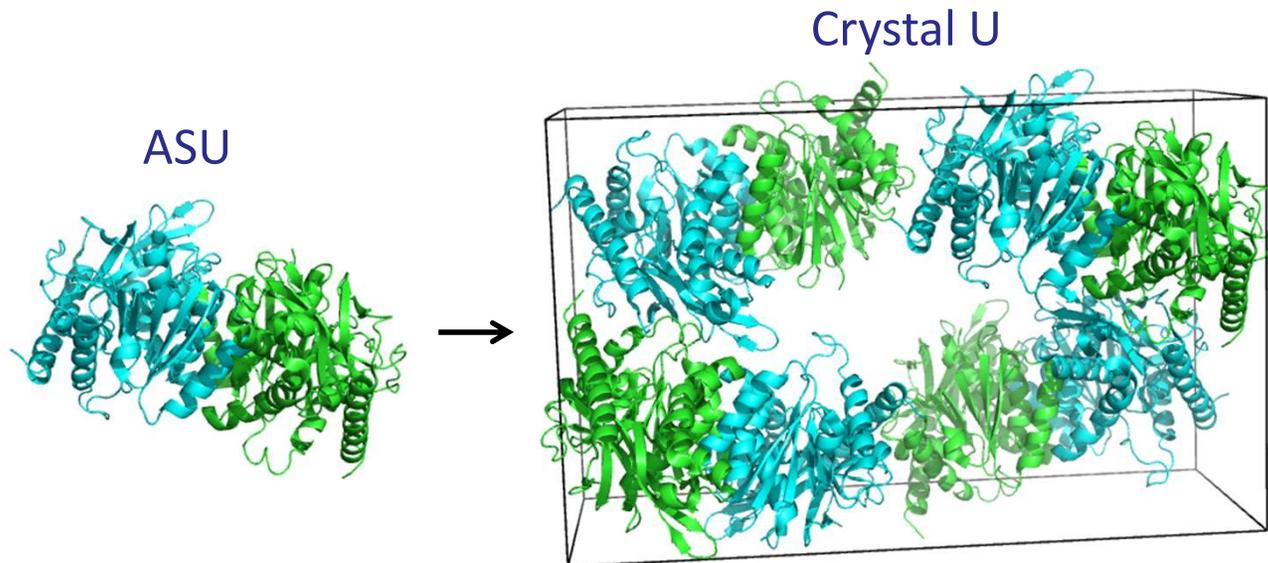
Crystalline environment

- **Artifacts of crystallization**

  - concerns conformation of some surface regions

  - often loops or side chains

 - can complicate the evaluation of effects of mutations on structure
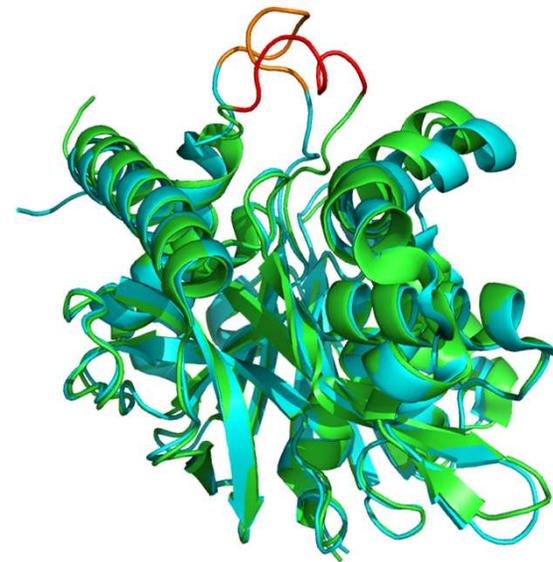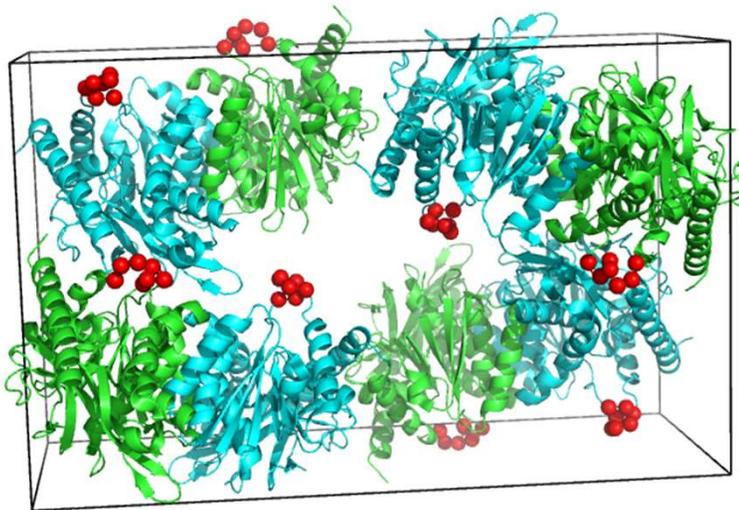
Quaternary structure in PDB database

□ **Biological unit**

- also: functional unit, biological assembly, quaternary structure

- the functional form of a protein

- depends on the environment, post-translational modifications of proteins and their mutations
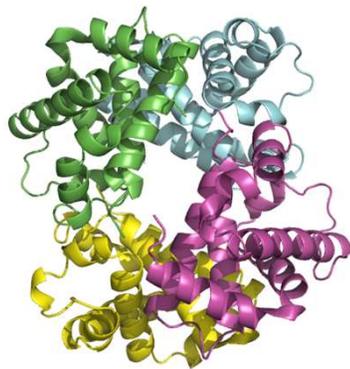

homotetramer
hemoglobin

Biological *versus* asymmetric unit

- **Biological unit can be formed from**

    - multiple copies of the ASU

    - one copy of the ASU

    - a portion of the ASU

ASU   Biol. U
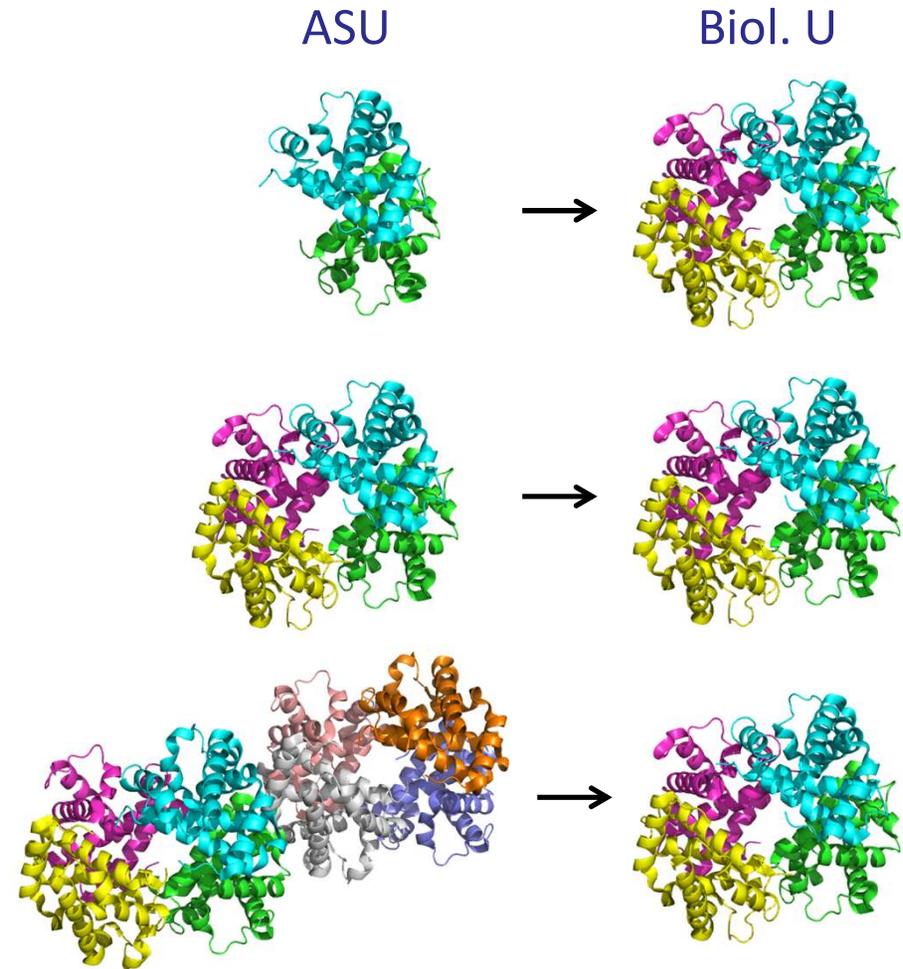
Biological *versus* asymmetric unit

❑ **Large assemblies**

▪ viral capsid

ASU   Biol. U
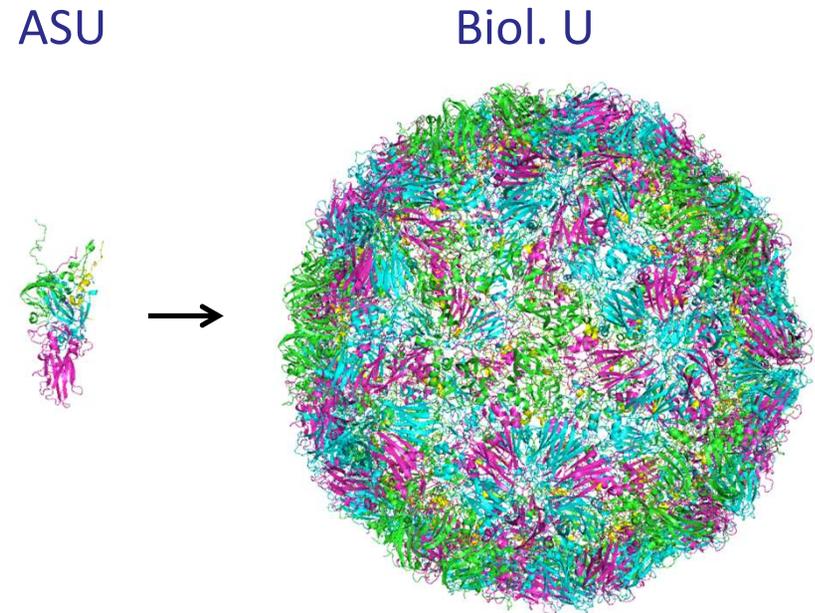
▪ filamentous bacteriophage PF1

ASU   Biol. U
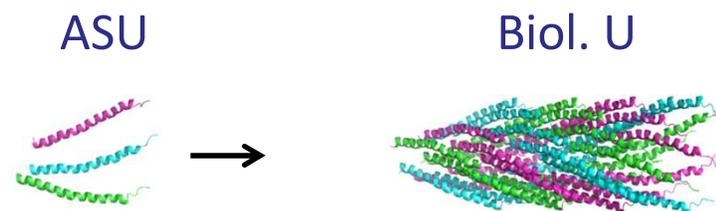
Complex or artifact?

❑ **Problem**

- most proteins in the PDB have three or more crystal contacts that sum up to 30% of the protein solvent accessible surface area

- how to recognize biologically relevant contacts from crystal one?
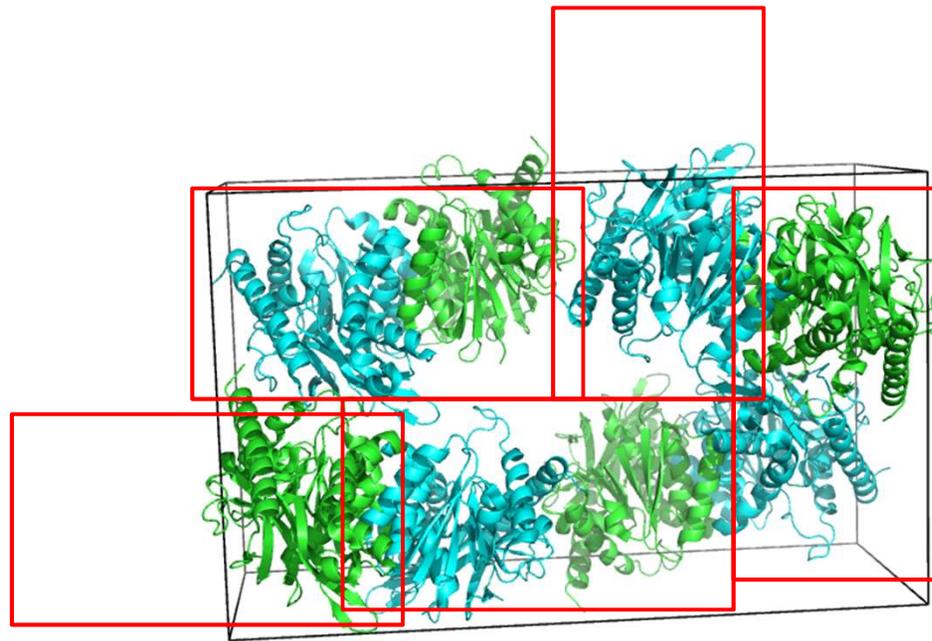
Complex or artifact?

- **Experimental knowledge of oligomeric state helps with identifying of the structure of native complex**

  - search literature

  - experimental methods

    - gel filtration, static or dynamic light scattering, analytical ultracentrifugation, native electrophoresis, …

- **How to get the structure of a biological unit?**

  - author-specified assembly

  - databases

  - predictive tools

Author-specified assembly

❑ **REMARK 350 in headers of PDB file**

- contains symmetry operations to reconstruct biological unit

  - sometimes the specific oligomers were not known at the time the ASU was published

  - some authors may have failed to specify the biological unit even when it was known

  - rarely the specified biological unit might be incorrect

- -> verify author-proposed biological unit also by other means

❑ **Employed by**

- MakeMultimer server

- RCSB PDB

Author-specified assembly

❑ **MakeMultimer server**

- ▪ http://watcut.uwaterloo.ca/makemultimer/

- ▪ generates a PDB file in which all protein chains are as separate files

- ▪ problems with non-crystallographic symmetries

Author-specified assembly

❑ **RCSB PDB**

- generates a PDB file in which all protein chains are as separate models -> complicates visualization and analysis

Prediction of 3D structure of complexes

- homology-based methods

- macromolecular docking

Homology based methods

- ❑ the model of a protein complex is built based on a similar protein complex with a known 3D structure

- ❑ assumption that the interaction information can be extrapolated from one complex structure to close homologs of interacting proteins

  - ▪ close homologs (≥ 40% sequence identity) almost always interact in the same way (if they interact with the same partner)

  - ▪ similarity only in fold was found to be only rarely associated with a similarity in interaction

- ❑ limited applicability (low number of templates)

Homology based methods

- ❑ HOMCOS (**Ho**mology **M**odeling of **Co**mplex **S**tructure)

  - ▪ http://strcomp.protein.osaka-u.ac.jp/homcos/

  - ▪ predicts 3D structure of heterodimers and homodimers by homology modeling

  - ▪ optionally, identifies potentially interacting proteins for user-provided sequence

  1. BLAST search to identify homologous templates in the latest representative dataset of heterodimer (homodimer) structures

  2. evaluation of the model validity by the combination of sequence similarity and knowledge-based contact potential energy

  3. generation of a script for building full atomic model by MODELLER

Prediction of 3D structure of complexes – homology based methods

Homology based methods

Homology based methods

Prediction of 3D structure of complexes – homology based methods

Macromolecular docking

- prediction of the best bound state for given 3D structures of two or more macromolecules

- difficult task
    - large search space - many potential ways in which macromolecules can interact
    - flexibility of the macromolecular surface and conformational changes upon binding

- can be facilitated by prior knowledge
    - e.g., known binding site → significant restriction of the search space
    - distance constraints on some residues

Macromolecular docking

- macromolecule representation

- searching

- scoring

Macromolecule representation

❑ representation of the macromolecular surface (applicable to both receptor and ligand)

- ▪ geometrical shape descriptors (set of spheres, surface normals, vectors radiating from the center of the molecule,…)

- ▪ Discretization of space: grid representation
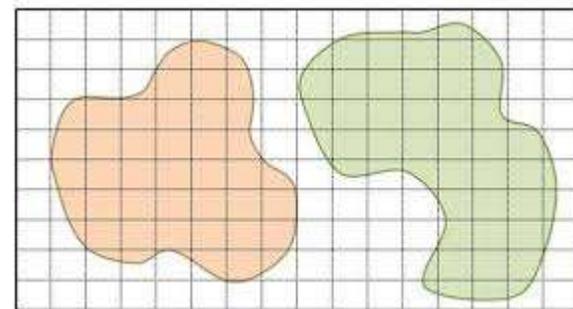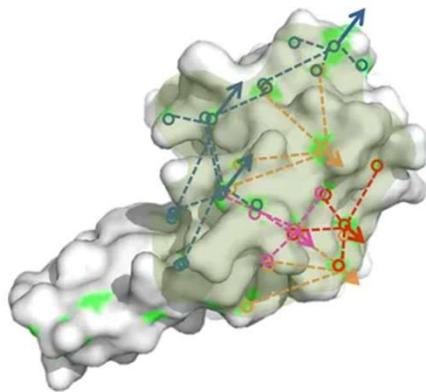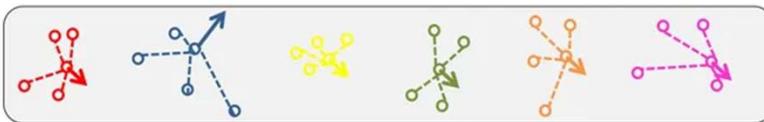
Macromolecule representation

□ macromolecule flexibility

- fully rigid approximation

- soft docking – employs tolerant "soft" scoring functions to simulate plasticity of otherwise rigid molecule

- explicit side-chain flexibility – optimization of residues by rotating part of their structure or rotation of whole side-chains using predefined rotamer libraries

- docking to molecular ensemble of protein structure – composed from multiple crystal structures, from NMR structure determination or from trajectory produced by MD simulation

Macromolecule representation

❑ macromolecule flexibility

- **rigid body** docking – basic model that considers the two macromolecules as two rigid solid bodies

- **semiflexible** docking – one of the molecules (typically the smaller one) is the only one considered flexible

- **flexible** docking – both molecules are considered flexible

Macromolecular docking - searching

- generally based on the idea of complementarity between interacting molecules (geometric, electrostatic or hydrophobic)

- the main problem is the dimension of the conformational space to be explored:
  - rigid docking: 6D (hard)
  - flexible docking: 6D + N_{fb} (impossible!)

- information on the rough location of the binding surface (experimental or predicted) → reduction of the search space

Macromolecular docking - searching

- ❑ **exhaustive search**

  - ▪ full search of the conformational space, i.e., trying every possible relative orientation of the two molecules

  - ▪ computationally very expensive – 6 degrees of freedom for rigid molecules (3 translations + 3 rotations)

  - ▪ grid approaches
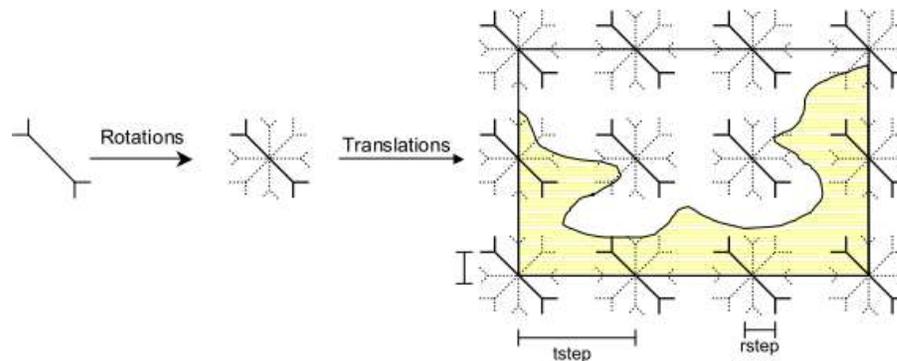
Macromolecular docking - searching

❑ **stochastic methods**

- Monte Carlo

- genetic algorithms

- …

Macromolecular docking - scoring

- ❑ scoring function

    - ▪ evaluation of a large number of putative solutions generated by searching algorithms

- ❑ methods often use a two-stage ranking

    1. approximate and fast-to-compute function – used to eliminate very unlikely solutions

    2. more accurate function – used to select the best among the remaining solutions

Macromolecular docking - scoring

- ❑ scoring functions

  - ▪ knowledge-based

  - ▪ empirical

  - ▪ force field-based

  - ▪ clustering-based – the presence of many similar solutions is taken as an indication of correctness (all solutions are clustered and the size of each cluster is used as on of the scoring parameters)

Macromolecular docking - scoring

❑ a combination of several parameters

- ▪ low free energy or pseudo-energy based on force field

- ▪ large buried surface area

- ▪ good geometric complementarity

- ▪ good H-bonding

- ▪ good charge complementarity

- ▪ polar/polar contacts favored

- ▪ polar/non-polar contacts disfavored

- ▪ many similar solutions (large clusters)

- ▪ ...

Macromolecular docking - programs

Web server/software and link	Docking method	Filtering and refinement
BDOCK [152] http://www.biotec.tudresden.de/~bhuang/bdock/bdock.html	FFT correlation based on shape complementarity, degree of burial and conservation	Altering the docking solutions with a scoring function
ClusPro [110] http://nrc.bu.edu/cluster/	FFT correlation using DOT [109]	Filtering with empirical potential and clustering, refinement by SmoothDock [111]
DOT [109] http://www.sdsc.edu/CCMS/DOT/	FFT correlation based on electrostatics and shape complementarity	Refinement by energy minimization
FireDock [153] http://bioinfo3d.cs.tau.ac.il/FireDock/	None (refinement server)	Refinement using an energy function
GRAMMX [108] http://vakser.bioinformatics.ku.edu/resources/gramm/grammx	FFT correlation based on shape complementarity, hydrophobicity and smoothed potentials	Clustering and knowledge-based scoring
HADDOCK [154] http://www.nmr.chem.uu.nl/haddock/	Data-driven docking approach based on biochemical and/or biophysical interaction data	None
HEX [155] http://www.csd.abdn.ac.uk/hex/	Spherical polar Fourier correlations	None
MolFit [156] http://www.weizmann.ac.il/ChemicalResearchSupport//molfit/home.html	FFT correlation based on chemical and shape complementarity	Clustering of the predicted conformations
PatchDock [114] http://bioinfo3d.cs.tau.ac.il/PatchDock/	Geometric hashing and pose-clustering	Ranking according to a geometric shape complementarity score
PyDock [157] http://mmb.pcb.ub.es/PyDock/	FFT based on electrostatics and desolvation energy	Ranking using an energy function
RosettaDock [115] http://rosettadock.graylab.jhu.edu/	Local docking by Monte Carlo search	Ranking using an energy function, clustering
ZDOCK [107] http://zlab.bu.edu/zdock/index.shtml	FFT correlation based on shape complementarity, desolvation energy and electrostatics	Refinement by energy minimization
3D-Dock [158] http://www.sbg.bio.ic.ac.uk/docking/	FFT correlation using FTDOCK [159]	Clustering, refinement of side-chains using Multidock [159]

Prediction of 3D structure of complexes – macromolecular docking

Macromolecular docking - programs

❑ Cluspro 2.0

▪ http://cluspro.bu.edu/

▪ performs a global soft rigid-body search using PIPER docking program (employ knowledge-based potential).

▪ the top 1,000 structures are retained and clustered to isolate highly populated low-energy binding modes

▪ a special mode for prediction of molecular assemblies of homo-oligomers

Macromolecular docking - programs

❑ PatchDock

▪ http://bioinfo3d.cs.tau.ac.il/PatchDock/index.html

▪ performs a geometry-based search for docking transformations that yield good molecular shape complementarity (driven by local feature matching rather than brute force searching of the 6D space):

1. the molecular surface is divided into concave, convex and flat patches

2. complementary patches are matched → candidate transformations

3. evaluation of each candidate transformation by a scoring function considering both geometric fit and atomic desolvation energy

4. clustering of the candidate solutions to discard redundant solutions

▪ results can be redirected to FireDock for refinement and re-scoring

Macromolecular docking - programs

❑ PatchDock

(A)

(B)

Convex patch

Concave patch

Macromolecular docking - programs

❑ FireDock

- http://bioinfo3d.cs.tau.ac.il/FireDock/index.html

- refines and re-scores solutions produced by fast rigid-body docking algorithms

- optimizes the binding of each candidate by allowing flexibility in the side-chains and adjustments of the relative orientation of the molecules

- scoring of the refined candidates is based on softened van der Waals interactions, atomic contact energy, electrostatic, and additional binding free energy estimations

Analysis of macromolecular complexes

- ❑ binding energy

- ❑ macromolecular interface

- ❑ interaction hotspots

Binding energy

❑ FastContact

- http://structure.pitt.edu/servers/fastcontact/

- rapidly estimates the electrostatic and desolvation components of the binding free energy between two proteins

- additionally, evaluates the van der Waals interaction using CHARMM and reports contribution of individual residues and pairs of residues to the free energy → highlight the interaction hot spots

```
----- SUMMARY ENERGIES ----------------------
Electrostatic (4r) Energy:   -18.3684946 kcal/mol
Desolvation  Free  Energy:    8.31365025 kcal/mol
van der waals (CHARMm19) :    -1734.5 kcal/mol
---------------------------------------------
Top 20 Min & Max ligand residues contributing to the binding free energy
 -2.628    89 ASN
 -2.586     6 LYS
 -2.209     9 TYR
 -2.135   125 LEU
 -2.114     2 PHE
 -1.832    45 ARG
 -1.684    87 ASN
```

Macromolecular interface

- ❑ the region where two protein chains or protein and nucleic acid chain come into contact

- ❑ can be identified by the analysis of the 3D structure of the macromolecular complex
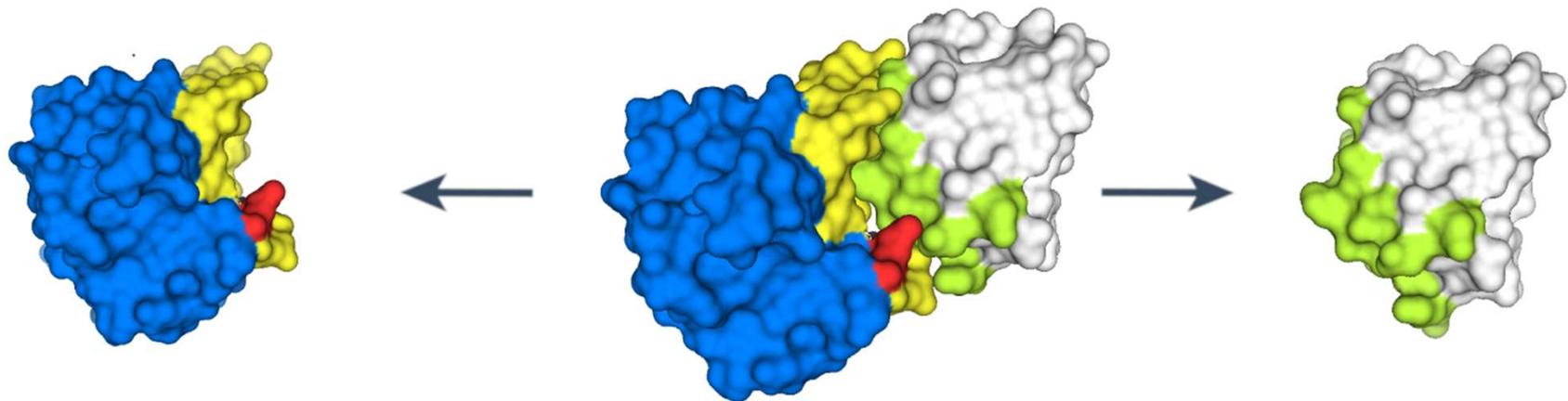
Interface analysis

- provides information about basic features of interacting macromolecular complexes (e.g., shape complementarity, chemical complementarity,...)

- provides information about interface residues

- acquired information is useful for a wide range of applications

  - design of mutants for experimental verification of the interactions

  - development of drugs targeting macromolecular interactions

  - understanding the mechanism of the molecular recognition

  - computational prediction of interfaces and complex 3D structures

  - ...

Interface analysis

- the most common approaches for a definition of interfaces:

  - methods based on the distance between interacting residues

  - methods based on the differences in the solvent accessible surface area (ASA) upon complex formation

  - computational geometry methods (using Voronoi diagrams)

- all three approaches provide very similar results

Interface analysis - tools

- PISA (**P**rotein **I**nterfaces, **S**urfaces and **A**ssemblies)

  - www.pdbe.org/pisa

  - an interactive tool for the exploration of macromolecular (protein, DNA/RNA and ligand) interfaces, prediction of probable quaternary structures, database searches of structurally similar interfaces and assemblies

  - overview and detailed characteristics of all interfaces found within a given structure (including those generated by symmetry operations)

  - provided characteristics including interface area, $\Delta^i G$, potential hydrogen bonds and salt bridges, interface residues and atoms, ...

Interface analysis - tools

❑ MolSurfer

- ▪ http://projects.villa-bosch.de/dbase/molsurfer/index.html

- ▪ visualization of 2D projections of protein-protein and protein-nucleic acid interfaces as maps showing a distribution of interface properties (atomic and residue hydrophobicity, electrostatic potential, surface-surface distances, atomic distances,...)

- ▪ 2D maps are linked with the 3D view of a macromolecular complex

- ▪ facilitates the study of intermolecular interaction properties and steric complementarity between macromolecules

Interface analysis - tools

□ MolSurfer
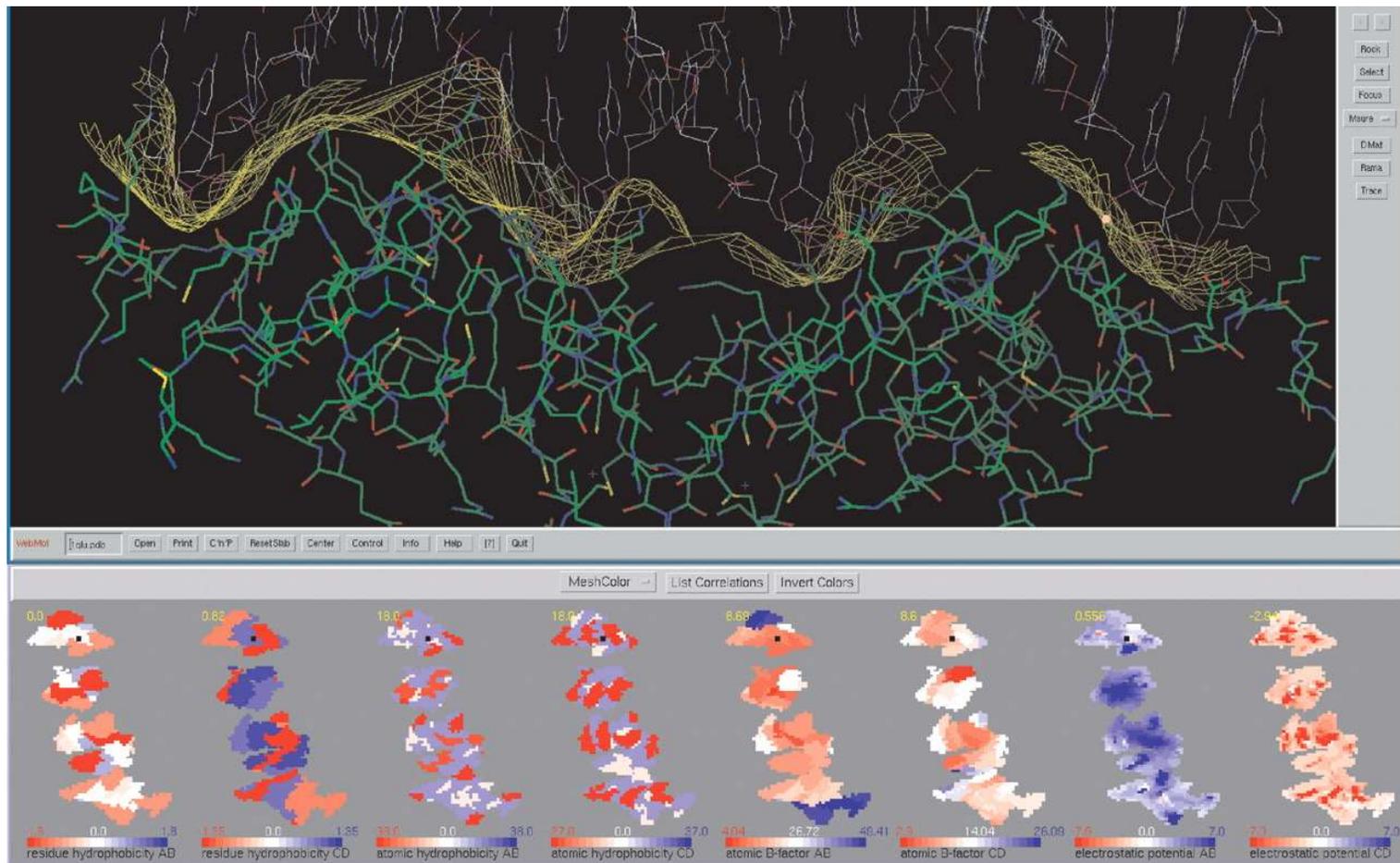
Interface analysis - tools

- ❏ Contact Map WebViewer

    - ▪ http://cmweb.enzim.hu/

    - ▪ represents residue-residue contacts within a protein or between

        proteins in a complex in the form of a contact map

- ❏ PIC (**P**rotein **I**nteraction **C**alculator)

    - ▪ http://pic.mbu.iisc.ernet.in/

    - ▪ identifies various interactions within a protein

        or between proteins in a complex

Interaction hotspots

- **Hot-spots**: residues contributing predominantly to the **binding free energy** of the complex

- knowledge of hot spots has important implications for:
  - understanding the **principles of protein interactions** (an important step in understanding recognition and binding processes)
  - **design of mutants** for experimental verification of the interactions
  - **development of drugs** targeting macromolecular interactions
  - …

Interaction hotspots

- ❑ hot spots are usually conserved and appear to be clustered in tightly packed regions in the center of the interface

- ❑ experimental identification by alanine scanning mutagenesis – if a residue has a significant drop in binding affinity when mutated to alanine it is labeled as a hot spot

- ❑ experimental identification of hot spots is costly and cumbersome → the computational predictions of hot spots can help!

Prediction of hotspots - tools

- most of the available methods are based on the 3D structure of the complex

- knowledge-based methods

  - combination of several physical and chemical features of residues

  - evolutionary conservation, ASA, residue propensity, structural location, hydrophobicity,...)

- energy-based methods

  - calculation of the change in the binding free energy ($\Delta\Delta G_{bind}$) of the complex upon *in silico* modification of a given residue to alanine

Prediction of hotspots - tools

❑ Robetta

- http://old.robetta.org/alascansubmit.jsp

- energy-based method

- performs *in silico* alanine scanning mutagenesis of protein-protein or protein-DNA interface residues

1. the side chain of each interface residue is mutated to alanine

2. all side chains within 5 Å radius sphere of the mutated residue are repacked; the rest of the protein remains unchanged

3. for each mutant, $\Delta\Delta G_{bind}$ is calculated (the residue with the predicted $\Delta\Delta G_{bind} \geq 1$ kcal/mol = hot spot)

Prediction of hotspots - tools

❑ Robetta

```
Tue Nov  6 00:20:55 PST 2012
virtual alanine scanning, Minimized_PfTPR1_23_1.alascan
pdb#    chain    int_id      res#     aa    DDG(complex)   DDG(complex,obs)   DG(partner)
  15       A         1         15      12        0.26             0.00            -0.11
  18       A         1         18       5        1.99             0.00             1.27
  45       A         0         45      16       -0.01             0.00             3.31
  46       A         1         46      12        1.53             0.00            -0.07
  53       A         1         53      16       -0.11             0.00            -0.57
  80       A         1         80      15        2.52             0.00             4.85
  83       A         1         83       2       -0.10             0.00             5.34
  86       A         1         86       7        0.29             0.00             0.34
 124       B         0        124      17       -0.02             0.00             0.60
 125       B         1        125       8        1.75             0.00             0.08
 126       B         1        126       4       -0.23             0.00            -0.41
 127       B         1        127       4        0.02             0.00            -0.60
 128       B         1        128      18        1.98             0.00            -0.45
 129       B         1        129       3       -0.29             0.00            -0.81
```

Analysis of macromolecular complexes – interaction hotspots

References I

❑ Liljas, A. *et al*. (2009). **Textbook Of Structural Biology**, World Scientific Publishing Company, Singapore.

❑ Goodsell, D. S. & Olson, A. J. (2000) Structural symmetry and protein function. *Annual Review of Biophysics and Biomolecular Structure* **29**: 105-153.

❑ Demachenko, A. P. (2001). Recognition between flexible protein molecules: induced and assisted folding. *Journal of Molecular Recognition* **14**: 42-61.

❑ Ali, M. H. & Imperiali, B. (2005) Protein oligomerization: How and why. *Bioorganic & Medicinal Chemistry* **13**: 5013-5020.

❑ Jahn, T. R. & Radford, S. E. (2008) Folding *versus* aggregation: Polypeptide conformations on competing pathways. *Archives of Biochemistry and Biophysics* **469**: 100-117.

❑ Csermely, P. *et al.* (2010) Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends in Biochemical Sciences* **35**: 539-546.

References II

❑ Bujnicki, J. (2009). **Prediction of Protein Structures, Functions, and Interactions**, John Wiley & Sons, Ltd., Chichester, p. 302.

❑ Tramontano, A. (2005). **The Ten Most Wanted Solutions in Protein Bioinformatics**, CRC Press UK, London, p. 186.

❑ Tuncbag, N., *et al.* (2009). A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Briefings in bioinformatics* **10**: 217-232.

❑ Ezkurdia, I., *et al.* (2009). Progress and challenges in predicting protein-protein interaction sites. *Briefings in bioinformatics* **10**: 233-246.

❑ Fernández-Recio, J. (2011). Prediction of protein binding sites and hot spots. *Computational molecular science* **6**: 680-698.

❑ Szilagyi, A., *et al.* (2005). Prediction of physical protein–protein interactions. *Physical biology* **2**: S1-S16.

❑ Moreira, I. S., *et al.* (2010). Protein-protein docking dealing with the unknown. *Journal of computational chemistry* **31**:317-342

References