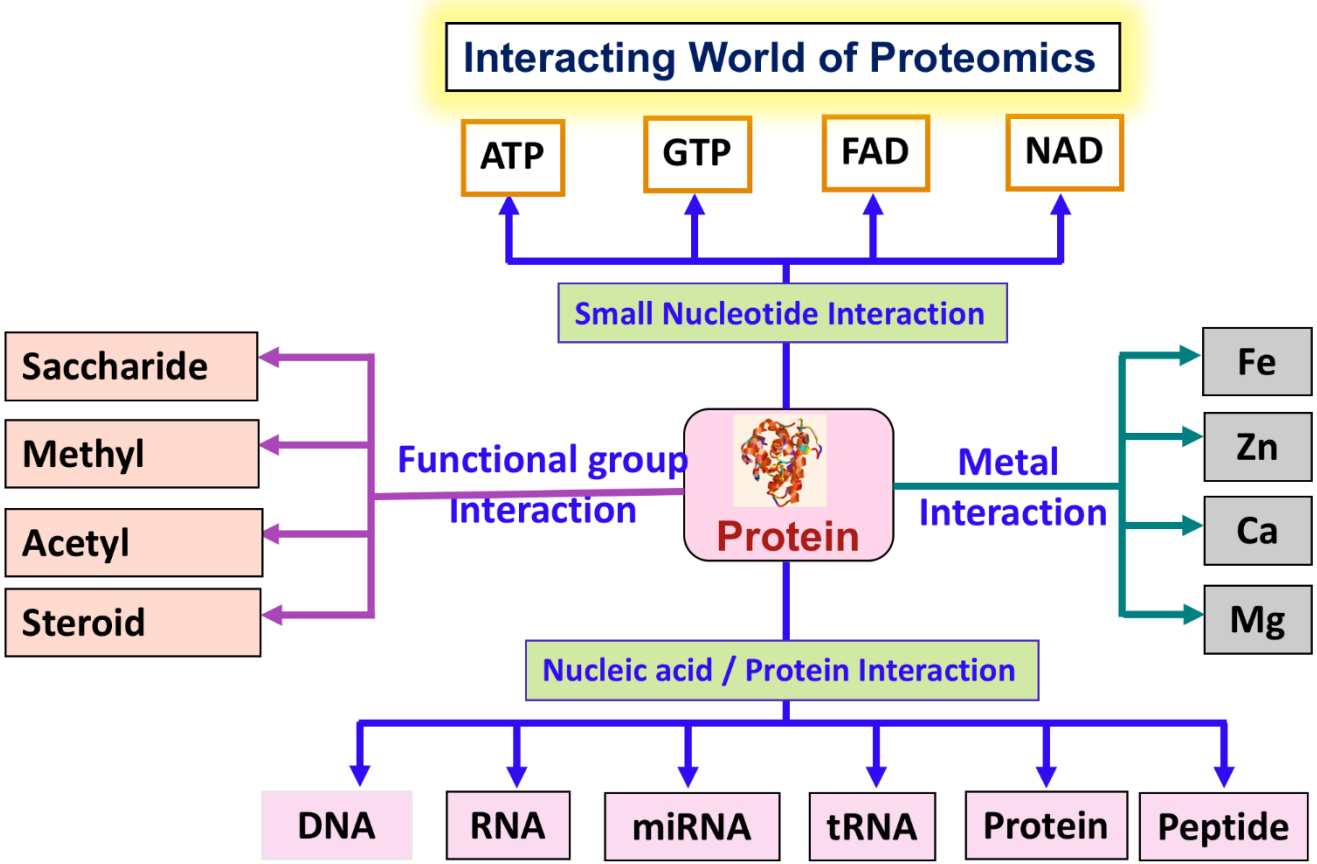
Prediction of interaction between biomolecules

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Dr. Jaspreet Kaur Dhanjal
Assistant Professor, Department of Computational Biology
Email ID: jaspreet@iiitd.ac.in

November 07, 2025

Introduction

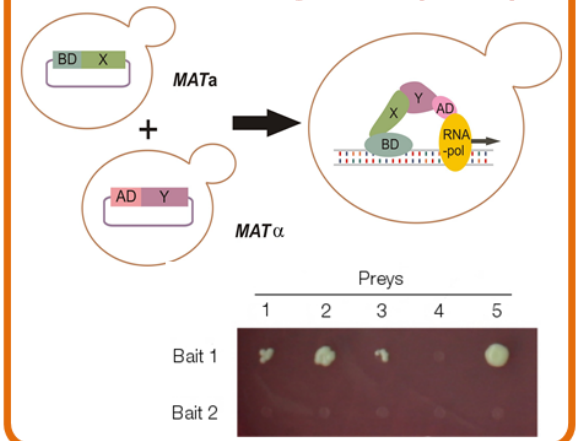


Experimental methods for finding protein interactions

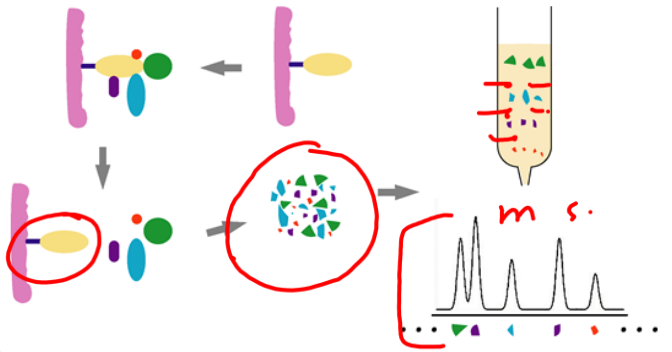
High-throughput

Low-throughput

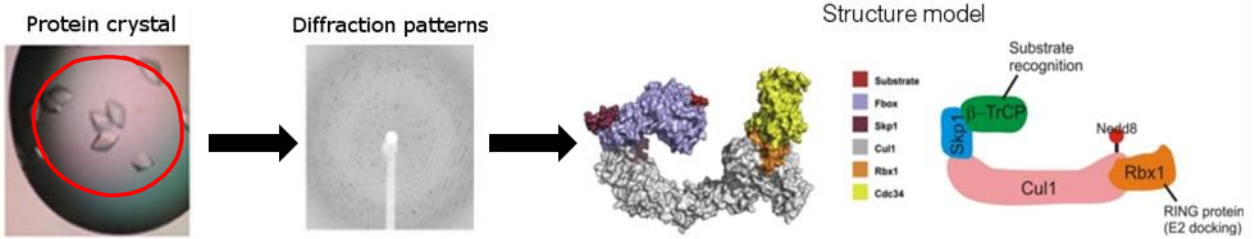
Yeast-two hybrid (Y2H)



Tandem affinity purification+ mass spectrometry (TAP-MS)



X-ray diffraction studies



Protein-protein interaction prediction

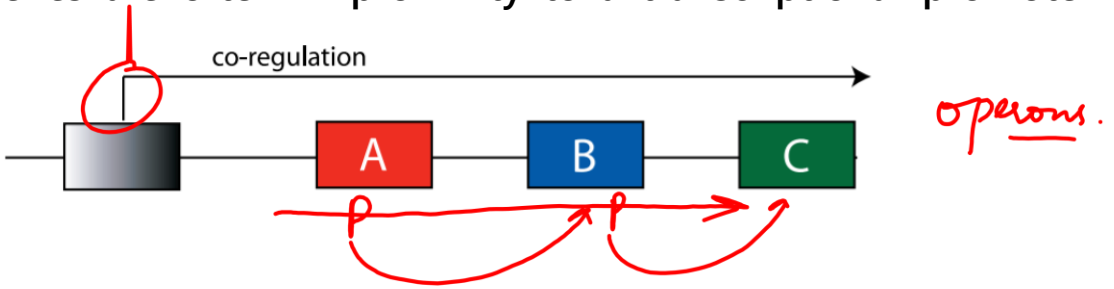
Different methods:

1. Gene cluster or gene neighborhood method
2. Rosetta stone method
3. Phylogenetic profile
4. Sequence-based co-evolution
5. Homology based inference
6. Association of structural motifs
7. Protein-protein docking
8. Machine learning-based methods

Gene cluster or Gene neighborhood method

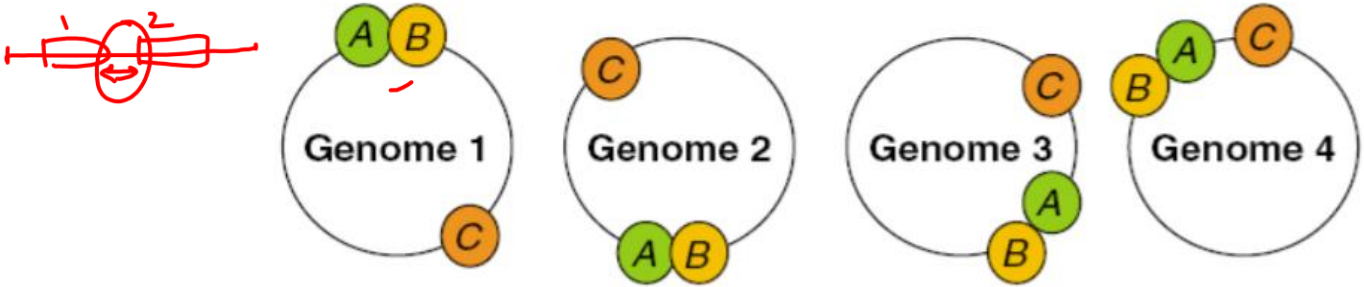
Gene cluster

Within bacteria, proteins of closely related function are often transcribed from a single functional unit known as operon. Operons contain two or more closely spaced genes located on the same DNA strand. These genes are often in proximity to a transcriptional promoter that regulates operon expression.



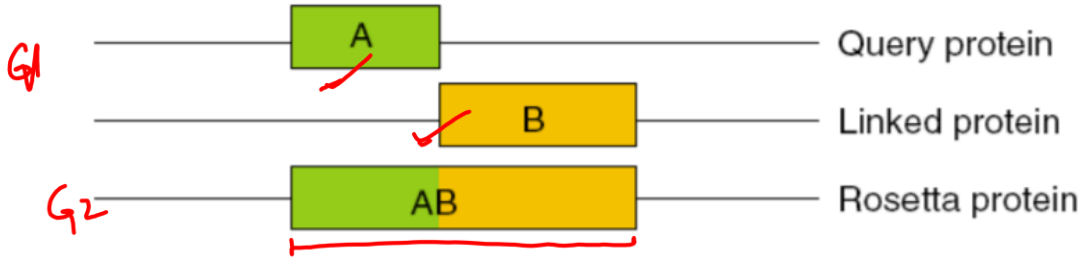
operons.

Gene neighborhood



Gene A is a neighbor of B in several genomes - potential functional link

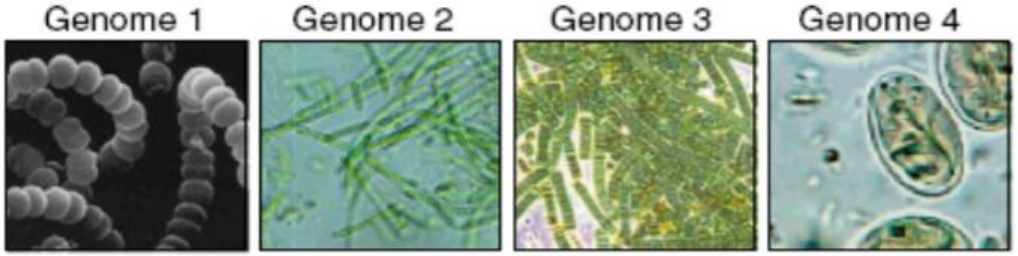
Rosetta stone method



- A, B – two domains that might interact
- Rosetta protein – protein containing both the domains together in an organism – indicates that these domains present as two different proteins in other organism may interact

Phylogenetic profile

Pairs of non-homologous proteins that are always both present or both absent in a genome suggest their functional dependence → possible interaction



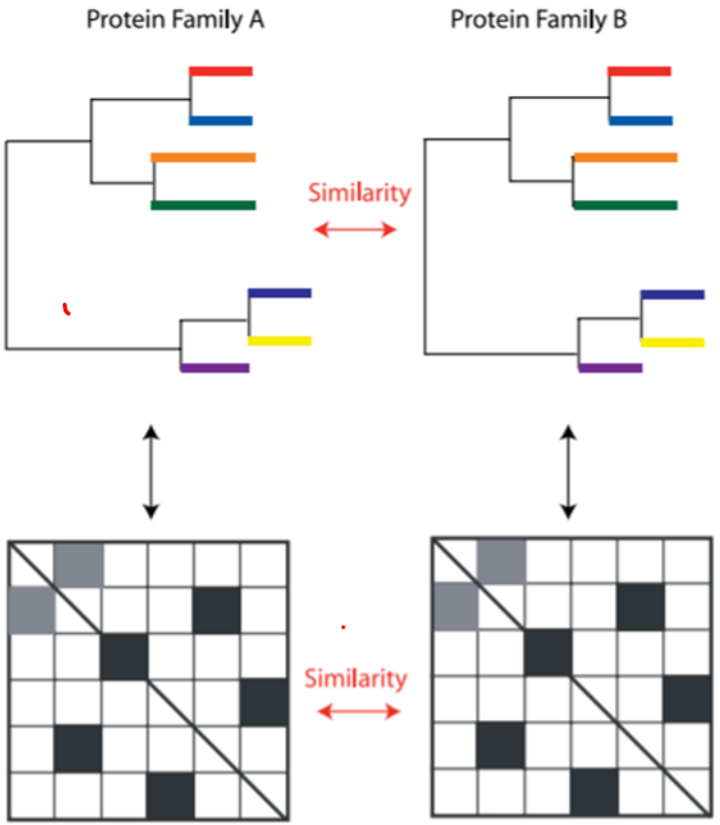
	Genome 1	Genome 2	Genome 3	Genome 4
Protein A	1	1	1	0
Protein B	1	1	1	0
Protein C	1	0	1	1
Protein D	1	1	0	1

	Genomes		
Proteins	EC	HI	BS
P1	0	1	1
P2	0	0	1
P3	1	0	0
P4	0	1	1

P1 and P4 are functionally linked

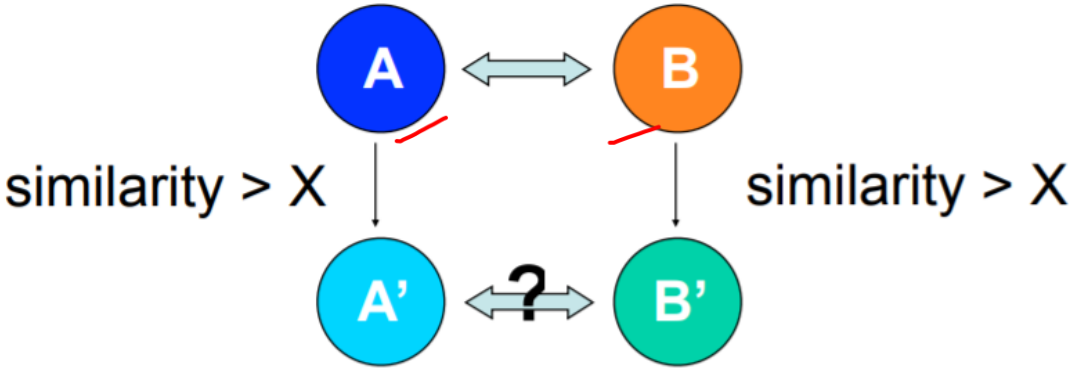
Sequence-based co-evolution

- Assuming that protein A and B interact.
- If both A and B are present in several organisms and perform the same role in these organisms they interact in all these organisms
- Evolution of A and B should be correlated



Template based prediction of protein-protein interaction

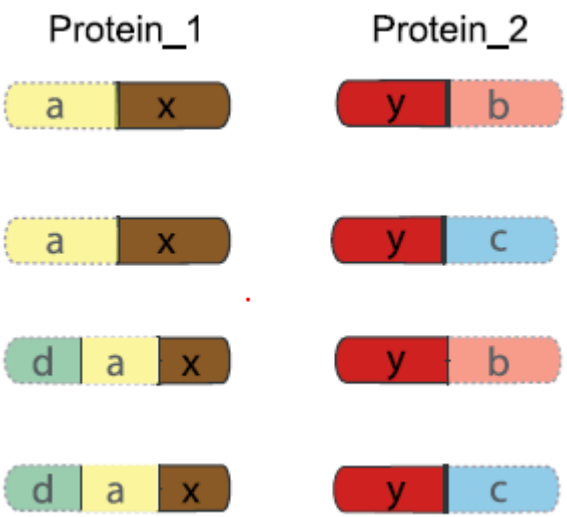
Homology based inference of protein-protein interface



A-B known experimentally
A'-B' inferred by homology

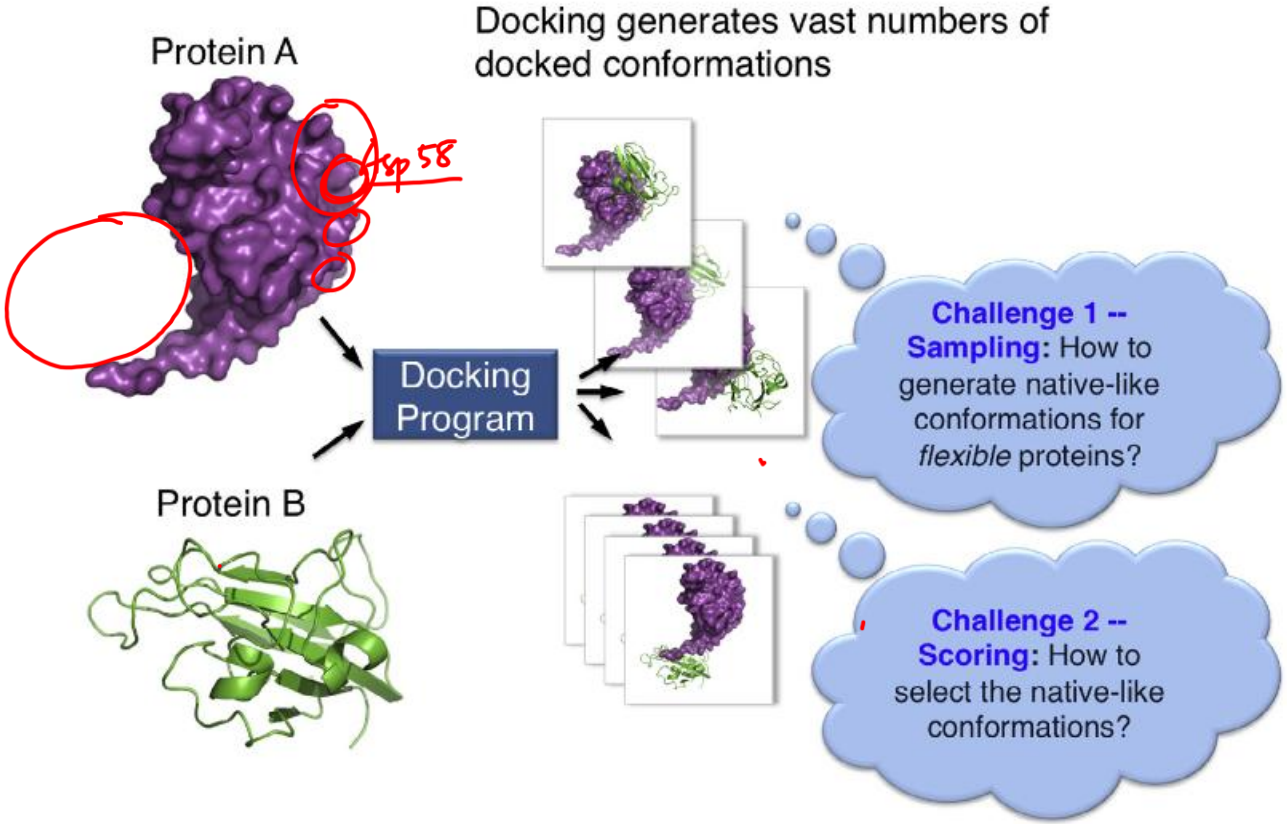
Association of structural motifs

- This methods looks for the characteristic sequence or structural motifs which distinguish
- interacting proteins from non-interacting.
- Correlated sequence signatures, or domains, that are found together more often than expected by chance are used as markers to predict a new type of protein interaction.
- For eg., protein interaction data is used to compute log-odds scores and to find correlated domains. The log-odds score is computed as:

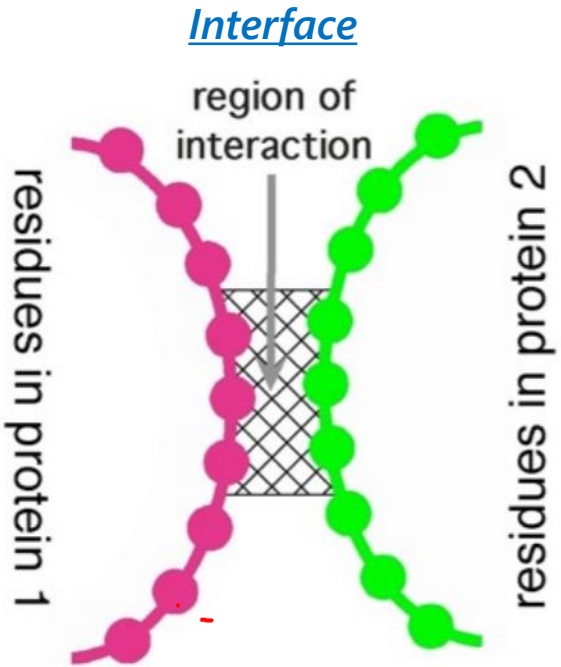


log2(Pij /PiPj), where Pij is the observed frequency of domains I and j occurring in one protein pair; Pi and Pj are the background frequencies of domains i and j in the data.

Protein-protein docking



Machine learning-based methods



Based on the required input of the predictors, machine learning interface predictors can be further classified into:

1) ***Structure-based methods***
   requiring information derived from 3D protein structures or models of the component proteins as input

2) ***Sequence-based methods***
   requiring only protein sequences as input

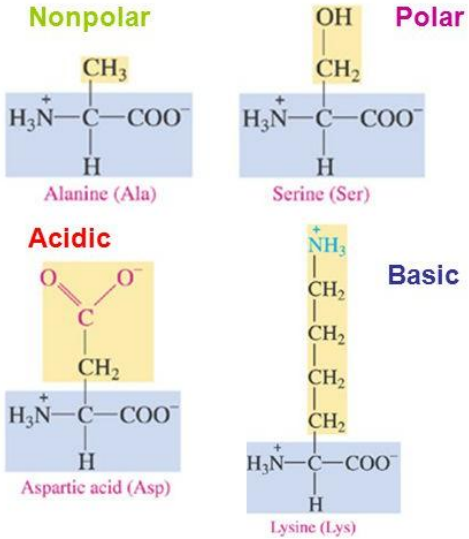
Machine learning-based methods

Characteristics of an interface

1) ***Amino acid types***: Each type of commonly occurring amino acid can be represented as a binary vector of size 20 by 1. For example, alanine can be represented as [1, 0, 0, 0, …. ,0].

Amino acids are classified as

- **nonpolar** (hydrophobic) with hydrocarbon side chains.
- **polar** (hydrophilic) with polar or ionic side chains.
- **acidic** (hydrophilic) with acidic (-COOH) side chains.
- **basic** (hydrophilic) with -NH_2 side chains.



Machine learning-based methods

Characteristics of an interface

2) ***Physicochemical properties***:
Commonly used physicochemical properties are hydrophobicity, charge and *van der* Waals volume.

A database of numerical indices representing various physicochemical properties of amino acids and pairs of amino acids is provided in AAindex.

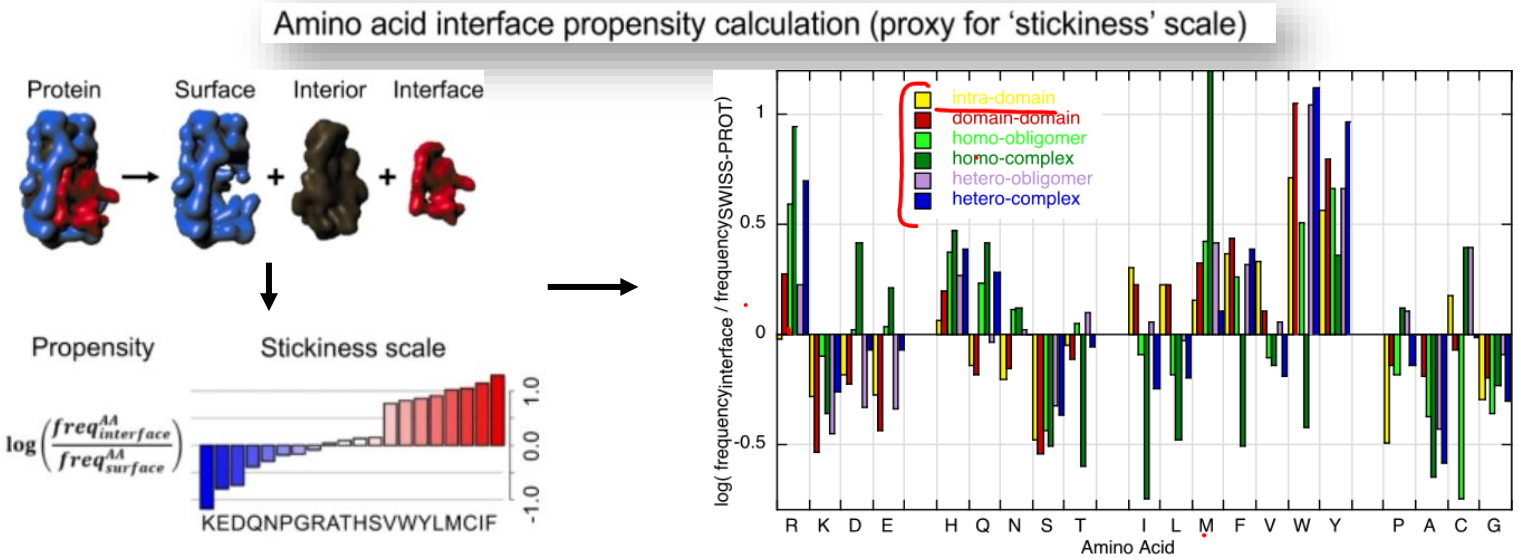
Amino Acid Code	Physicochemical Property (*cf.* Equation (11)) [a]						
	$\Phi^{(1)}$	$\Phi^{(2)}$	$\Phi^{(3)}$	$\Phi^{(4)}$	$\Phi^{(5)}$	$\Phi^{(6)}$	$\Phi^{(7)}$
	H1	H2	V	P1	P2	SASA	NCI
A	0.62	−0.5	27.5	8.1	0.046	1.181	0.007187
C	0.29	−1	44.6	5.5	0.128	1.461	−0.03661
D	−0.9	3	40	13	0.105	1.587	−0.02382
E	−0.74	3	62	12.3	0.151	1.862	0.006802
F	1.19	−2.5	115.5	5.2	0.29	2.228	0.037552
G	0.48	0	0	9	0	0.881	0.179052
H	−0.4	−0.5	79	10.4	0.23	2.025	−0.01069
I	1.38	−1.8	93.5	5.2	0.186	1.81	0.021631
K	−1.5	3	100	11.3	0.219	2.258	0.017708
L	1.06	−1.8	93.5	4.9	0.186	1.931	0.051672
M	0.64	−1.3	94.1	5.7	0.221	2.034	0.002683
N	−0.78	2	58.7	11.6	0.134	1.655	0.005392
P	0.12	0	41.9	8	0.131	1.468	0.239531
Q	−0.85	0.2	80.7	10.5	0.18	1.932	0.049211
R	−2.53	3	105	10.5	0.291	2.56	0.043587
S	−0.18	0.3	29.3	9.2	0.062	1.298	0.004627
T	−0.05	−0.4	51.3	8.6	0.108	1.525	0.003352
V	1.08	−1.5	71.5	5.9	0.14	1.645	0.057004
W	0.81	−3.4	145.5	5.4	0.409	2.663	0.037977
Y	0.26	−2.3	117.3	6.2	0.298	2.368	0.023599

[a] H1, hydrophobicity; H2, hydrophilicity; V, volume of side chains; P1, polarity; P2, polarizability; SASA, solvent accessible surface area; NCI, net charge index of side chains.

Machine learning-based methods

Characteristics of an interface

3) ***Interface propensity***: The different physicochemical properties of amino acids result in differential interaction propensities. The higher the interface propensity, the more likely an amino acid is to appear on the interface as opposed to elsewhere on the protein surface. Such propensities are usually derived from an analysis of known structures in the PDB.



Machine learning-based methods

Characteristics of an interface

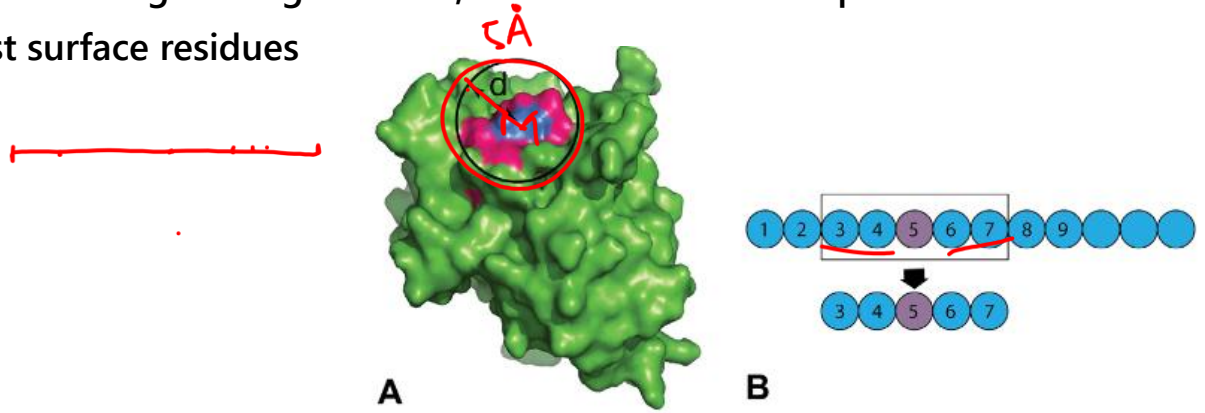
3) ***Interface propensity***: Coexistence of residues at the interface is also evaluated.

For each target residue in a given protein structure, a set of neighboring residues (spatial neighbors) on the protein surface (surface patch) can also be calculated.

There are two common ways to define a surface patch:

(A) based on a fixed radius, in which the surface patch consists of the target residue and any surface residues within a fixed radius from the target residue;

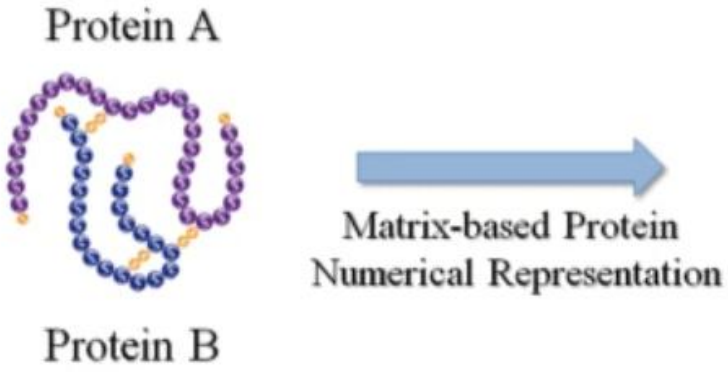
(B) based on a fixed number of neighboring residues, in which the surface patch consists of the target residue and its *K* nearest surface residues



Machine learning-based methods

Characteristics of an interface

4) ***Evolutionary information***: Interfacial residues are important functional sites and tend to be conserved among homologs or undergo correlated mutations. There are different ways to encode sequence conservation, and a widely used approach is to construct PSSMs (Position Specific Scoring Matrices) from multiple sequence alignments (MSAs).



Machine learning-based methods

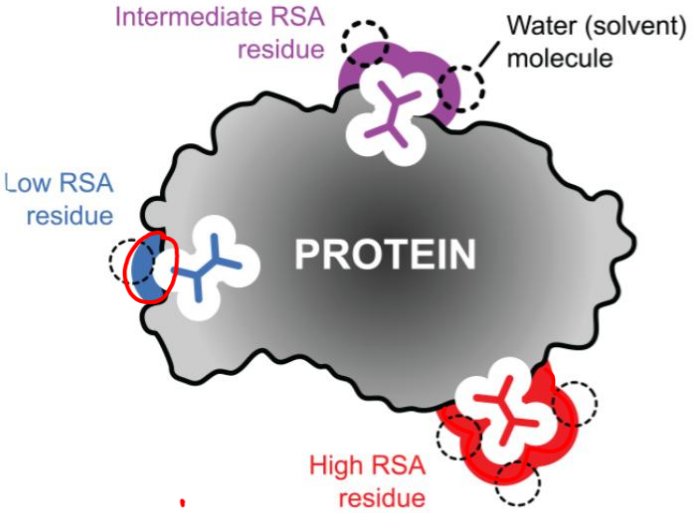
Characteristics of an interface

5) ***Relative solvent accessibility***: Most proteins recognize and interact with other proteins through their surface residues (i.e., residues with relatively high solvent accessible surface area) unless the interacting proteins undergo large conformational changes upon binding.

Relative accessible surface area, $RASA = \frac{ASA\_residue\_in\_protein}{ASA\_free\_residue}$

where, *ASA_residue_in_protein* is accessible surface area of the residue in the protein structure, and *ASA_free_residue* is the accessible surface area of this residue in a "free" state.

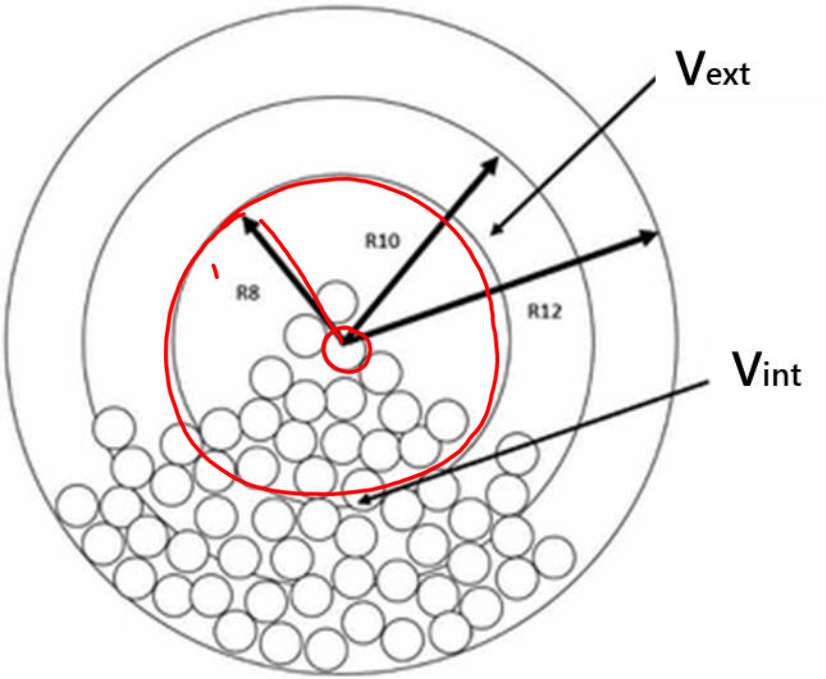
A residue is generally regarded as a surface residue if its RASA is larger than 5%. Solvent accessibility of a residue in a protein can be calculated using software like STRIDE.



Machine learning-based methods

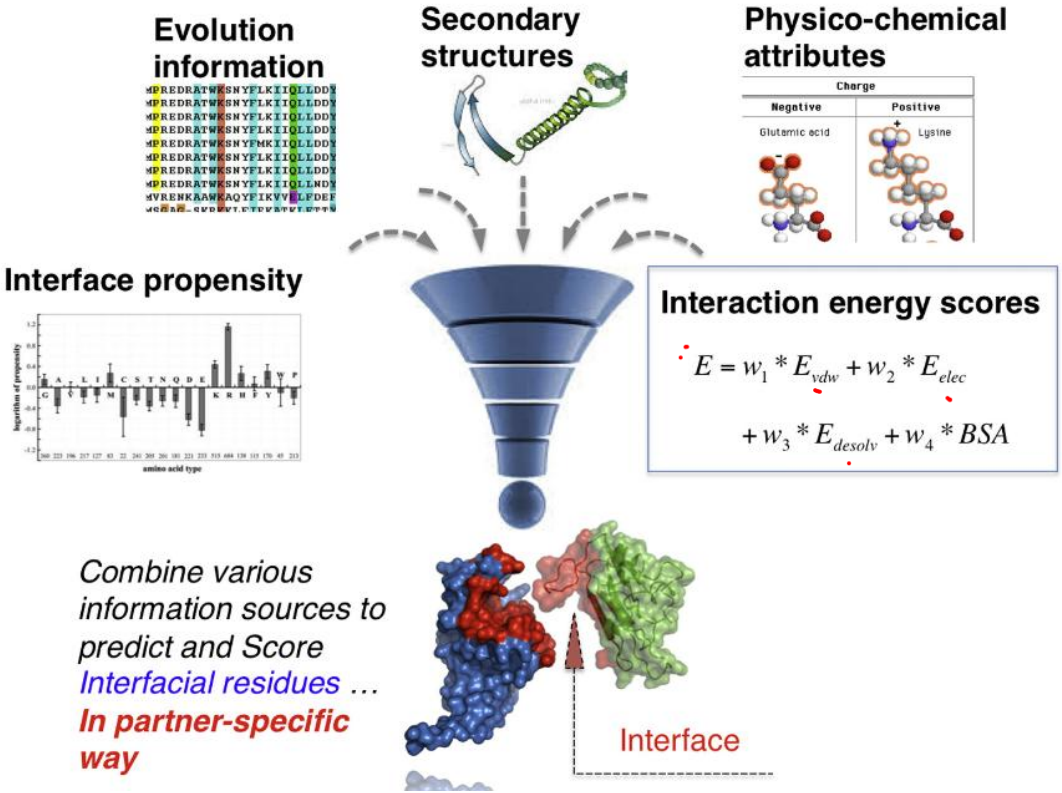
Characteristics of an interface

6) ***Surface shape:*** The shape of a protein surface is also a useful indicator of interacting sites. One widely used measure for the concavity or convexity of the neighborhood of an atom in a protein is the CX value. To calculate the CX value of an atom, a sphere is centered on the target atom, and CX=Vext/Vint, where Vint is the volume occupied by the protein, and Vext is the free volume in the sphere.



Machine learning-based methods

3D structure based protein interface prediction



Protein-protein interaction prediction

Different methods:

1. Gene cluster or gene neighborhood method
2. Rosetta stone method
3. Phylogenetic profile
4. Sequence-based co-evolution
5. Homology based inference
6. Association of structural motifs
7. Protein-protein docking
8. Machine learning-based methods