Instructions for the group exercise on Gaussian distribution

Binding interactions between biomolecules are the basis of all complex intracellular networks

Molecular binding is used at inside living cells multiple levels, Each level (see Figure 1) is characterized by its its own molecular interaction network:

- Regulatory network (top level): regulation of gene expression mediated by binding between proteins (transcription factors) and DNA
- Protein-Protein Interaction network (middle level): binding interactions between proteins. Used for signaling as well as formation of multi-protein molecular machines
- Metabolic network (bottom level): Binding interactions between proteins (enzymes) and small molecules (metabolites)

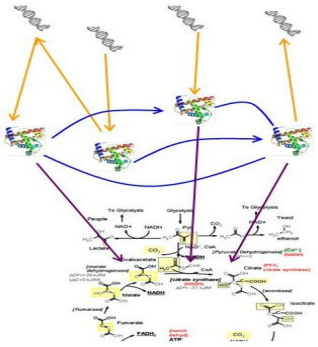
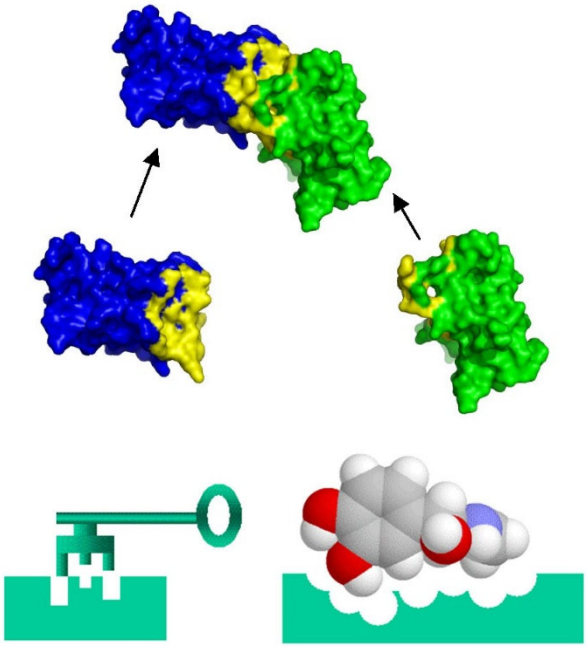


Figure 1 Three levels of intracellular biomolecular networks

Energy of Protein-Protein Binding Interactions

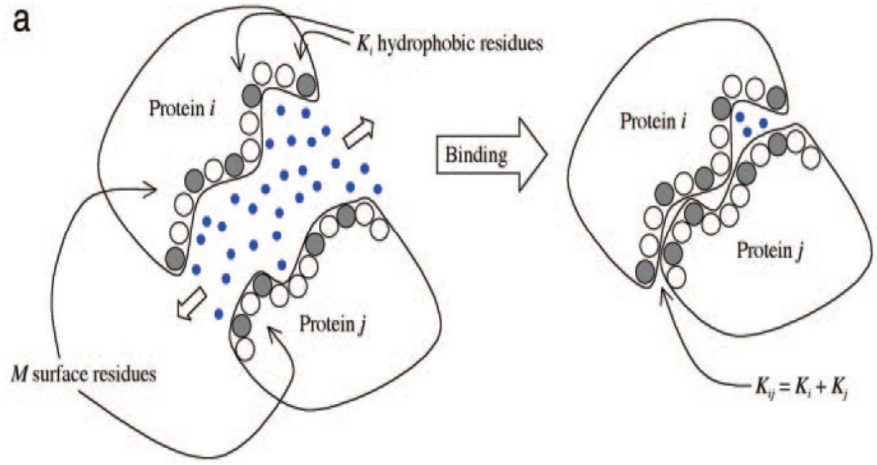
Proteins and other biomolecules (metabolites, drugs, DNA) can bind each other either specifically (strong binding, large absolute value $|E_{ij}|$ of negative binding energy E_{ij}) or non-specifically (weak binding, small absolute value $|E_{ij}|$ of negative binding energy E_{ij}).

Specific bindings are described by Lock-and-Key theory.



Non-specific bindings are mediated by random contacts.

Most of the binding energy is explained by hydrophobic amino acid residues shielded from water.



Since the number of amino acids (residues) at the contact interface is large and their hydrophobicities are independent, one expects binding energy to follow Gaussian distribution: $PDF(E_{ij}) \sim \exp(-(E_{ij}-\mu)^2/2\sigma^2)$. Indeed, E_{ij} – sum of hydrophobicities of many independent residues

Assignment:

- In Matlab load the file PINT_binding_energy.mat with binding_energy variable giving the binding energy E_{ij} (in units of kT at room temperature) for 430 pairs of interacting proteins from human, yeast, etc. Data was collected from the PINT database and analyzed in J. Zhang, S. Maslov, E. Shakhnovich, Molecular Systems Biology (2008).
- Calculate mean and standard deviation of binding energies as we did in class: mean_bidning_energy=sum(r2)/Stats; std_bidning_energy =sum((??-??))/??. What is the analog of r2 and Stats in this exercise?
- Plot a PDF of E_{ij} using [counts, bins]=hist(binding_energy); Use linear x-axis and logarithmic y-axis.
- Fit Gaussian (Normal) distribution to E_{ij} using dfittool(binding_energy) command in Matlab
    - Take snapshots of different properties by changing "Display Type" dropdown menu
    - Fit Gaussian distribution without any exclusions
    - Use "Exclude" button to generate the new exclusion rule to drop all points with X<-23 from the fit
    - Use "New Fit" button to generate a new "Normal" fit with the exclusion rule you just created
    - Select "probability plot" from "Display type" dropdown menu to evaluate the quality of the plot. Where does the probability plot deviate from a straight line?
- Can you propose a plausible biological explanation for deviations of very strong binding energies from the Gaussian?