LOSCHMIDT LABORATORIES

Analysis of protein structures

Outline

- Residue solvent accessibility
- Protein solubility
- Molecular interactions
- Functional sites
  - Binding sites
  - Transport pathways

Residue solvent accessibility

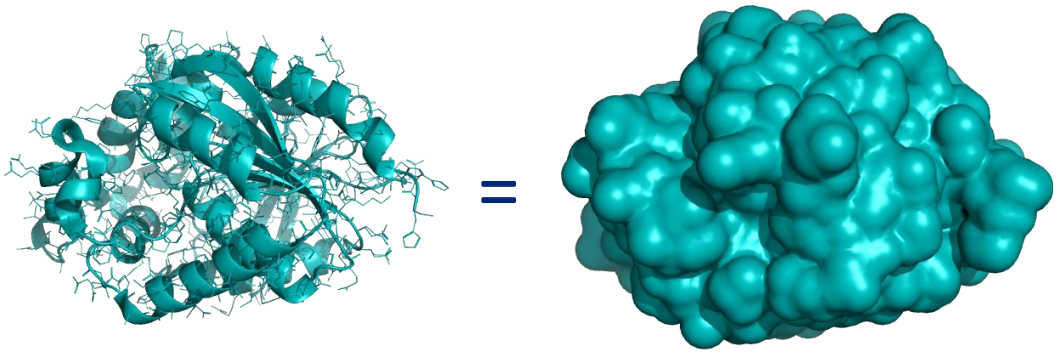
- Solvent accessible surface area

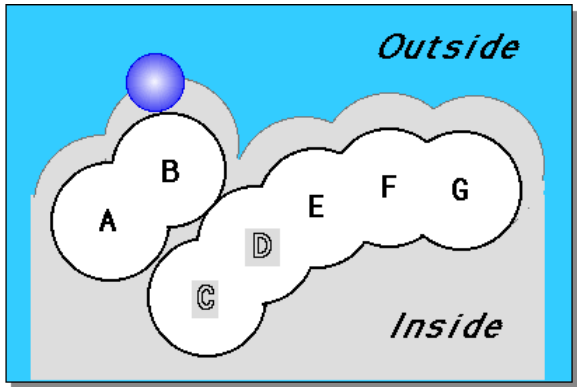
Residue solvent accessibility

- Solvent accessible surface area (ASA, SASA or SAS, in $Å^2$)

  → It quantifies the extent to which a residue in a protein structure is accessible to the solvent

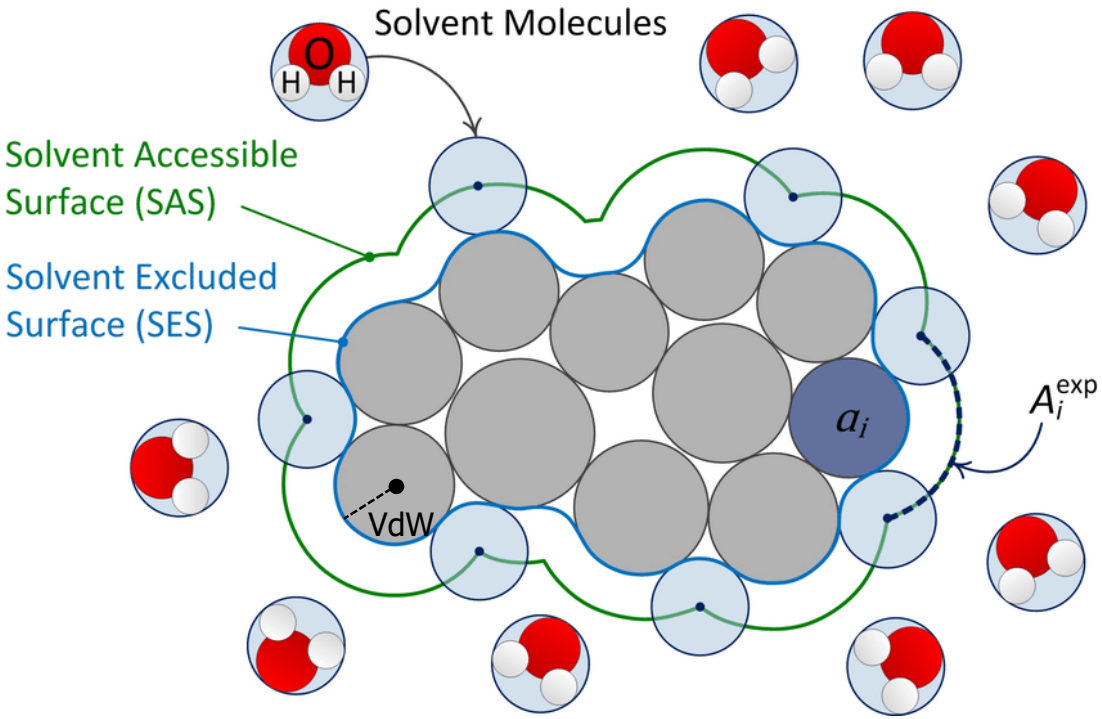
- Typically calculated by rolling a spherical probe of a particular radius over a protein surface and summing the area that can be accessed by this probe on each residue

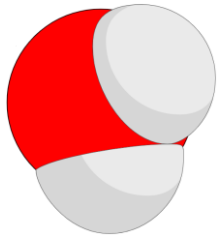




Residue solvent accessibility

- Solvent accessible surface area (ASA, SASA or SAS, in Å^2)
- Solvent excluded surface (SES) – also known as molecular surface, or Connolly surface area



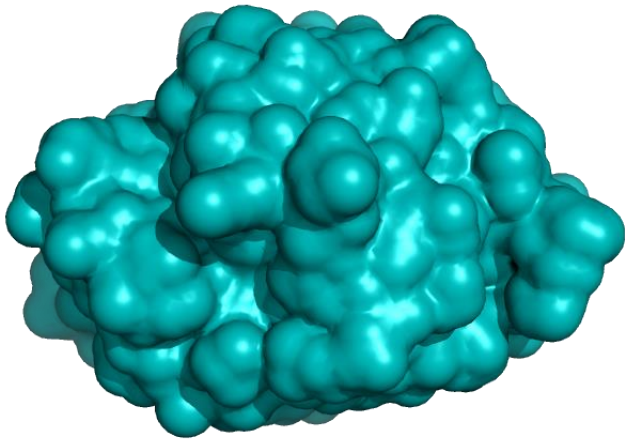


Water radius ≈ 1.4 Å

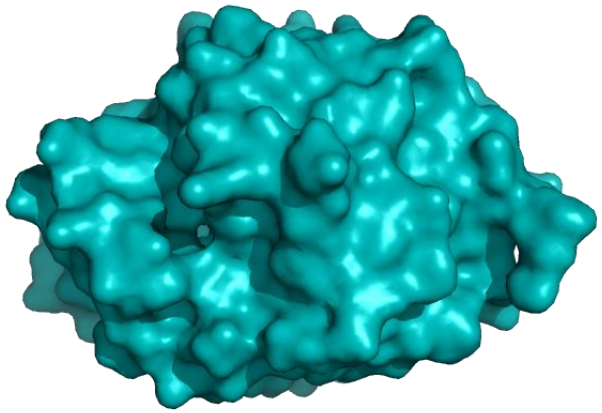
A_i^{exp} = exposed area
VdW = Van der Waals radius

Residue solvent accessibility

- Solvent accessible surface area (ASA, SASA or SAS, in $Å^2$)
- Solvent excluded surface (SES) – also known as molecular surface, or Connolly surface area – usually represented in "surface" visualization



SASA



SES

Residue solvent accessibility

- Relative accessible surface area (rASA)
  - Ratio of the actual accessible area of a given residue

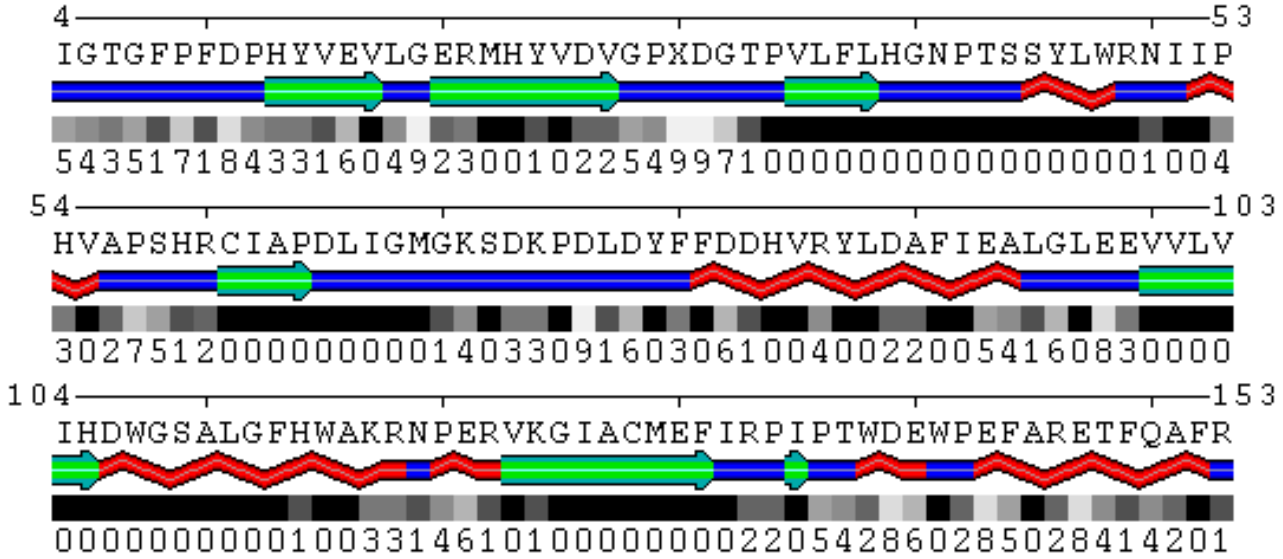
    $$rASA = ASA / ASA_{MAX}$$

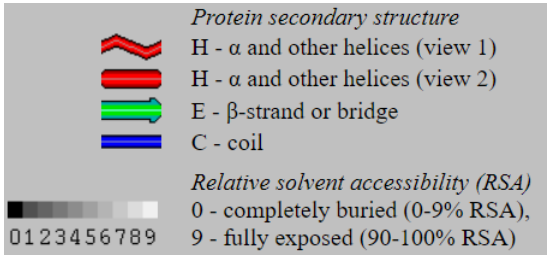
  - Enables comparison of accessibility of different amino acids (e.g., long extended vs. spherical amino acids)
- Simplified two state description
  - Buried vs. exposed residues
  - *Threshold* for differentiating surface residues vs. buried is not well defined (usually rASA = 15–25 %)
  - rASA < *threshold* => buried

    rASA ≥ *threshold* => exposed

Residue solvent accessibility – programs

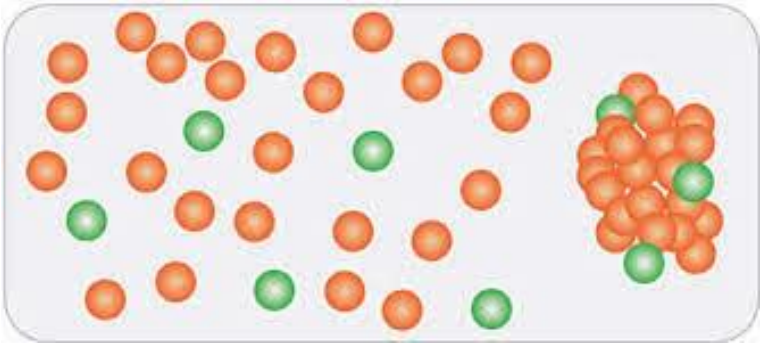
- POLYVIEW-2D (PDB) / SABLE (sequence)
  - https://polyview.cchmc.org/ ; https://sable.cchmc.org/
  - Visualization tool for structural and functional annotations of proteins, including solvent accessibility
  - Residue SASA calculated by DSSP and transformed to rASA

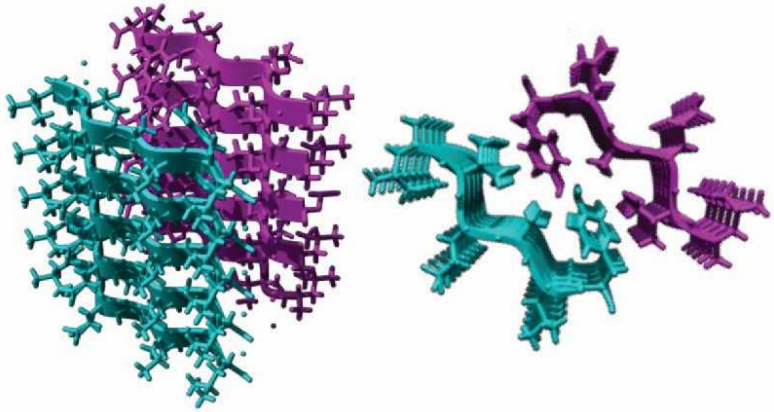




Protein solubility

- Concentration of protein in saturated solution that is in equilibrium with solid phase
- For proteins expressed in the lab, it depends on
  - Hydrophilic/hydrophobic balance of the solvent-exposed residues
  - Aggregation-prone regions (APRs) – mainly hydrophobic residues prone to form beta-structures
  - Protein expressibility in the cells

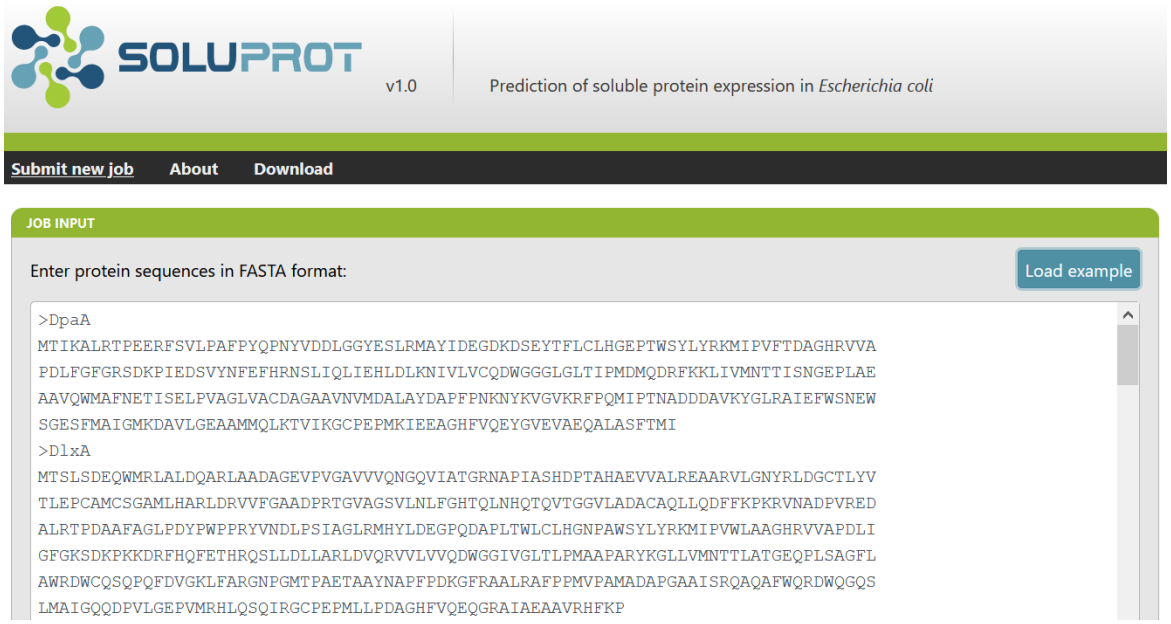




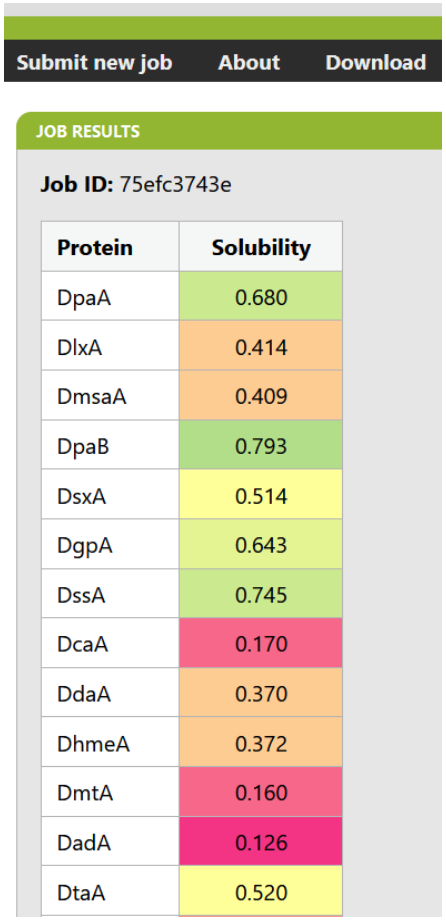
Cross-beta spines of amyloid fibrils

Protein solubility

- SoluProt
  - https://loschmidt.chemi.muni.cz/soluprot/
  - Soluble expression of protein sequences in *E.coli*
  - Based on machine learning



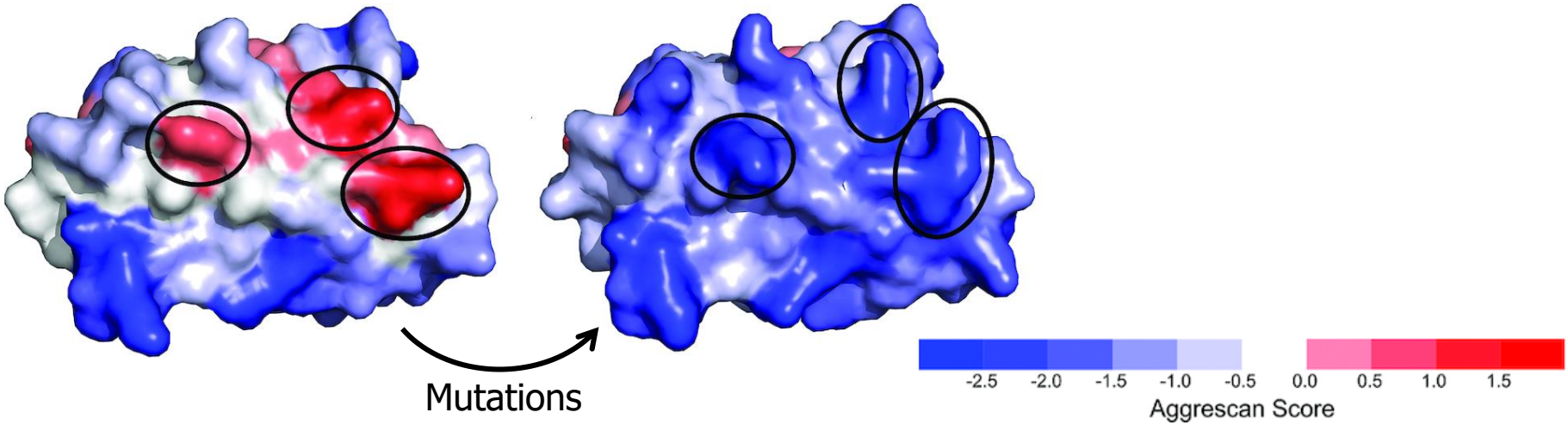
Input



Output

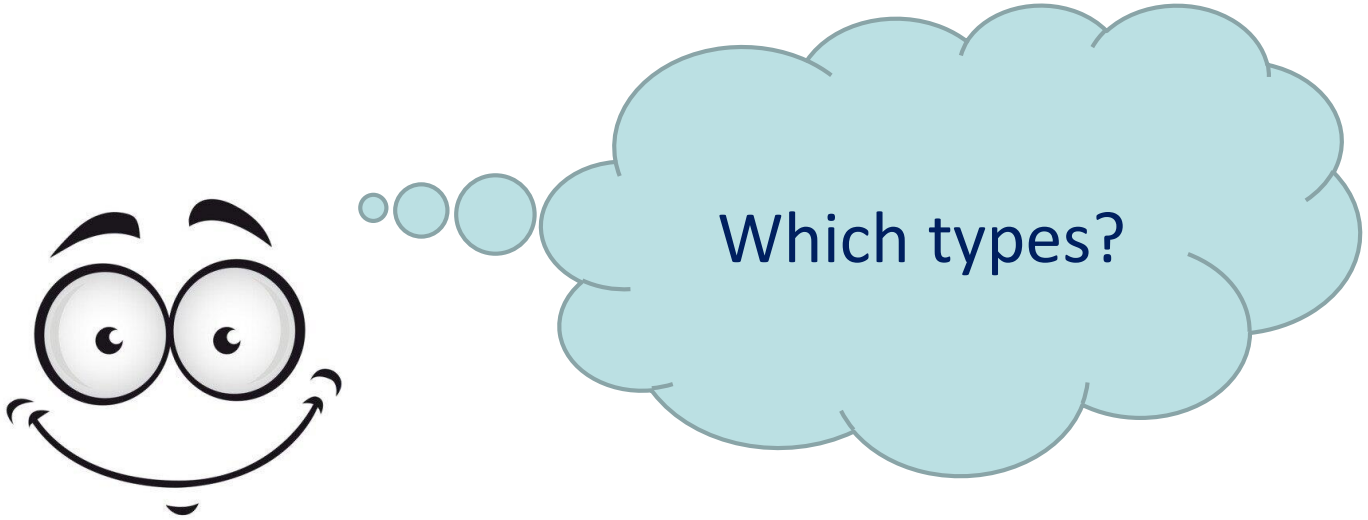
Protein solubility

- Aggrescan3D
  - http://biocomp.chem.uw.edu.pl/A3D2/
  - Predicts the aggregation propensities by identifying APRs
  - Can introduce mutations and the predict impact on stability and aggregation-propensity
  - Can account for protein flexibility (“dynamic mode”)



Molecular interactions

- Intra-molecular – within the same protein structure
- Inter-molecular – between proteins in an assembly
- Essential to understand the molecular basis for function and stability of proteins and their complexes



Types of interactions

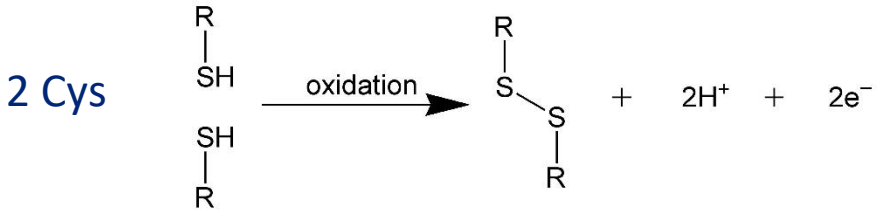
- Charge-charge (ionic) interactions
  - Present in charged residues; ex. salt bridges

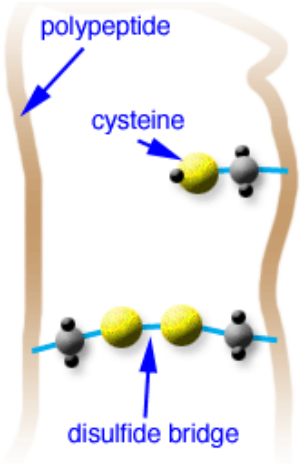
Salt bridge

- Hydrogen bonds (H-bonds)
  - Donor and acceptor atoms sharing hydrogen
- Aromatic (π-π) interactions
  - Attractive interaction between aromatic rings
- Van der Waals (vdW) interactions
  - Between any two atoms; more important for non-polar residues
- Hydrophobic interactions
  - Entropic origin; important for non-polar/hydrophobic residues

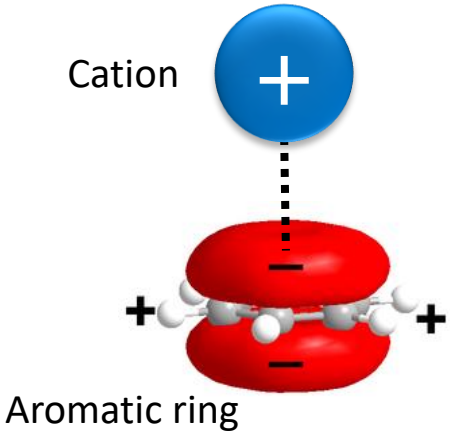
Types of interactions

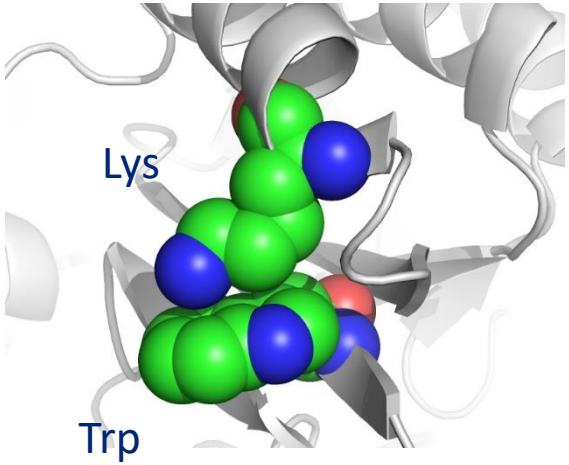
- Disulfide bonds (cysteine bridges)





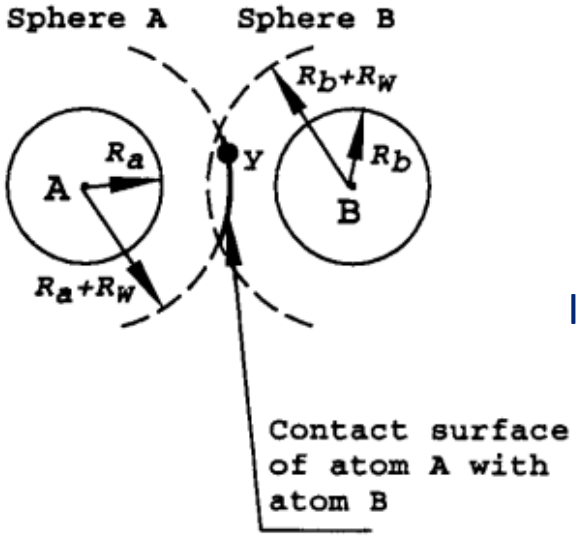
- Cation-π interactions
  - Electrostatic interaction of a positively charged residue (Lys or Arg) with an aromatic residue (Phe, Trp, or Tyr)





Molecular interactions – how to identify?

- Criteria for recognizing various types of interactions
  - Geometric rules (distances, angles)
  - Atom types
  - Energetics (physicochemical rules)
  - Contact surface area between atoms

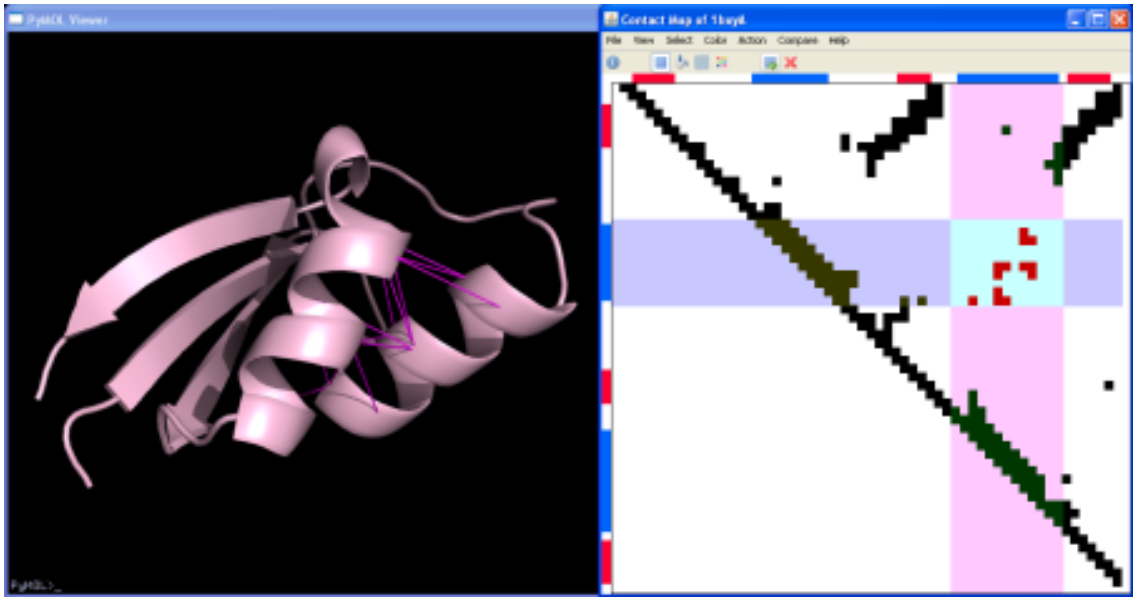


If $SASA_{Total} < SASA_A + SASA_B$

$\Rightarrow$ Interaction

Molecular interactions – programs

- CMView
  - https://www.bioinformatics.org/cmview/
    - Represents residue-residue contacts within a protein or between proteins in a complex in the form of a contact map
    - 3D visualization using PyMol

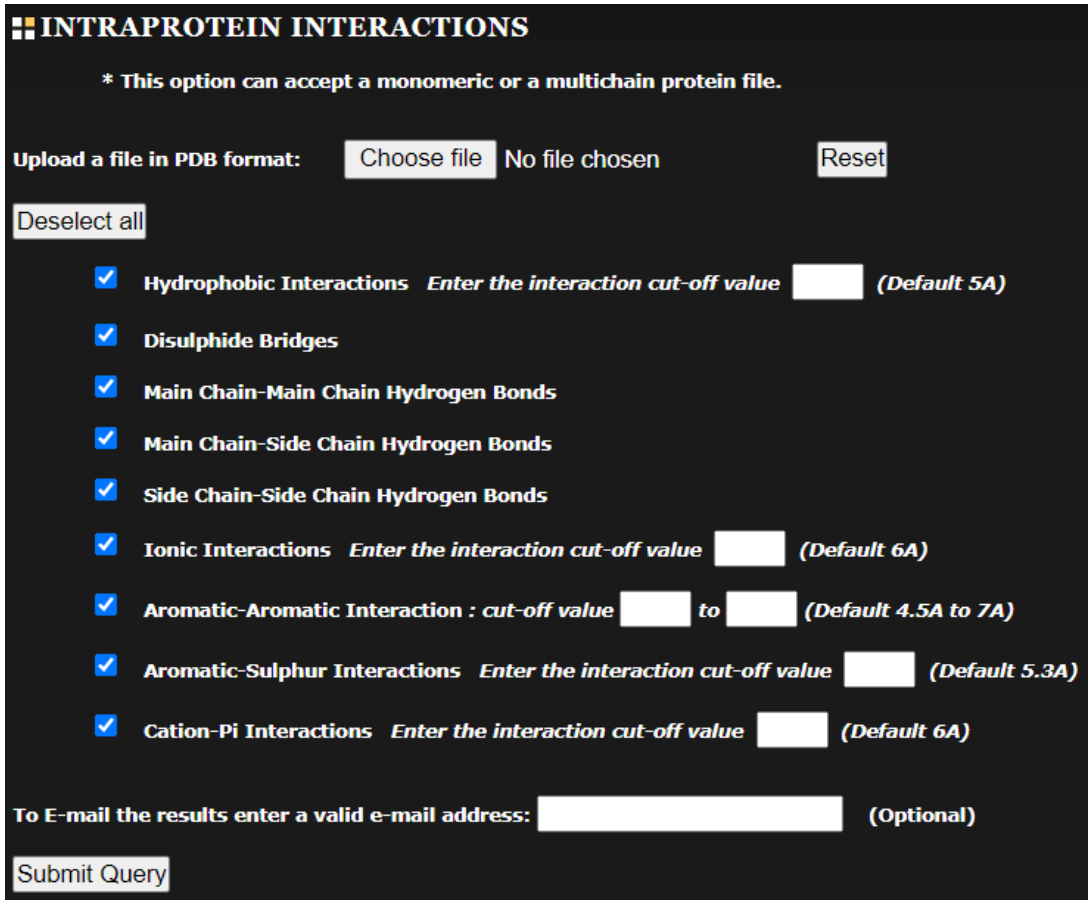


Molecular interactions – programs

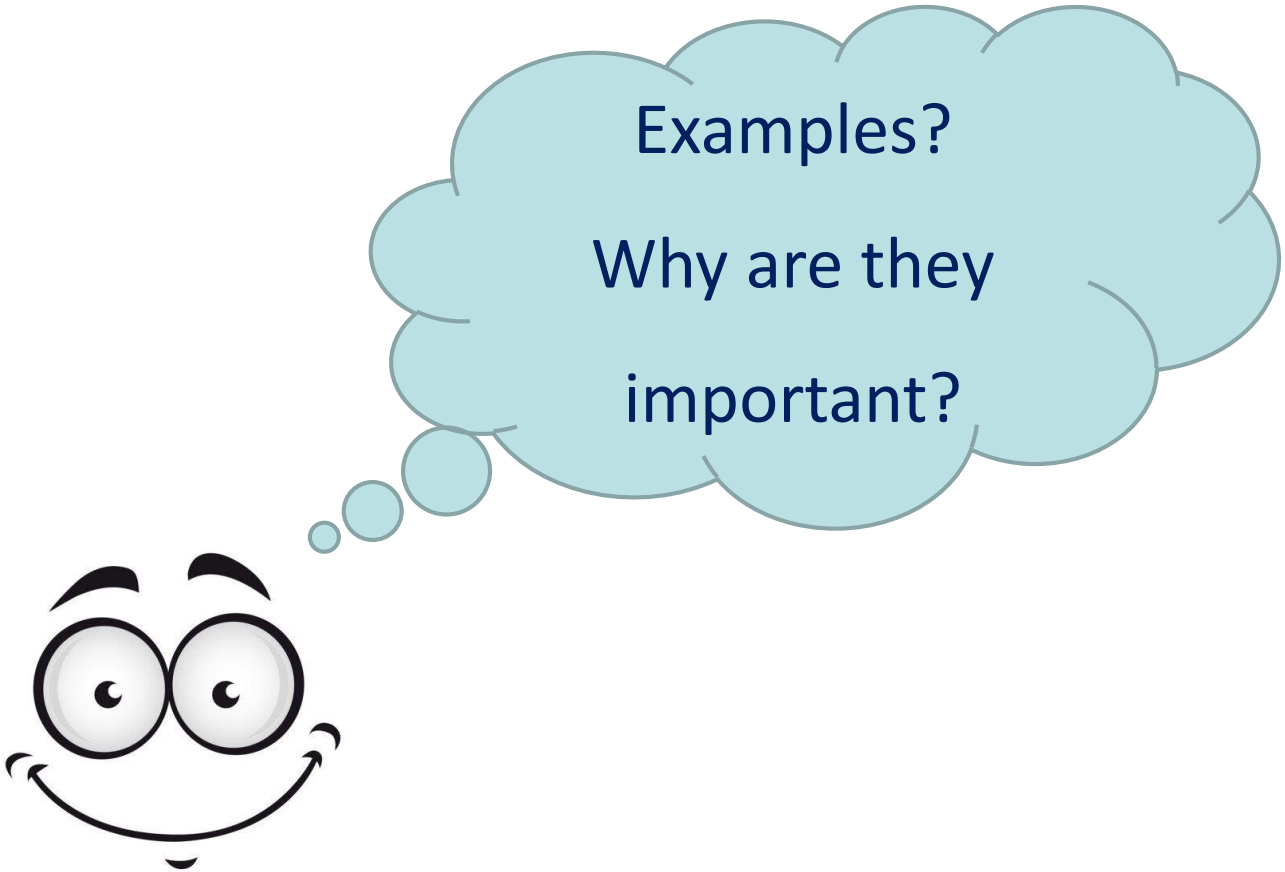
- PIC (**P**rotein **I**nteractions **C**alculator)
  - http://pic.mbu.iisc.ernet.in/
  - Identifies various interactions – hydrophobic interactions, ionic (charge-charge) interactions, hydrogen bonds, aromatic–aromatic, aromatic–sulfur, cation–π interactions, and disulfide bonds, within a protein or between proteins in a complex
  - Uses standard criteria (atom types and geometry)

Molecular interactions – programs

- PIC (**P**rotein **I**nteractions **C**alculator)
  - http://pic.mbu.iisc.ernet.in/



Functional sites



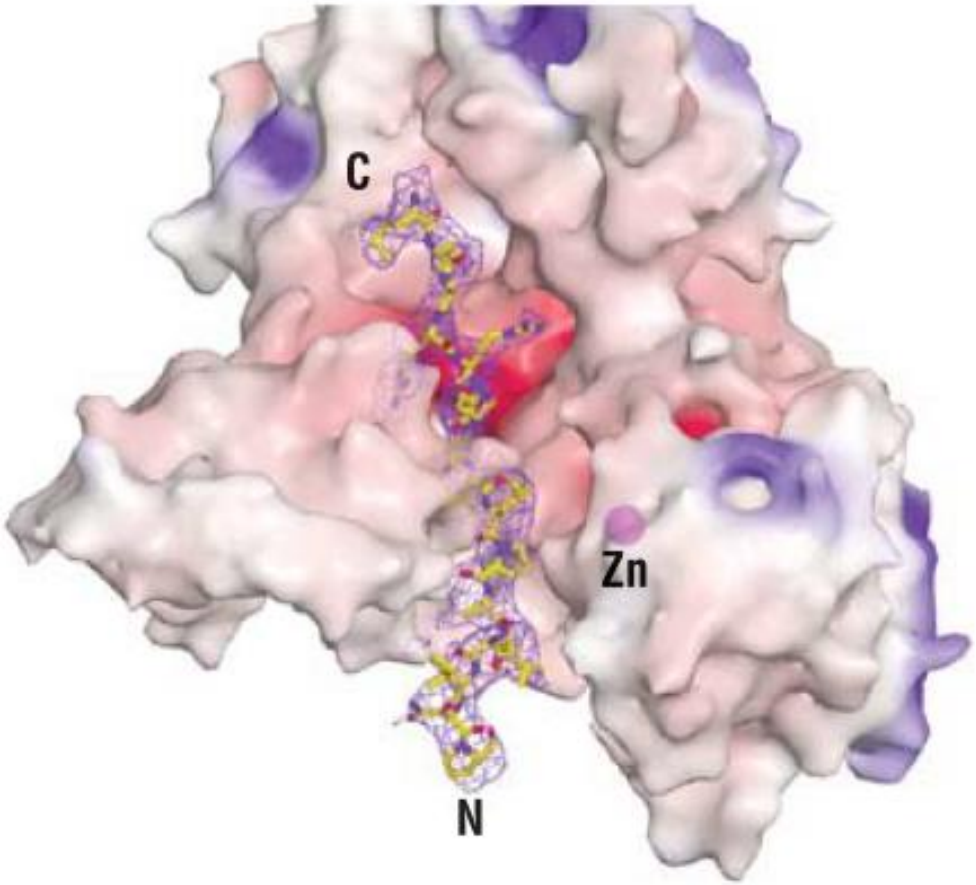
Functional sites

- Binding sites
  - Binding sites for small molecules
  - Binding sites for macromolecules
- Transport pathways
  - Voids
  - Tunnels
  - Channels

Binding sites

- Sites on the protein that provides the complementarity for the bound molecule (ligand)
  - Binding site – its function is molecular recognition
  - Active/catalytic site – special case of the binding site – its function is to promote chemical catalysis (break/formation of covalent bonds)
- Binding involves the formation of non-covalent interactions between the protein and the bound molecule
- Bound molecule – small molecule or macromolecule
- Binding is usually very specific – complementarity in shape and charge distribution between the site and bound molecule

Binding sites

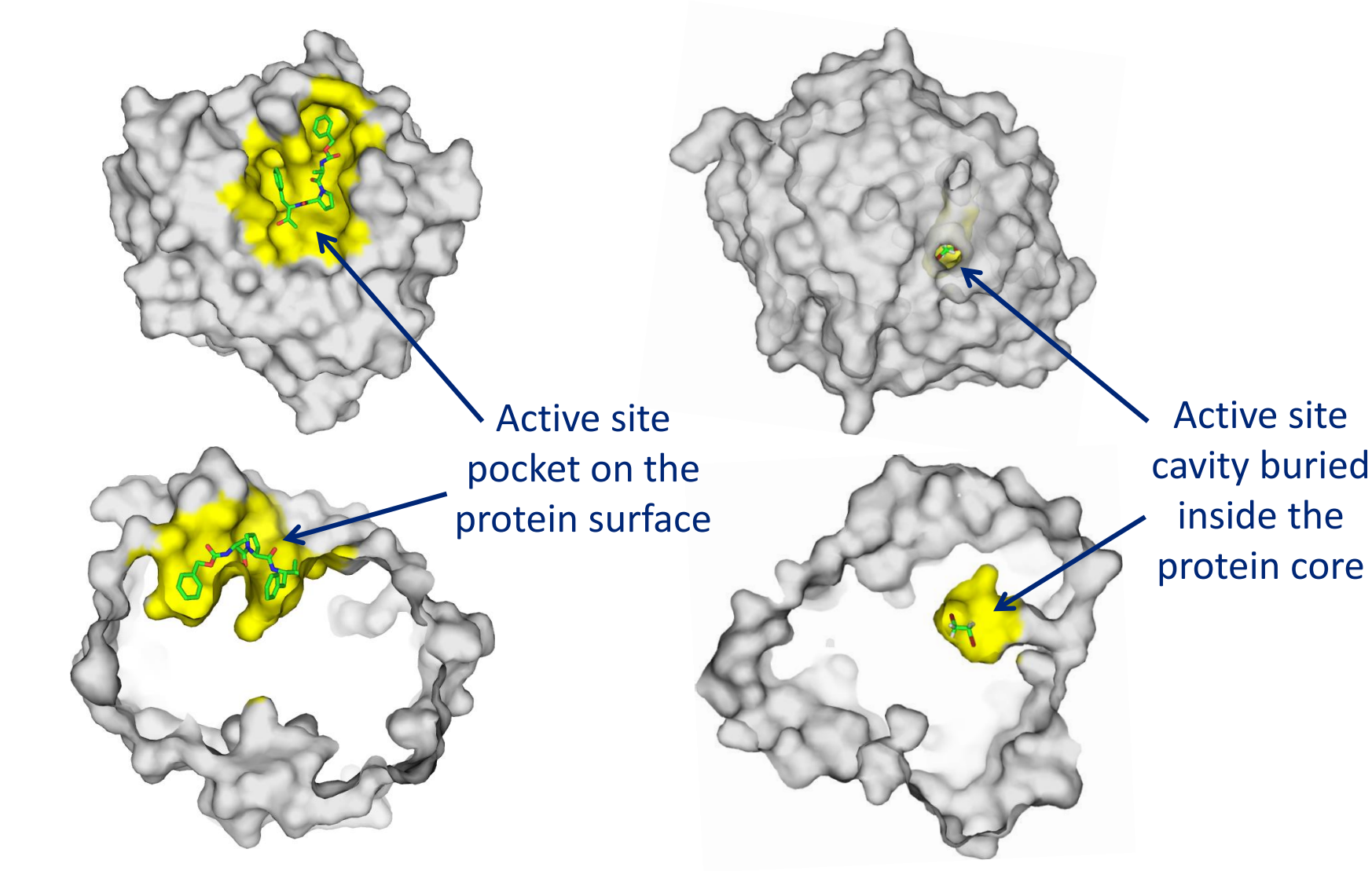


Complementarity in shape and charge distribution between the active site and substrate

Binding sites for small molecules

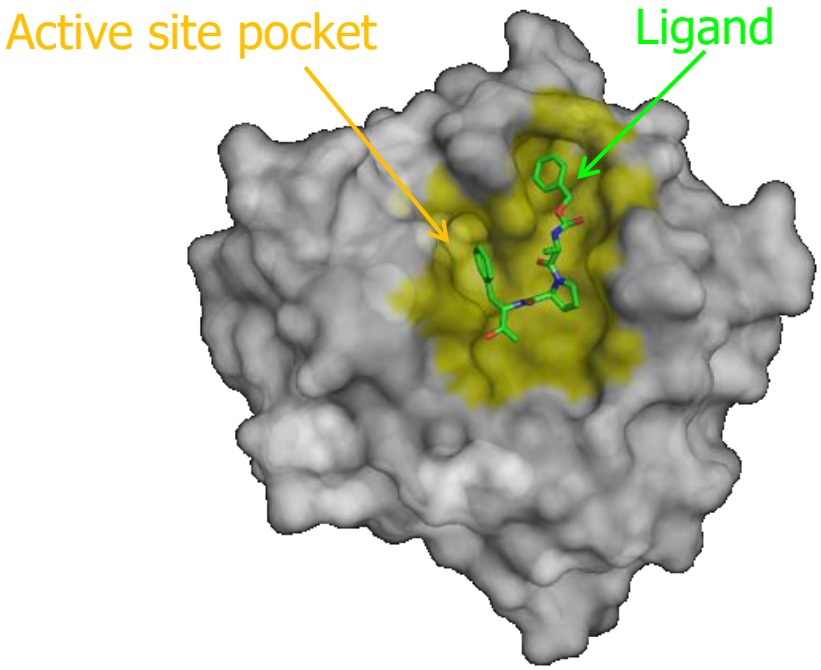
- Usually: internal cavities, surface pockets or clefts
  - Concave regions
  - Provide microenvironment different from that of the bulk solvent (e.g., many residues with negative charge → very strong electrostatic field enabling binding of highly charged ligands)
  - Often identifiable by a simple examination of the protein structure
- Highly conserved by evolution
- Low desolvation energy
- Characteristic physicochemical properties

Binding sites for small molecules



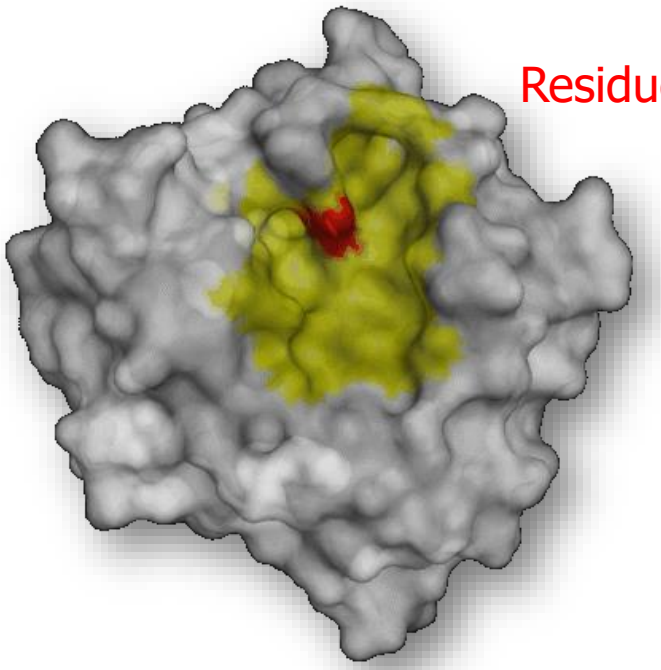
Binding sites for small molecules

- Can be very ligand-specific



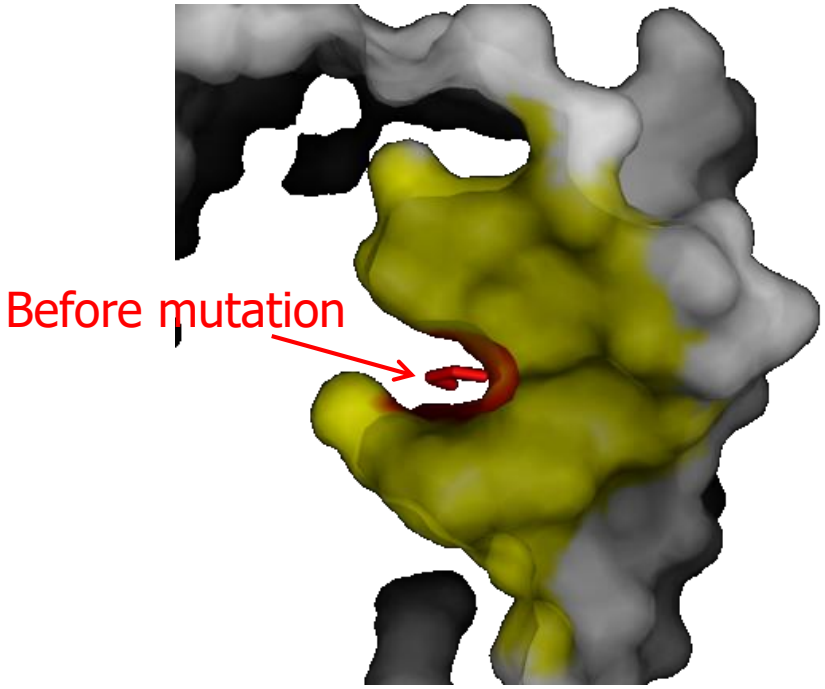
Binding sites for small molecules

- Can be very ligand-specific



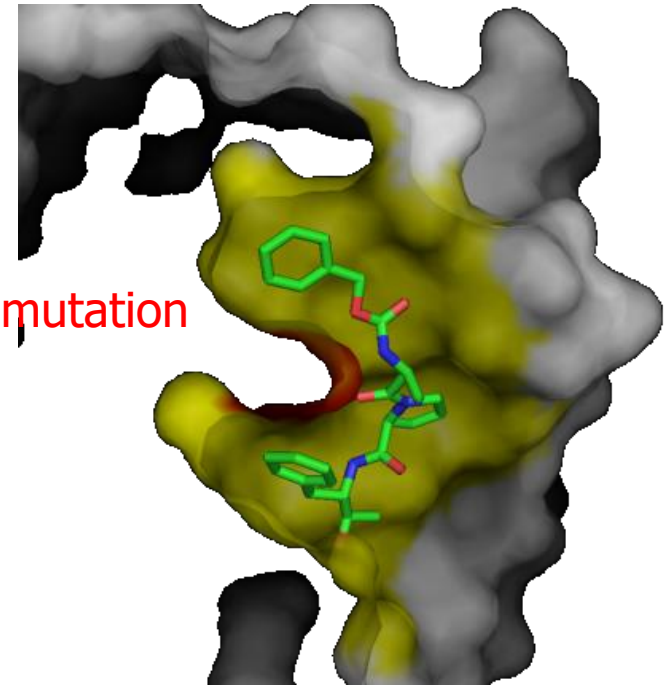
Binding sites for small molecules

- Can be very ligand-specific



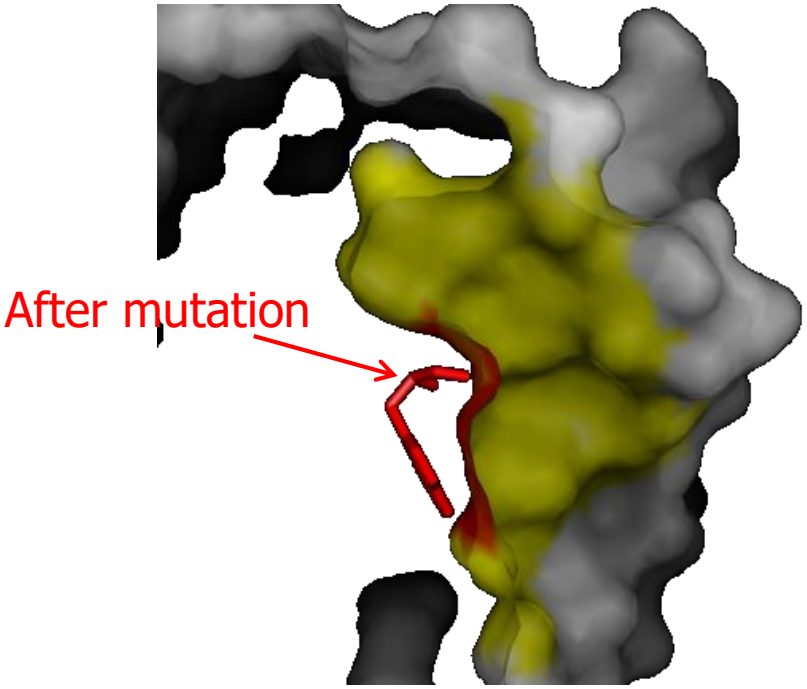
Binding sites for small molecules

- Can be very ligand-specific



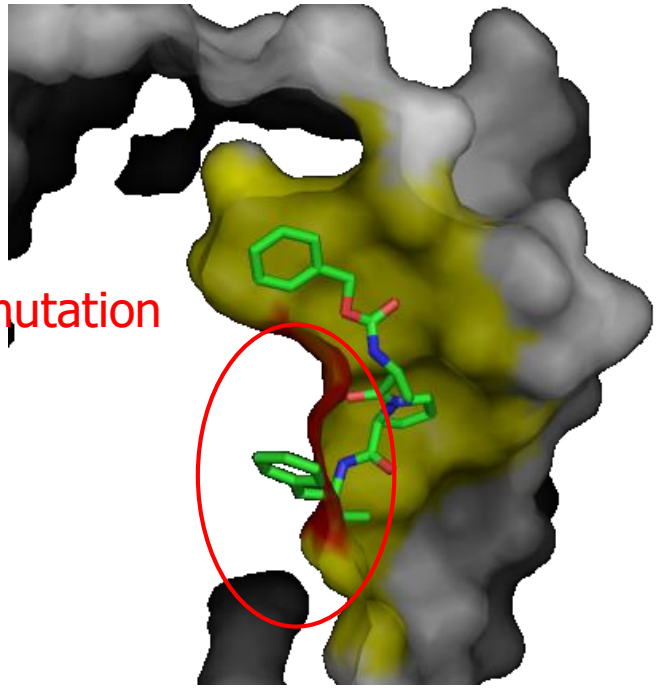
Binding sites for small molecules

- Can be very ligand-specific



Binding sites for small molecules

- Can be very ligand-specific



No longer a good fit!

Binding sites for small molecules

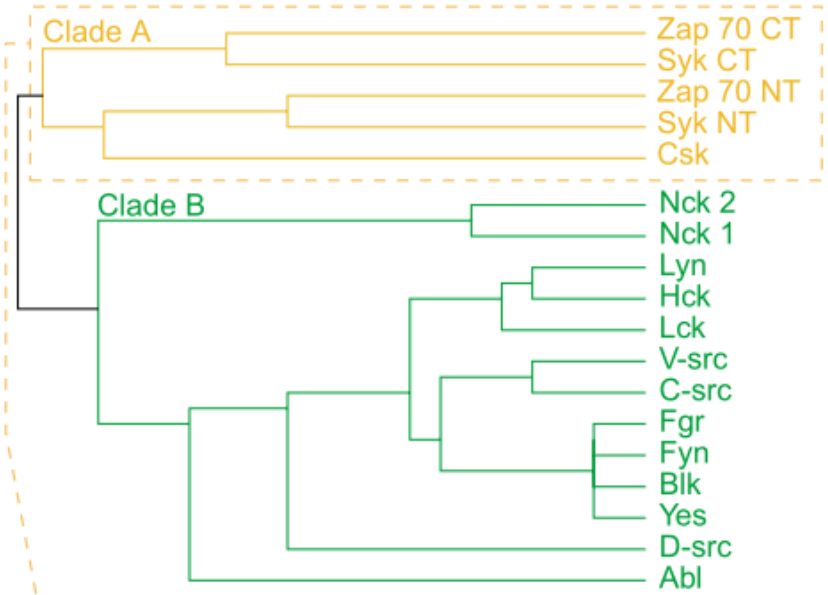
- Approaches to identify binding sites:
  - Evolutionary conservation
  - Physical detection of "pockets"
    - Geometry based methods
    - Energy based methods
  - Binding site similarity
    - Template-based methods
    - Microenvironment-based methods

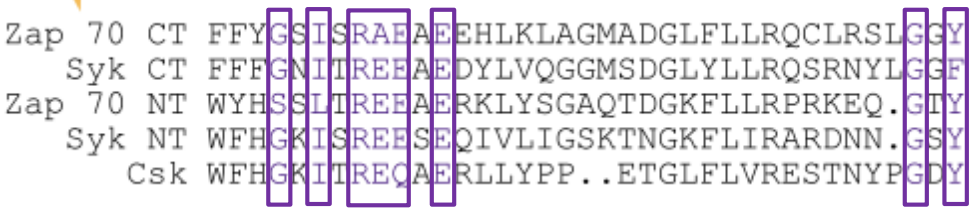
Evolutionary conservation

- Residues important for protein function or stability tend to be highly conserved over evolution
- Residue conservation in a set of related proteins can be derived from a multiple sequence alignment (MSA)
- Mapping of conservation on structure can reveal patches of conserved surface residues – potential binding sites
- Protein interior usually more conserved than surface – not suitable for prediction of buried cavities
- Not very specific – better to combine with other features

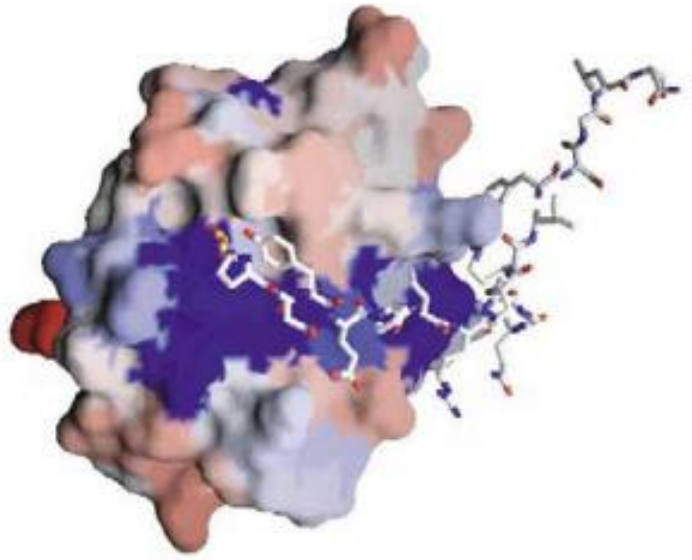
Evolutionary conservation

Phylogenetic analysis





Conservation scoring



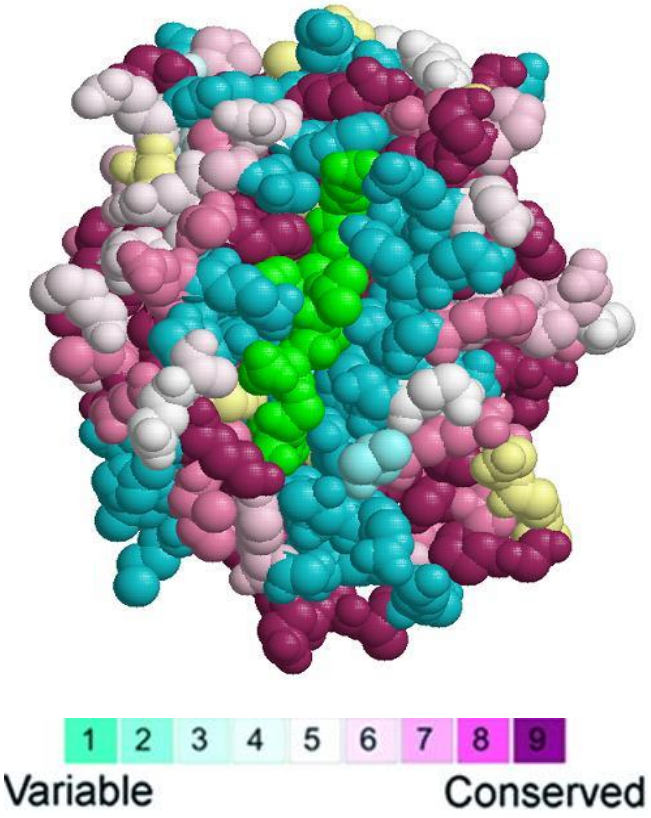
Map of evolutionary data on the structure

Evolutionary conservation

- ConSurf
  - http://consurf.tau.ac.il/
  - Estimates the level of evolutionary conservation of individual positions in protein and maps this information onto its 3D structure
  - Conservation score is derived based on the site-specific evolutionary rates calculated for each position by Rate4Site software
  - ConSurfDB – pre-calculated conservation scores for all structures from wwPDB

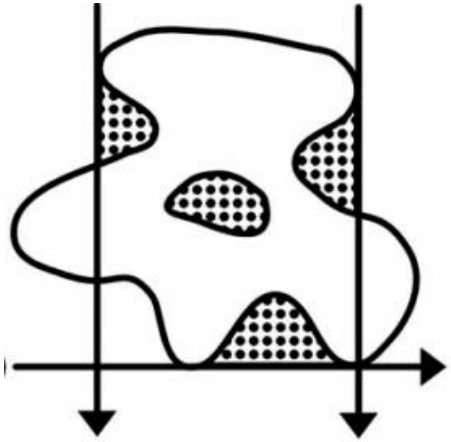
Evolutionary conservation

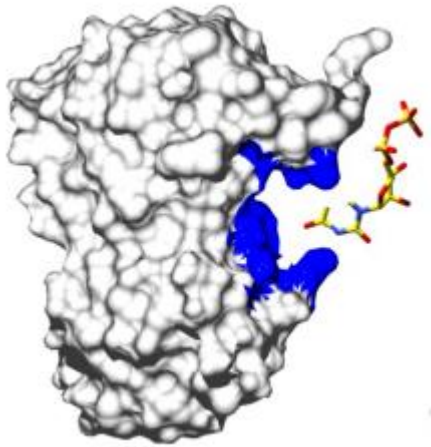
- ConSurf



Physical detection of “pockets”

- Analyze the protein surface for pockets (clefts, cavities)

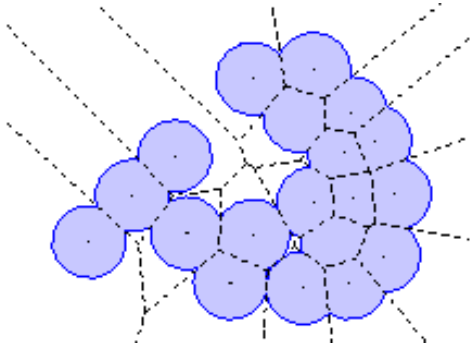




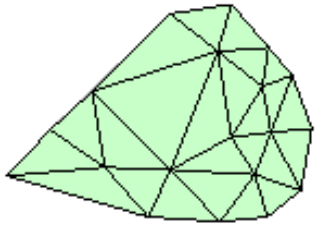
- Geometry-based methods
  - Define favorable cleft regions based on steric assessments
- Energy-based methods
  - Define favorable cleft regions based on energetic evaluations

Geometry-based methods

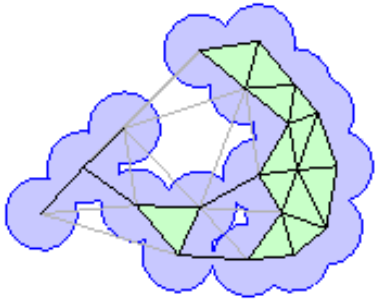
- Computed Atlas of Surface Topography of proteins (CASTp)
  - http://sts.bioe.uic.edu/castp
  - Uses computational geometry methods including Delaunay triangulation, alpha shape and discrete flow theory
  - Measures the volume and surface area of each pocket and cavity using the ASA model and molecular surface (Connolly) model



Voronoi diagram



Delaunay triangulation



Alpha shape

Energy-based methods

- Pockets are defined by energetic criteria
- Evaluate the interaction energy between the protein and a molecular fragment – probe (e.g., a methyl, hydroxyl, amine, etc.) to locate energetically favorable binding sites
- Can be combined with other methods to assess the *ligandability* (ability of a cavity to bind ligands)

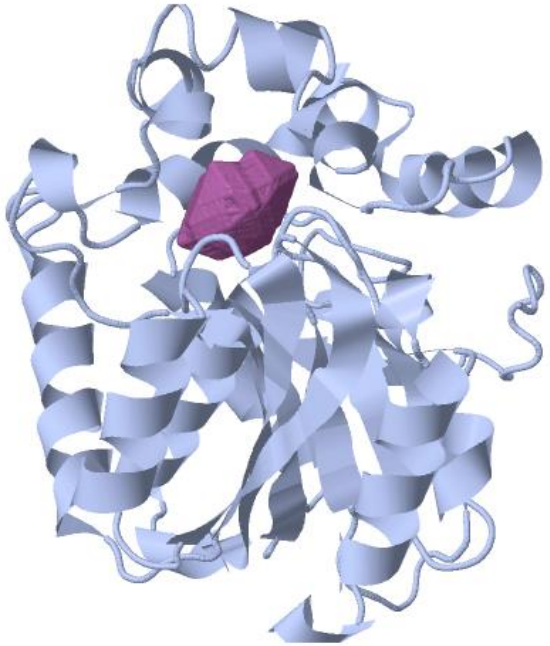
Note: *druggability* is referred to the likelihood of finding orally bioavailable small molecules that bind to a particular target in a disease-modifying way. *Ligandability* is a requirement but not sufficient condition for *druggability*.

Energy-based methods

- Cavity Plus
  - http://www.pkumdl.cn/cavityplus
  - Applies Cavity program to detect the potential binding sites and rank them with ligandability and druggability scores
  - Extracts pharmacophore features within the cavities

Energy-based methods

- Cavity Plus

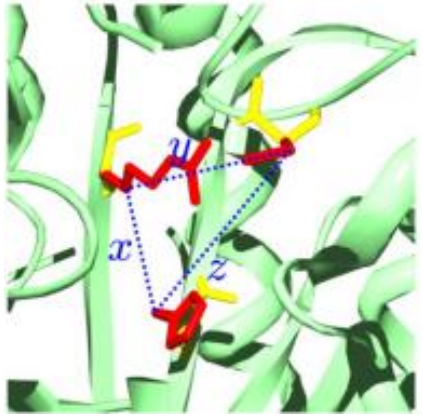


Cavity Results

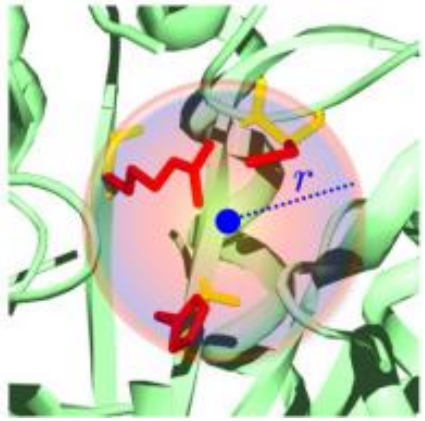
No.	Pred. Max pKd	Pred. Avg pKd	DrugScore	Druggability	Surface	Residues
1	10.19	6.11	493.00	less druggable	☑	More
2	8.87	5.66	-745.00	Undruggable	☐	More
3	8.16	5.42	-420.00	Undruggable	☐	More
4	7.87	5.32	-750.00	Undruggable	☐	More
5	7.11	5.06	-1105.00	Undruggable	☐	More
6	6.54	4.86	-992.00	Undruggable	☐	More
7	5.90	4.64	-1123.00	Undruggable	☐	More

Binding site similarity

- Prediction of binding sites is based on the similarity with other (known) binding sites



- Template-based methods
  - Binding sites are represented by 3D templates
  - Based on similarity with homologous proteins
- Microenvironment-based methods
  - Based on description of local environment, such as type of residues, their distances, solvent accessibility and physicochemical properties



Template-based methods

- Definition and construction of 3D templates of features
  - Local structural motifs, patterns and descriptors that characterize the binding sites (e.g., functional groups, shape, solvent accessibility, etc.)
  - Capture the essence of the binding sites in protein
  - Usually apply constraints on atom types and occasionally sequential relationships
- Search a database for structures using template as a query
  - Identification of structures with a given binding site
- Compare the query structure against a 3D template database
  - Identification of potential binding sites in the query structure

Template-based methods

- PINTS
  - http://www.russelllab.org/cgi-bin/tools/pints.pl
  - To compare a protein structure against a database of 3D patterns (templates), as well as 3D templates against a database of protein structures
  - Additionally allows comparison of two structures
  - The 3D template database includes ligand-binding sites and SITE annotations from PDB files

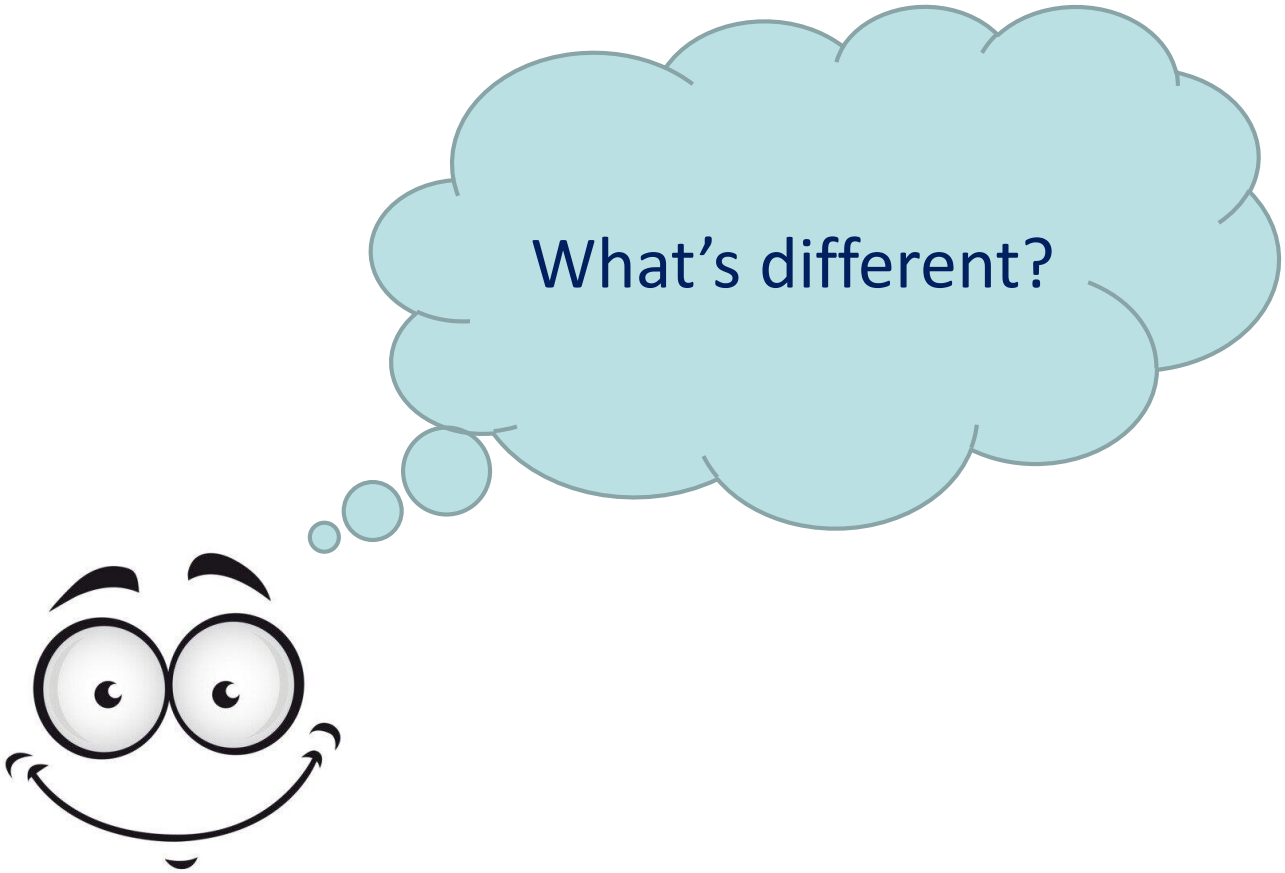
Template-based methods

- ProFunc
  - http://www.ebi.ac.uk/thornton-srv/databases/profunc/
  - Aims to identify the most likely function of a protein from its 3D structure
  - Uses several methods, including fold matching, residue conservation, surface cleft analysis, and functional 3D templates (templates for enzyme active sites, ligand-binding templates, DNA-binding templates, reverse template comparison vs. structures in wwPDB)

Template-based methods

- Mechanism and Catalytic Site Atlas
  - https://www.ebi.ac.uk/thornton-srv/m-csa/
  - Database that provides information about the active sites, catalytic residues and reaction mechanisms in enzymes with experimentally determined 3D structure
  - Defines catalytic residues as the residues directly involved in some aspect of the enzymatic reaction
  - Provides 3D templates for catalytic sites in the database

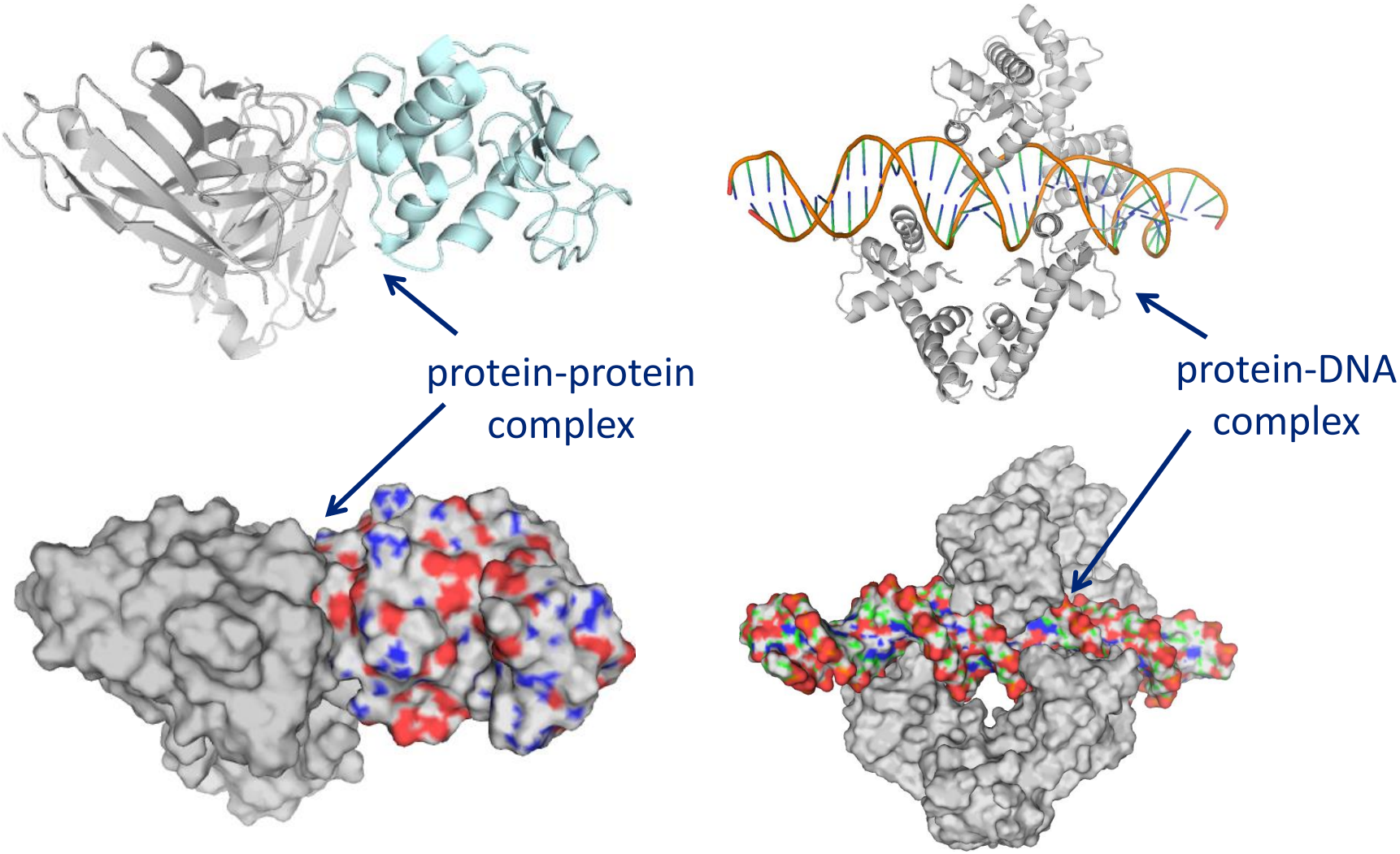
Binding sites for macromolecules



Binding sites for macromolecules

- Typically protruding loops, large surface clefts but also flat binding sites – flatter than binding sites for small molecules
  - Recognition of a macromolecule involves interactions over a large continuous surface area or several discrete binding regions
  - Difficult to identify by a simple examination of the protein structure
- High evolutionary conservation
- Low desolvation energy
- Characteristic physicochemical properties
- DNA binding sites have characteristic motifs and positive charged electrostatic patches

Binding sites for macromolecules



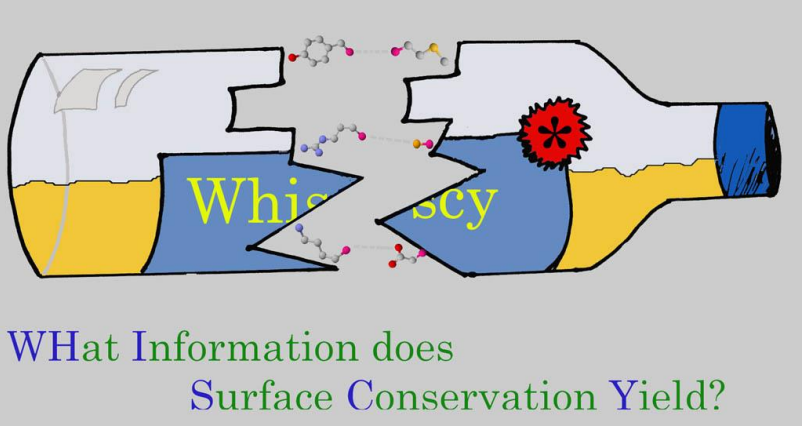
Binding sites for macromolecules

- Approaches to identify binding sites
  - Evolutionary conservation
  - Knowledge-based

- Meta-servers (tools that combine several methods)

Evolutionary conservation methods

- Same principles as for binding sites of small molecules (see above)
- WHISCY
  - https://wenmr.science.uu.nl/whiscy/
  - Predicts protein-protein interface using conservation and structural information (interface propensities for each residue at the surface are used to adjust the score)



Knowledge-based methods

- Combine multiple interface features
  - Conservation
  - Residue propensity for being at protein-protein interfaces
  - Physicochemical properties
  - Structural properties
- Use known binding sites for parameterization or training → empirical scoring functions and machine learning methods

Knowledge-based methods

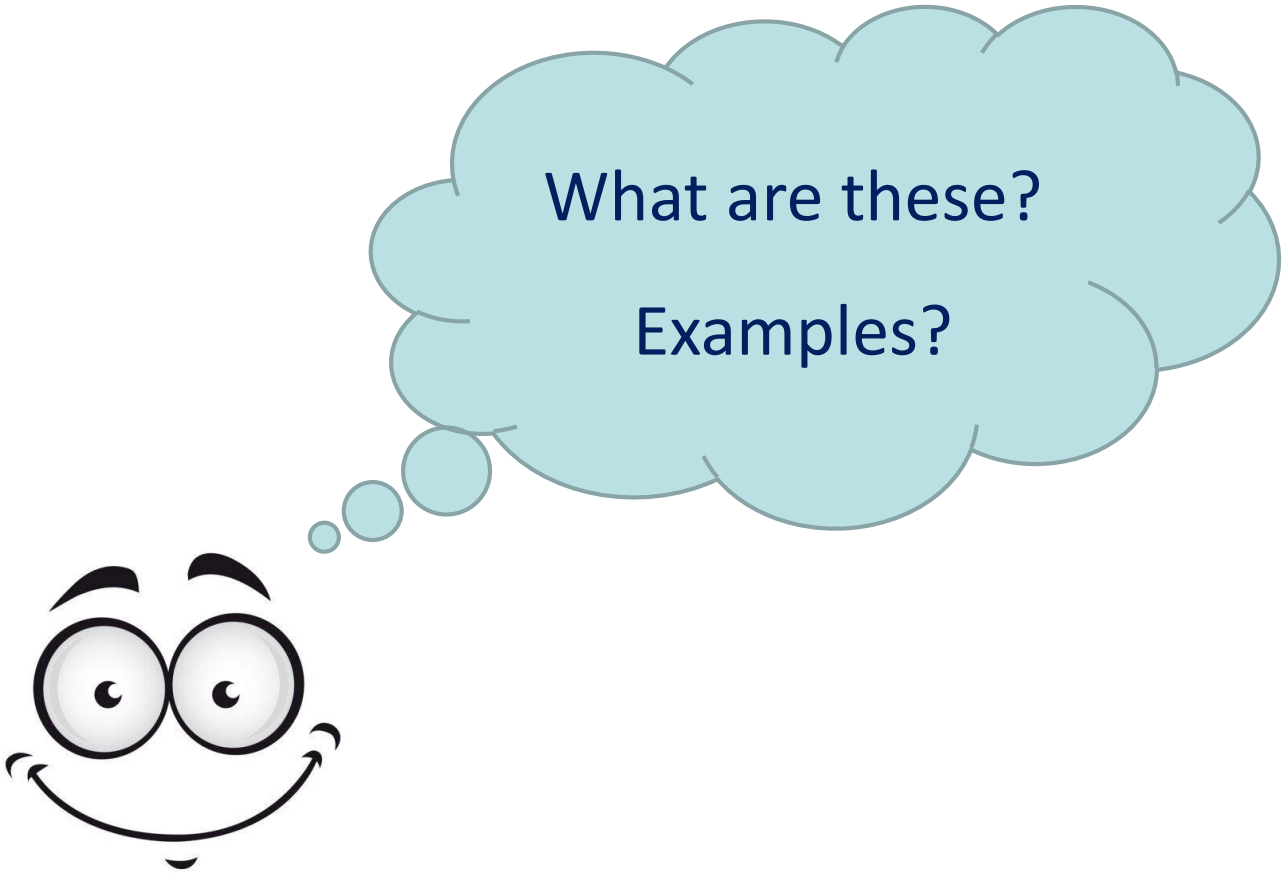
- CONS-PPISP (consensus Protein-Protein Interaction Site Predictor)
  - http://pipe.scs.fsu.edu/ppisp.html
  - Utilizes machine learning to predict protein binding sites
  - Trained on position-specific sequence profiles and solvent accessibilities of each residue and its spatial neighbors
- Patch Finder Plus
  - http://pfp.technion.ac.il/
  - Utilizes machine learning primarily to find DNA binding regions
  - Identifies the largest positive electrostatic patch on a protein surface – combination of residue frequency, composition and conservation, surface concavity, accessible area and H-bond potential

Meta-servers

- Combine multiple methods to improve prediction accuracy

- META-PPISP (Protein Protein Interaction Site Predictor)
  - http://pipe.scs.fsu.edu/meta-ppisp.html
  - Combines cons-PPISP, ProMate and PINUP
- PI^2PE (Protein Interface/Interior Prediction Engine)
  - http://pipe.scs.fsu.edu/
  - Pipeline to use five different predictors including cons-PPISP, meta-PPISP and DISPLAR

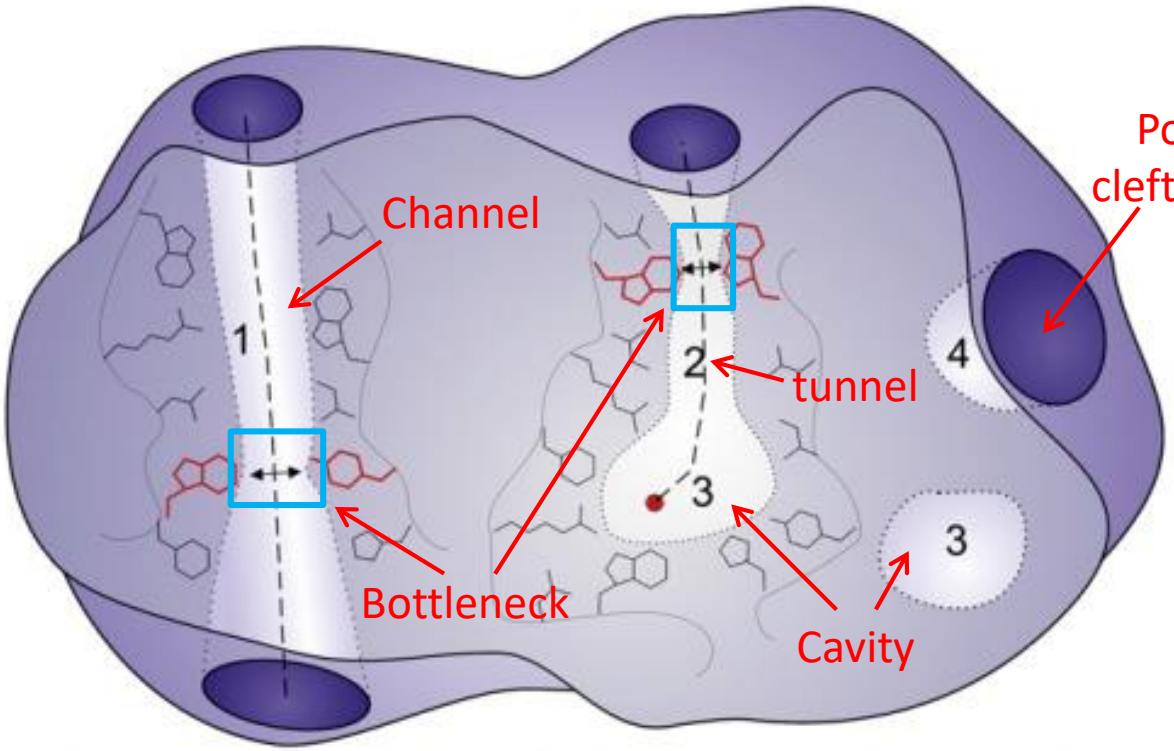
Transport pathways

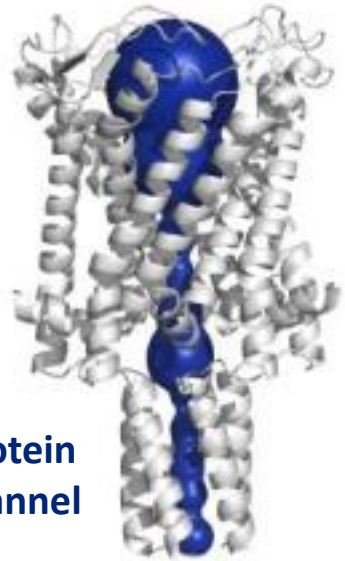


Transport pathways

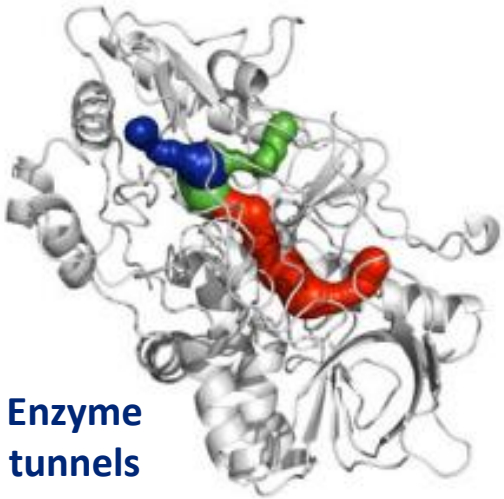
- Mediate transport of ions and small molecules in proteins – an essential role in functioning of large variety of proteins
  - Channels/pores – transport of substances across membranes
  - Tunnels – exchange of ligands between buried active/binding site cavities and the bulk solvent
  - Intramolecular tunnels – transport of reaction intermediates between two distinct active sites in bifunctional enzymes
- The permeability to different substances depends on their size (radii), shape (length and curvature), amino acid composition (physicochemical properties) and dynamics

Transport pathways





Protein channel

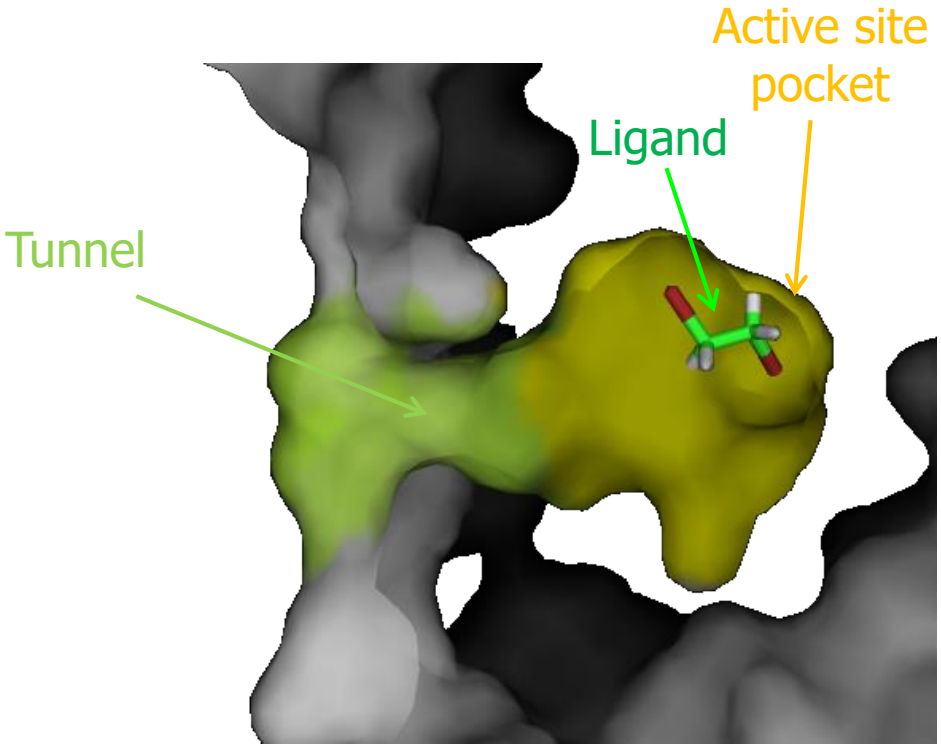


Enzyme tunnels

- Bottleneck – the narrowest part of the tunnel/channel; it has critical importance for the selectivity

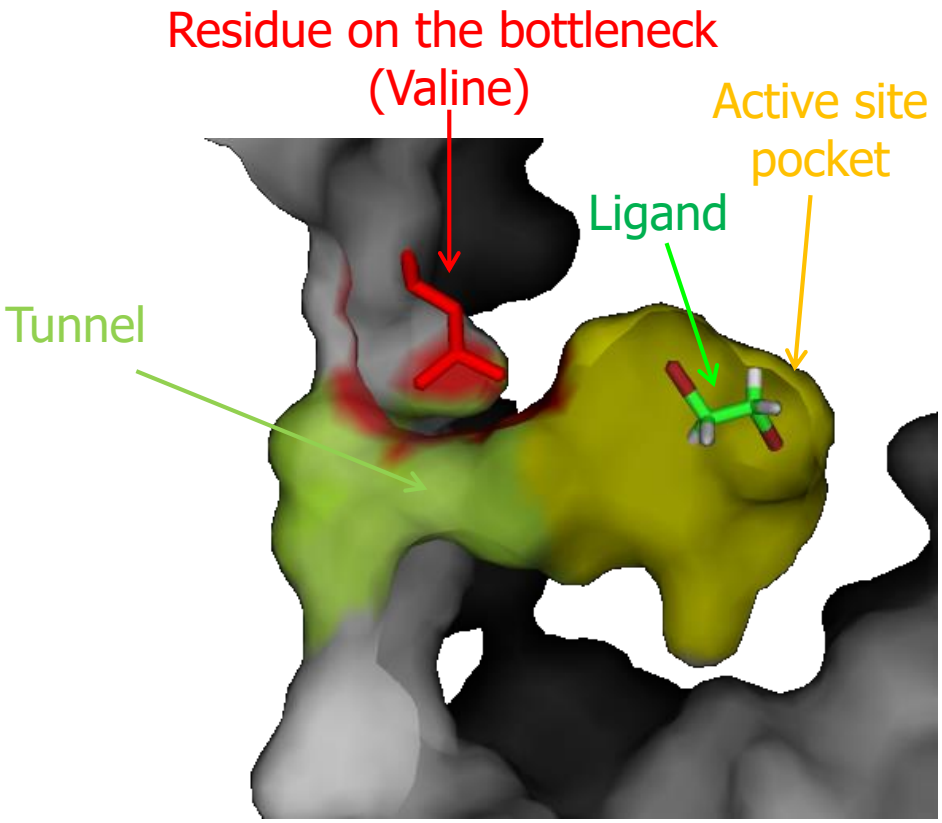
Transport pathways

- Dependence on the residues



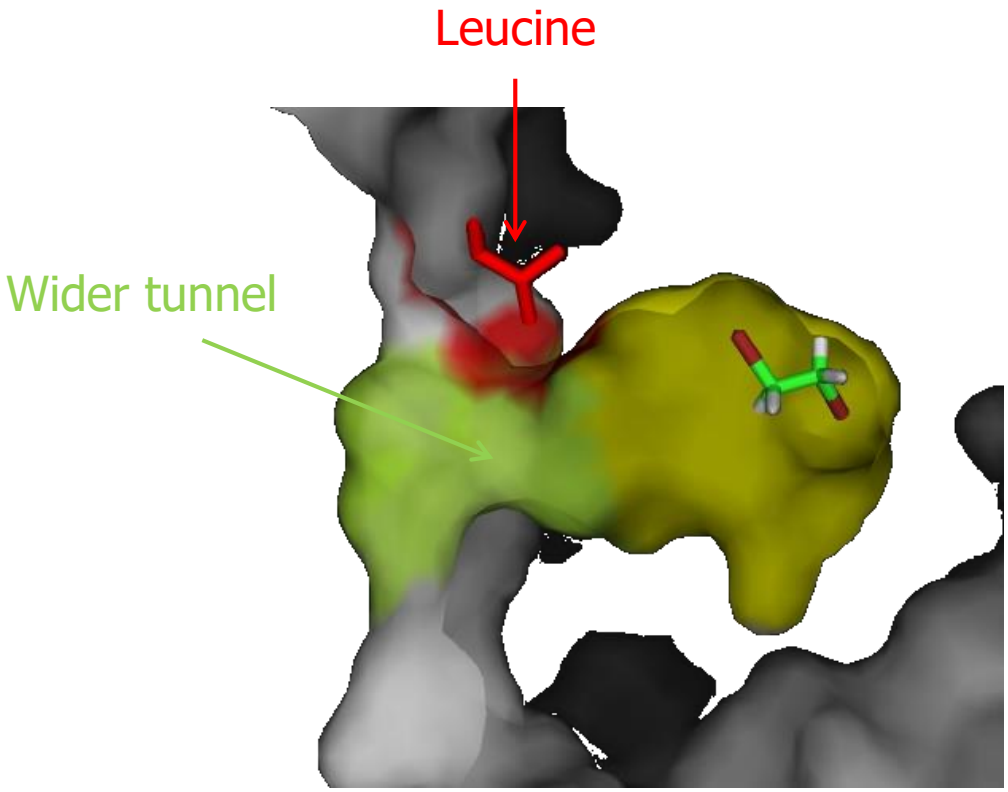
Transport pathways

- Dependence on the residues



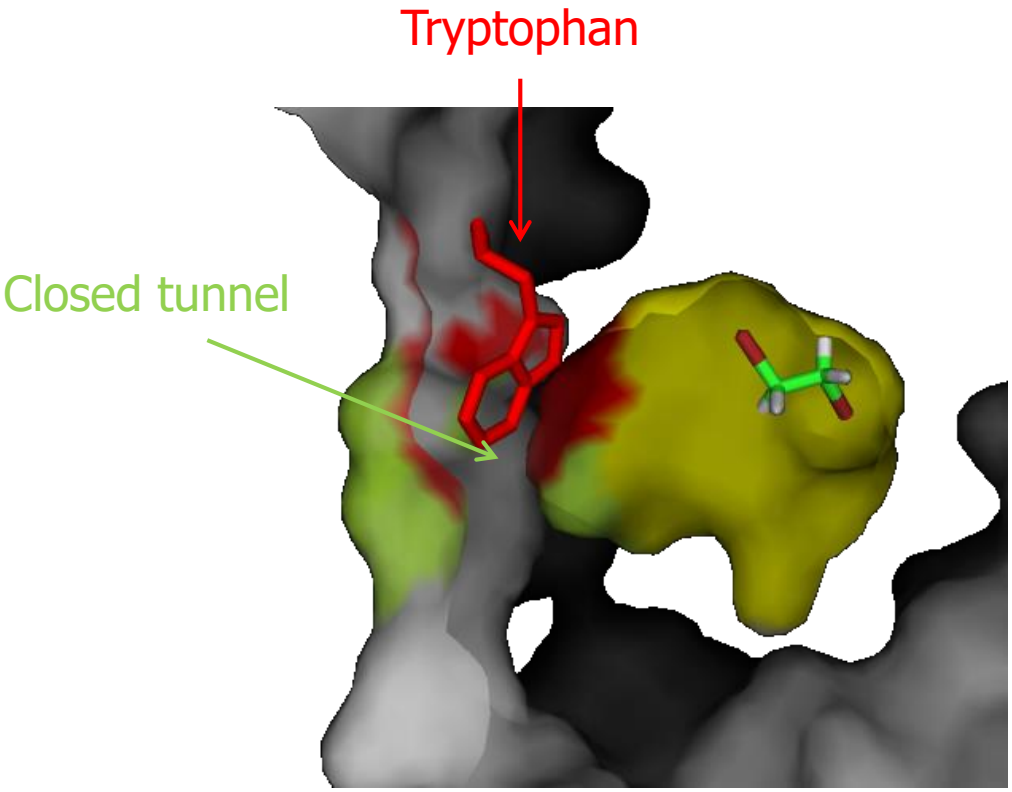
Transport pathways

- Dependence on the residues



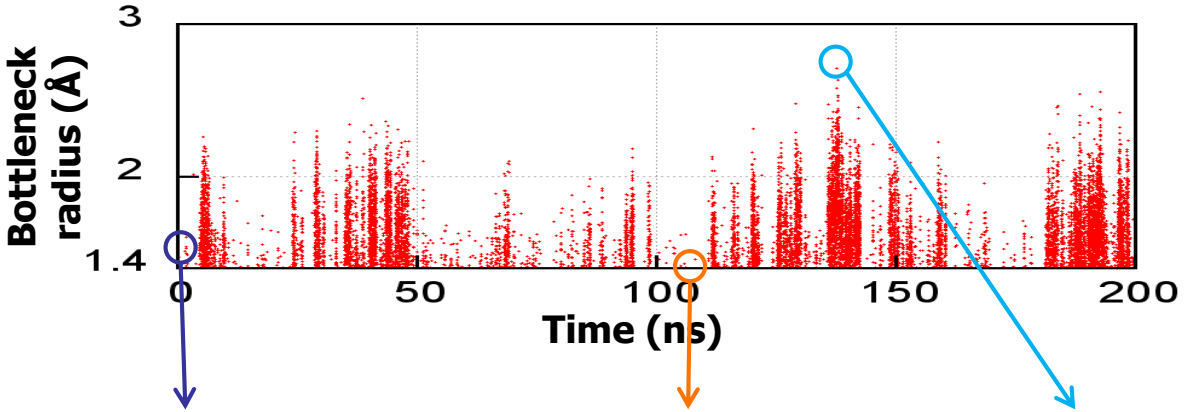
Transport pathways

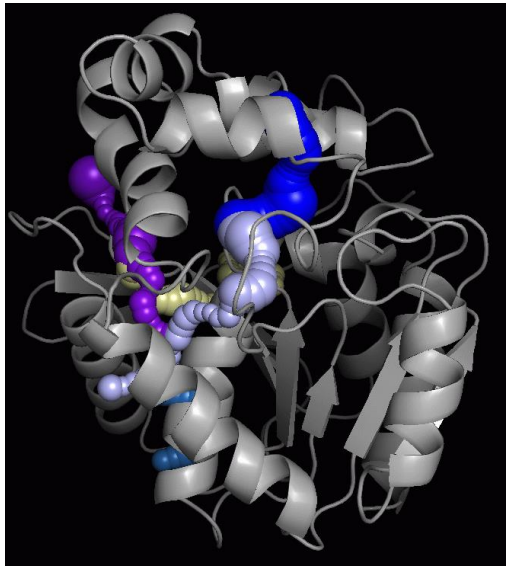
- Dependence on the residues

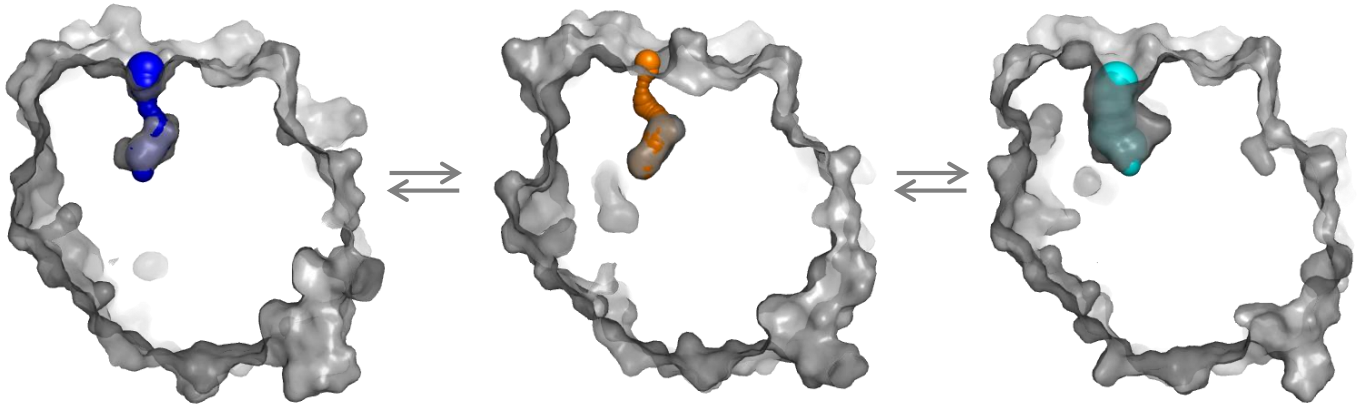


Transport pathways

- Dependence on protein dynamics







Prediction of transport pathways

- Identification of overall voids in proteins
- Identification of tunnels
- Identification of channels

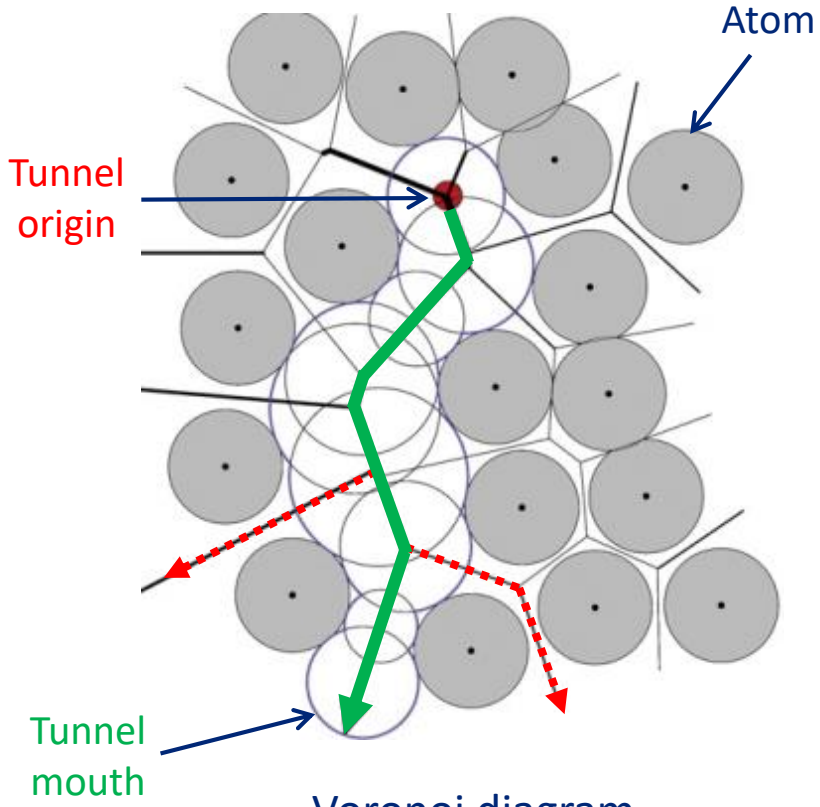
Identification of overall voids

- Methods that aim to accurately represent all types of voids in a protein structure, including channels, tunnels, surface clefts, pockets as well as internal cavities
- Usually provide very limited information on tunnel and channel characteristics – the identified voids have to be separated from each other
- Geometry-based methods for pocket detection
  - HOLLOW – http://hollow.sourceforge.net/
  - 3V – http://3vee.molmovdb.org/
  - fPocket, LIGSITE^{csc}, PASS, CASTp, SURFNET, POCASA …

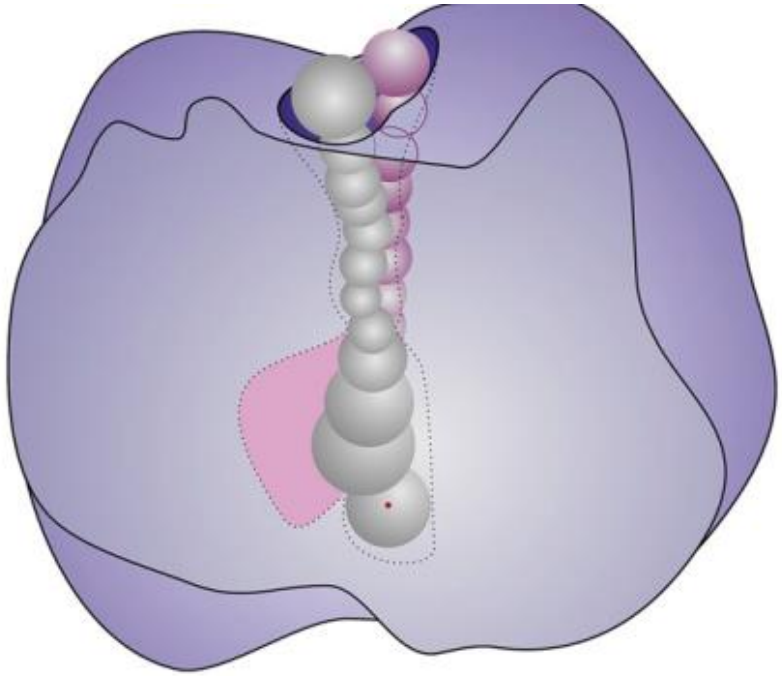
Identification of tunnels

- Methods that calculate tunnels connecting occluded cavities with the surrounding bulk solvent
- Identify the pathways from a cavity to the protein surface
- Voronoi diagrams described by the skeleton of voids between atoms to find all theoretically possible pathways connecting the starting point with the bulk solvent
- Diagrams of optimal pathways using Dijkstra's algorithm, based on criteria defined by a cost function
- The probe size defines the lowest radius threshold
- Tunnel geometry is approximated by a sequence of spheres

Identification of tunnels



Voronoi diagram



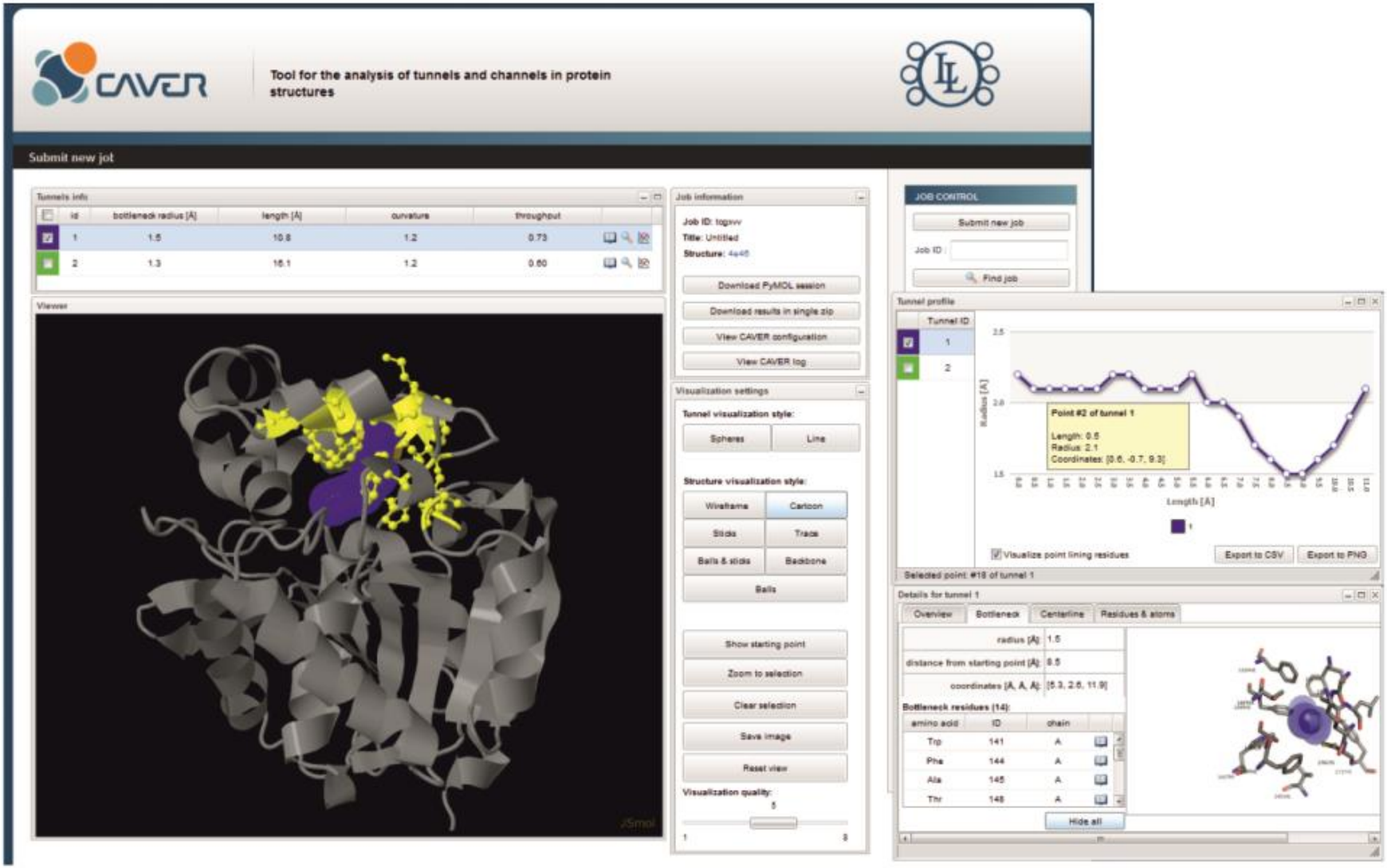
Common limitation: the tools identify two spherical tunnels instead of one asymmetric tunnel

- Probe size: the minimum radius specified for the tunnel search
- Allowed pathway according to the selected probe
- Disallowed pathways

Identification of tunnels - programs

- CAVER 3.0
  - http://caver.cz/
  - Command-line stand-alone and PyMOL plugin
  - GUI with CAVER Analyst 2
  - For static structures and dynamic ensembles
- CAVER Web
  - http://loschmidt.chemi.muni.cz/caverweb/
  - Interactive guide-through web server
  - Optimized protocol for detection of biologically relevant tunnels
  - Based on CAVER 3.0 program

Identification of tunnels - programs

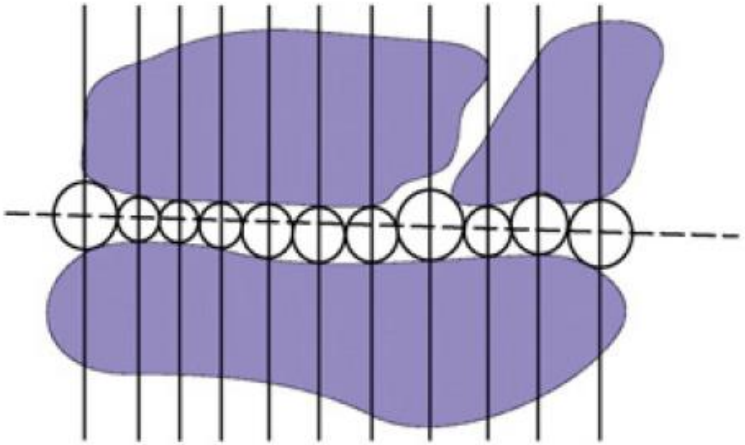


Identification of channels

- Methods that calculate channels (or pores) penetrating throughout the proteins
- Not suitable to identify tunnels leading from occluded cavities
- Usually analyze just one channel per structure
- Usually need information about approximate position and direction of the channel (channel axis) – user-provided or automatically identified

Identification of channels - programs

- POREWALKER
  - http://www.ebi.ac.uk/thornton-srv/software/PoreWalker/
  - Identifies channel axis by heuristic iterative approach (based on the axes of transmembrane secondary structures)
  - Protein is divided into equally-spaced slices perpendicular to the axis; the largest spheres fitting the channel are identified

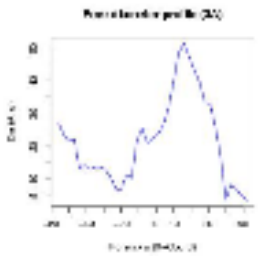


Identification of channels - programs

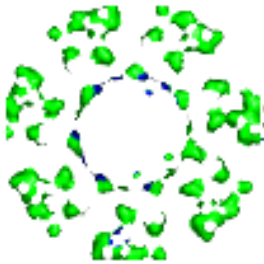
- POREWALKER

Pore analysis results

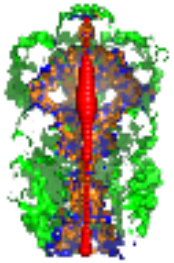
Overview of the available results:



Pore shape



Slideshow of the pore cross-sections



Pore visualization

Features of the cavity

References

- Gu, J. & Bourne, P. E. (2009). **Structural Bioinformatics, 2nd Edition**, Wiley-Blackwell, Hoboken, p. 1067.
- Laurie, A. T. & Jackson, R. (2006). Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Current Protein and Peptide Science* **7**: 395-406.
- Campbell, S. J. *et al.* (2003). Ligand binding: functional site location, similarity and docking. *Current opinion in structural biology* **13**: 389-395.
- Xin, F. & Radivojac, P. (2011). Computational methods for identification of functional residues in protein structures. *Current protein and peptide science* **12**: 456-469.
- Leis, S. *et al.* (2010). *In silico* prediction of binding sites on proteins. *Current medicinal chemistry* **17**: 1550-1562.
- Fernández-Recio, J. (2011). Prediction of protein binding sites and hot spots. *Computational molecular science* **6**: 680-698.
- 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.
- Brezovsky, J. *et al.* (2012). Software tools for identification, visualization and analysis of protein tunnels and channels. *Biotechnology advances.* In press: doi:10.1016/j.biotechadv.2012.02.002