

Dynamical Analysis of Networks: How to Identify Important Nodes with Applications to Protein Engineering



Yi Mao
National Institute for
Mathematical and Biological Synthesis (NIMBioS)
University of Tennessee

DIMACS/CCICADA Workshop
on Stochastic Networks: Reliability, Resiliency, and Optimization

October 13, 2011

- ◆ Introduction to proteins and protein modeling
- ◆ Mathematical framework for network modeling
- ◆ Luciferase bioluminescence

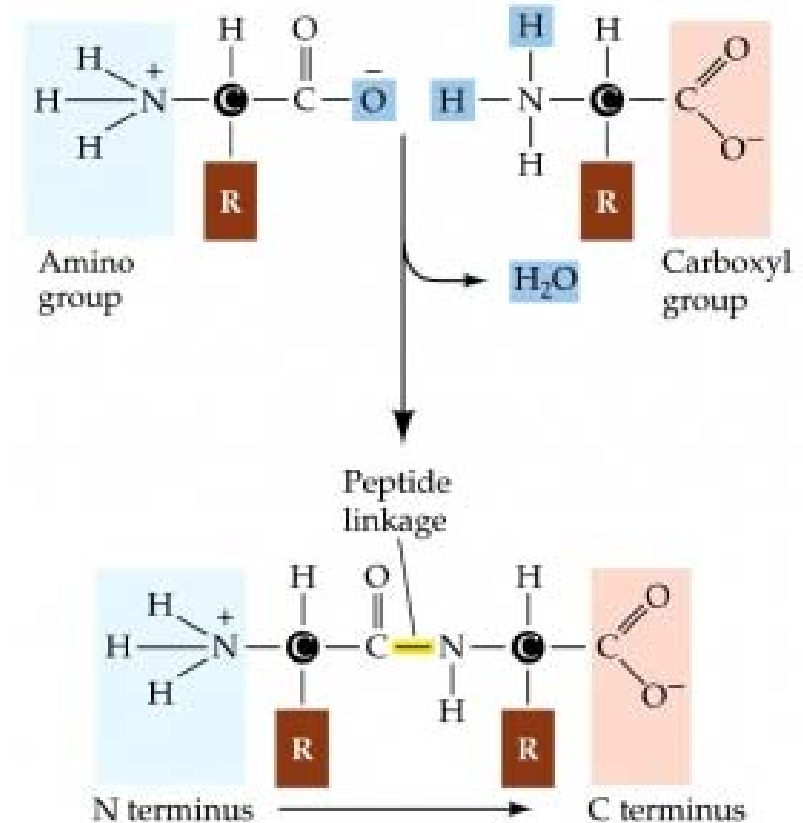
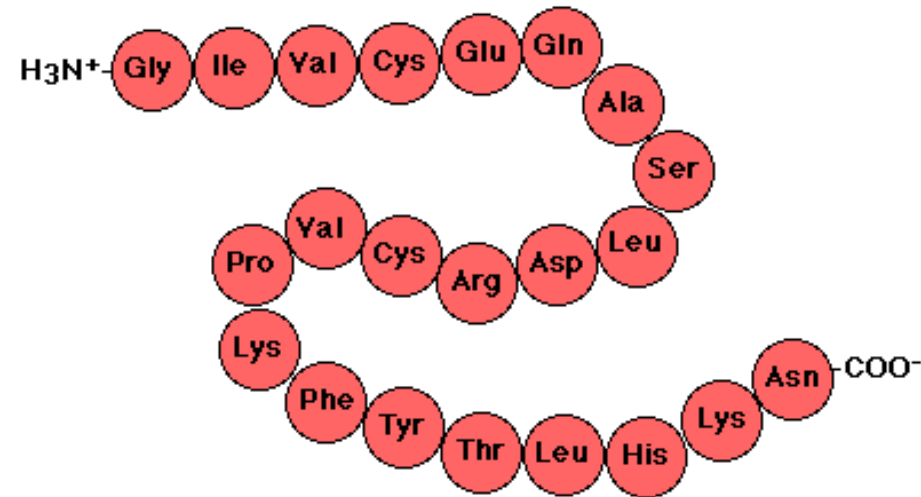
Mathematical challenge in biology: a lesson in complexity

- ◆ High dimensionality
- ◆ Nonlinearity
- ◆ Stochasticity

Introduction to proteins

Protein sequence

- Sequence: the order of amino acids (20 amino acids)

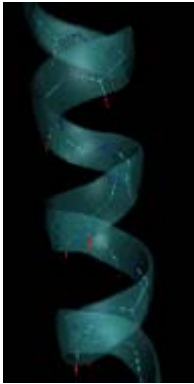


- Mutation: change in sequence

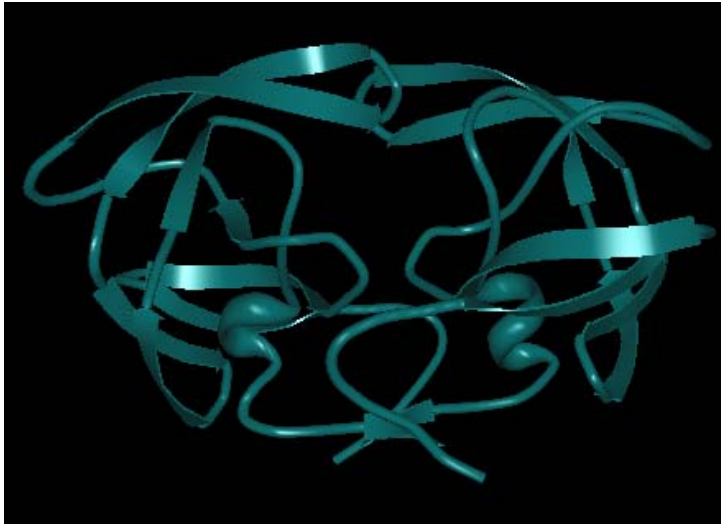
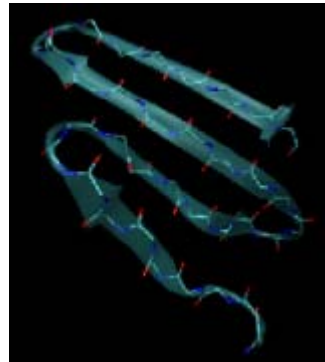
Protein structure

- Secondary structure

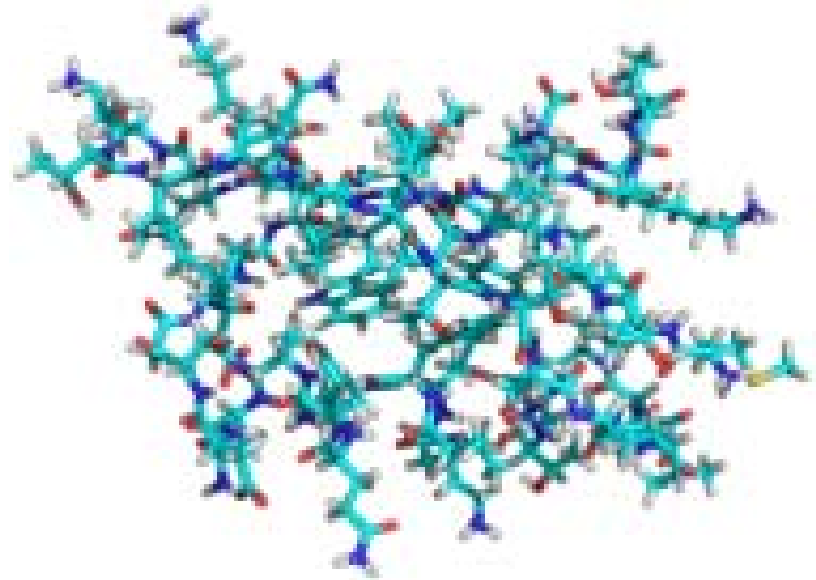
α helix



β sheet

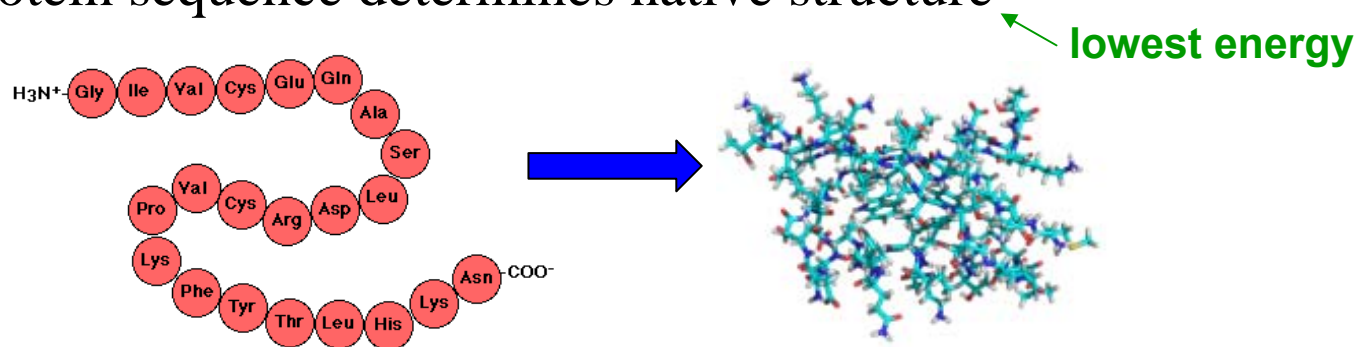


- Tertiary structure



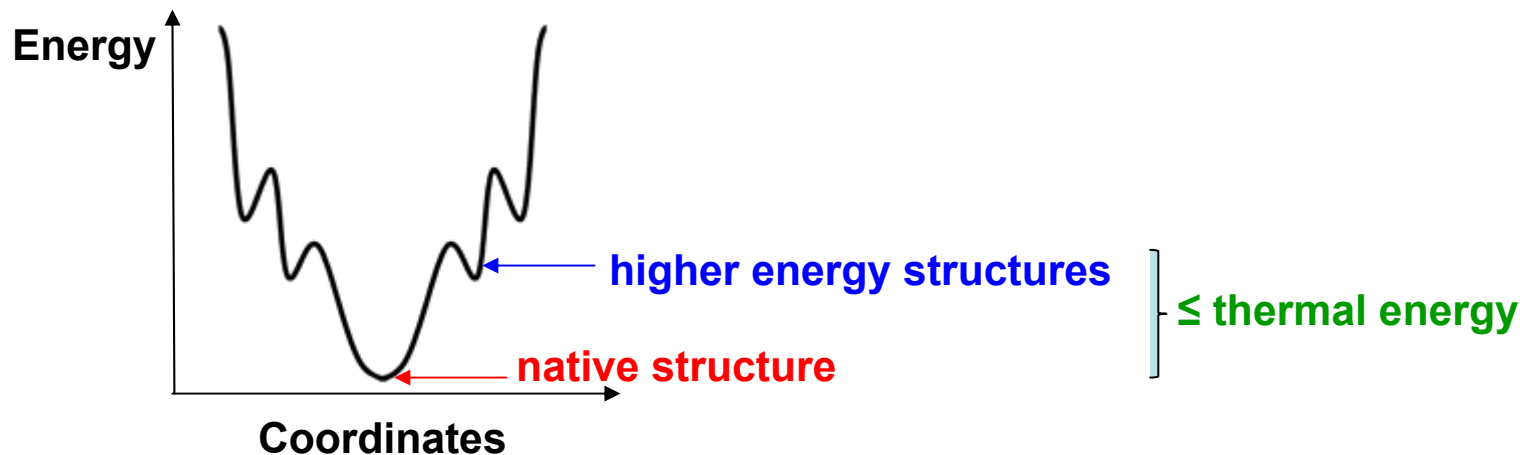
Protein folding: from sequence to structure

- Protein sequence determines native structure



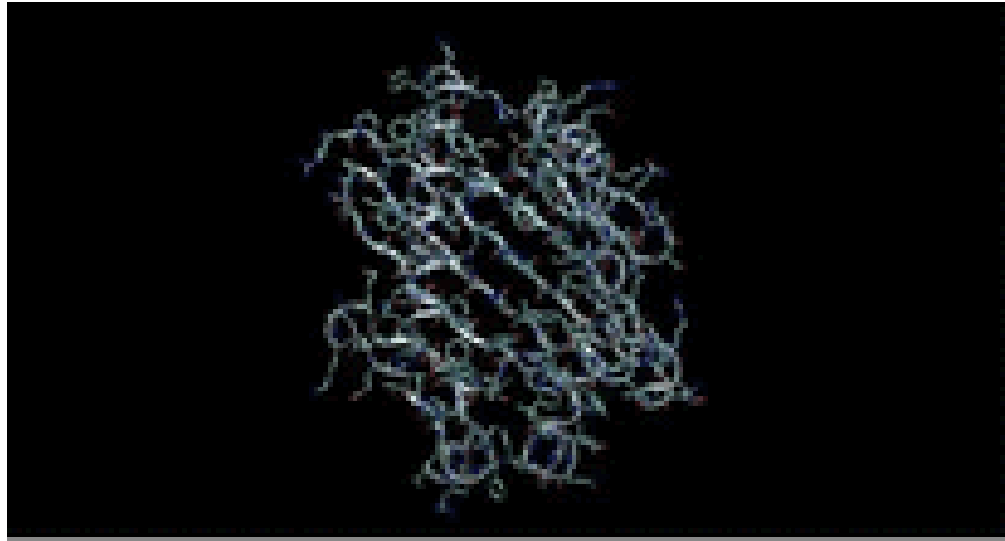
- Computational approach:

modeling the atomic interactions and sampling different structures



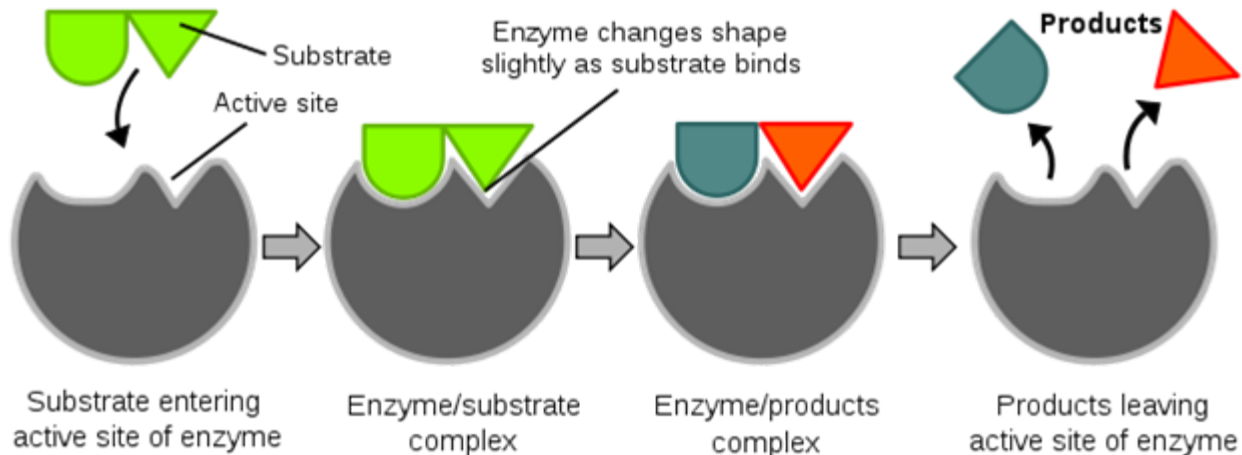
Protein dynamics

- Proteins occupy an ensemble of conformations at room temperature



Protein function

- ◆ Proteins are the most functionally diverse molecules in living organisms
- ◆ Proteins function by binding to other molecules (ligands, proteins).
- ◆ Enzyme proteins catalyzes the chemical reactions by binding to reactant (substrate).



Active site crucial to enzymatic activity

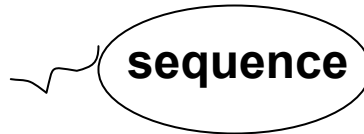
- ◆ The major targets of prescription drugs are proteins.

Biological question: protein sequence ↔ function

- change in sequence → what's the change in function?
- how to change sequence ? ← desirable function

◆ Experimental approach: mutagenesis experiments

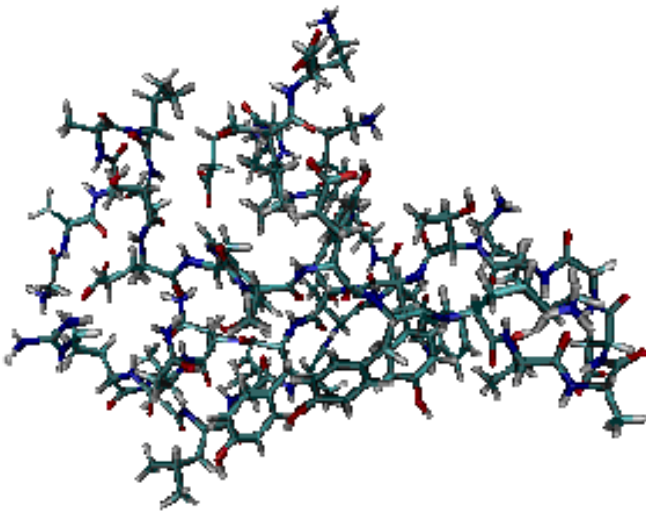
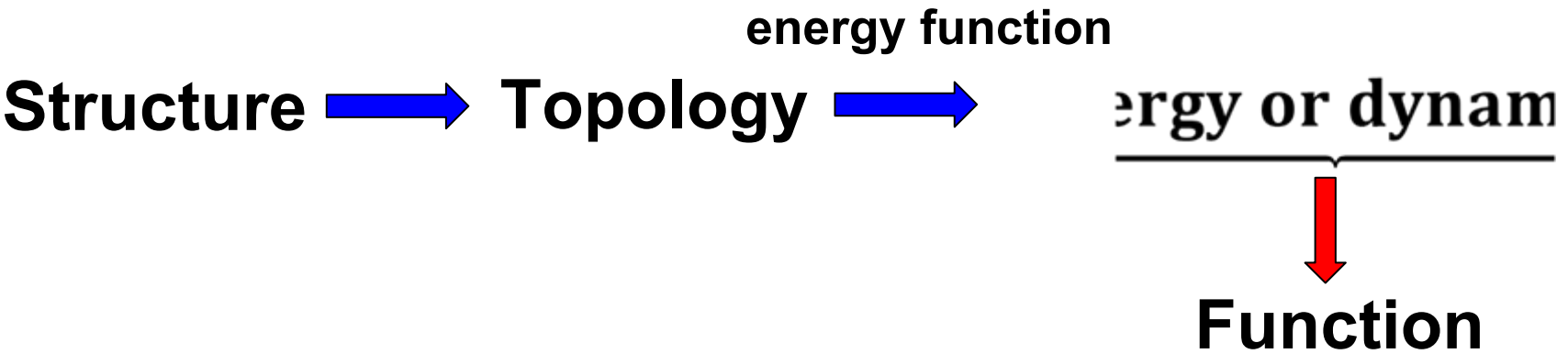
Mutation



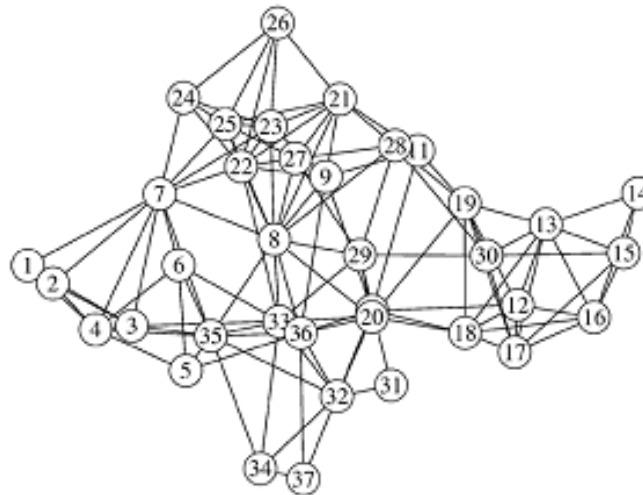
Replace one residue type by another

limits: lack of details and adequate sampling

Structure modeling of protein



Atomistic model



Elastic network model

Structure based modeling of protein: bottom-top approach

- ◆ Protein as a system of interacting components
atoms or residues?

- ◆ Protein modeled by classical mechanics

$$m_i a_i = \frac{-\partial V}{\partial r_i}$$

- ◆ Modeling starts with structure, generates information
about **dynamics** and **energetics**.



structural details
of a biological process



quantitative measure
of molecular interactions

Use protein modeling to address: protein sequence ↔ function

- change in sequence → what's the change in function?

Answer: protein function is computed as a chemical or physical property (based on energetics)

- how to change sequence ? ← desirable function

Use protein modeling to address: protein sequence $\xrightarrow{?}$ function

- change in sequence \rightarrow what's the change in function?

Answer: protein function is computed as a chemical or physical property (based on energetics)

- ~~how to change sequence \leftarrow desirable function?~~

how to identify important residues?

Answer: Important residues **interact** strongly with protein's functional sites.



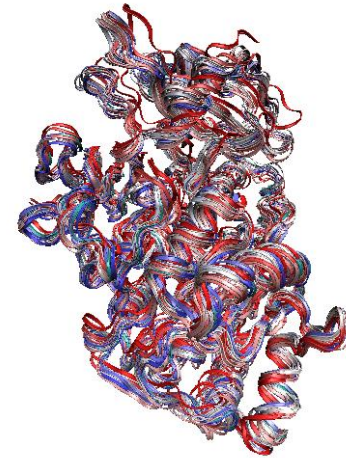
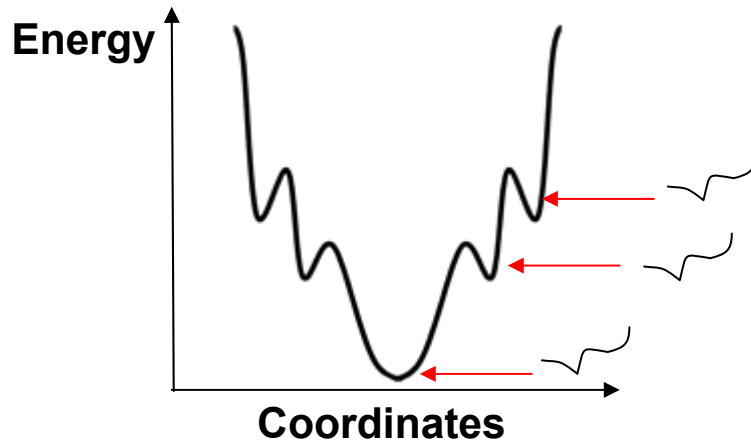
dynamic correlations

Structure vs dynamics

◆ For a given sequence

- Protein native structure \leftrightarrow energy minimum
- Protein dynamics \leftrightarrow shape of energy function

◆ Mutation effects may affect the shape of energy function



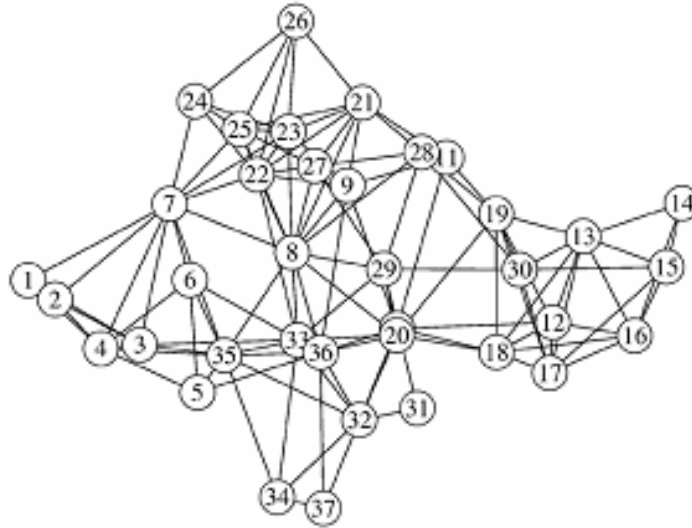
Protein dynamics

◆ Computation of protein dynamics

- How to sample the protein energy space efficiently?

Elastic network modeling of proteins

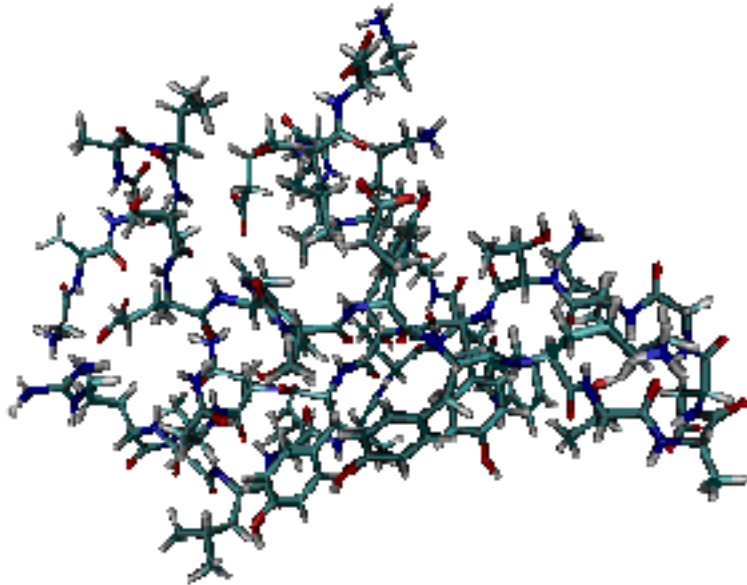
Elastic network model



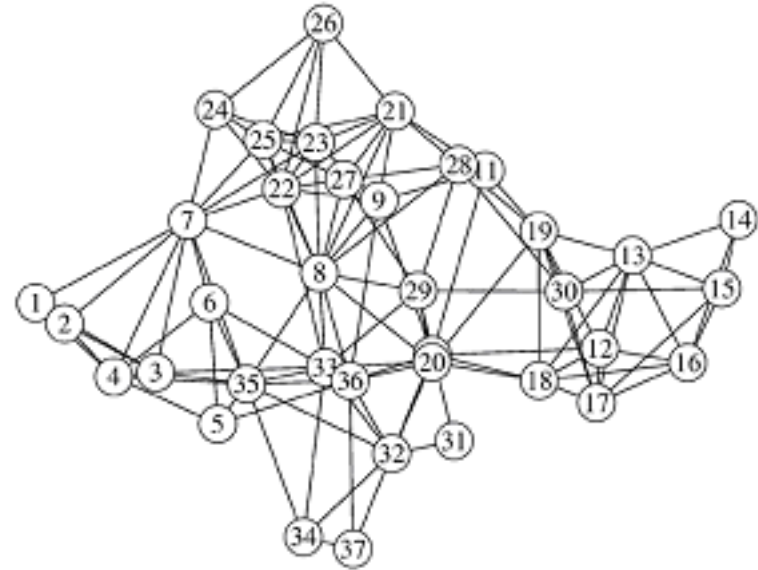
- Starts with protein crystal structure
- Each amino acid is replaced by node (C α atom)
- Any pair within a cutoff distance is governed by

$$V = \frac{1}{2} C (d_{ij} - d_{ij}^0)^2$$

Atomistic model vs elastic network model



Atomistic model



Elastic network model

◆ Elastic network model is

- computationally simple
- energetic inaccurate

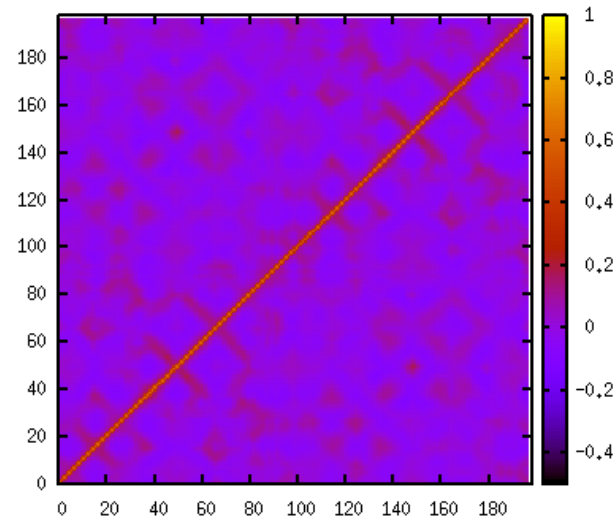
Dynamic measures

- Fluctuation and covariance $R_{ij} = \langle (R_i - \langle R_i \rangle)(R_j - \langle R_j \rangle) \rangle$

- Correlation
$$C_{ij} = \frac{\langle R_{ij} \rangle}{[\langle R_{ii} \rangle \langle R_{jj} \rangle]^{1/2}}$$

- Correlation matrix

$$\begin{bmatrix} C_{11} & C_{12} & C_{13} & 6 \\ C_{21} & C_{22} & C_{23} & 6 \\ C_{31} & C_{32} & C_{33} & 6 \\ 7 & 7 & 7 & 9 \end{bmatrix}$$



ENM: protein dynamics in a closed form

$$R_{ij} = \langle (R_i - \langle R_i \rangle)(R_j - \langle R_j \rangle) \rangle = \left(\frac{1}{Z}\right) \int (R_{ij}) e^{-E/k_B T} d\Delta R$$

$$= \frac{k_B T}{C} [H^{-1}]_{ij}$$

- Hessian matrix H**

$$\begin{bmatrix} H_{11} & 6 & H_{1N} \\ 7 & 9 & 7 \\ H_{N1} & 6 & H_{NN} \end{bmatrix} \quad H_{ij} = \begin{bmatrix} \frac{\partial^2 E}{\partial x_i \partial x_j} & \frac{\partial^2 E}{\partial x_i \partial y_j} & \frac{\partial^2 E}{\partial x_i \partial z_j} \\ \frac{\partial^2 E}{\partial y_i \partial x_j} & \frac{\partial^2 E}{\partial y_i \partial y_j} & \frac{\partial^2 E}{\partial y_i \partial z_j} \\ \frac{\partial^2 E}{\partial z_i \partial x_j} & \frac{\partial^2 E}{\partial z_i \partial y_j} & \frac{\partial^2 E}{\partial z_i \partial z_j} \end{bmatrix}$$

From elastic network model $\frac{\partial^2 E}{\partial x_i \partial y_j} = \frac{-c(x_i - x_j)(y_i - y_j)}{d_{ij}^2}$

- Normal mode analysis** $[H^{-1}]_{ij} = \sum_k \left[\frac{v_k v_k^T}{\lambda_k} \right]_{ij}$

v_k and λ_k are the eigenvector and eigenvalue of mode k

Comparison with experiments

- Debye-Waller factor: $\beta = \frac{8\pi^2}{3} \langle R_{ii} \rangle$

Figure 1

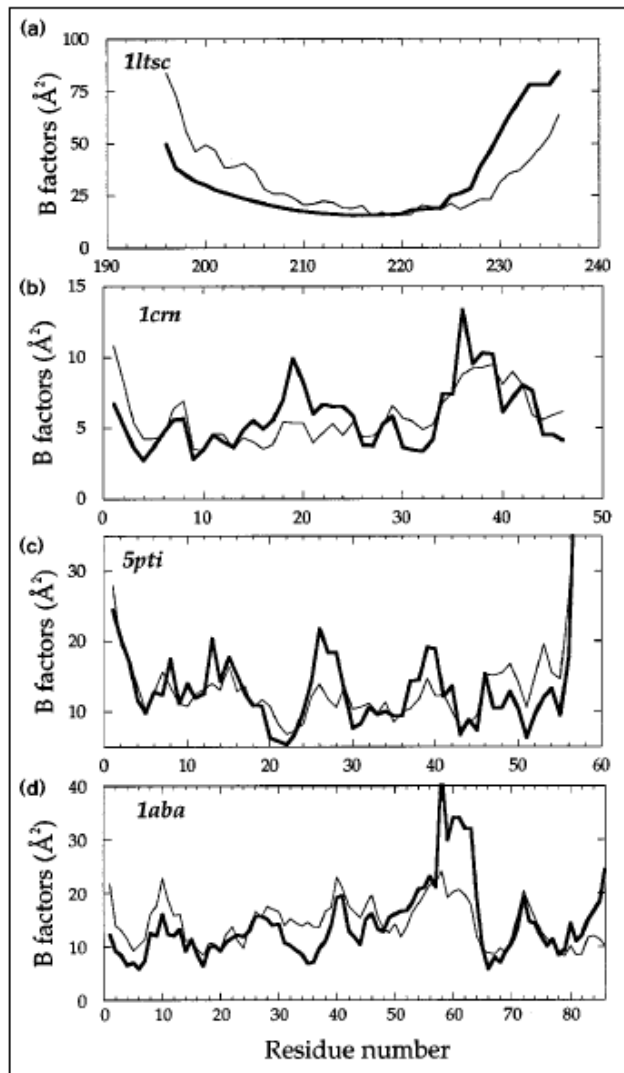
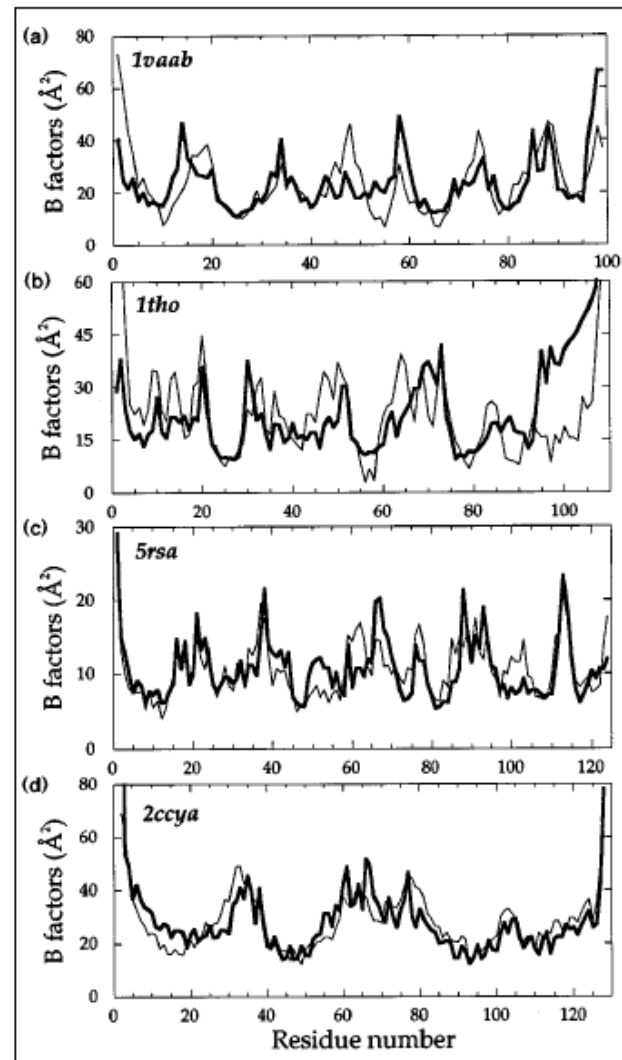


Figure 2



— Exp.
— Theory

Use protein dynamics to identify important residues

◆ **Fact:** Proteins carry out function by binding to other molecules.

◆ **Hypotheses**

- Functionally important residues interact strongly with the functional sites.
- Residues involved in the conformational changes of binding are functionally important.

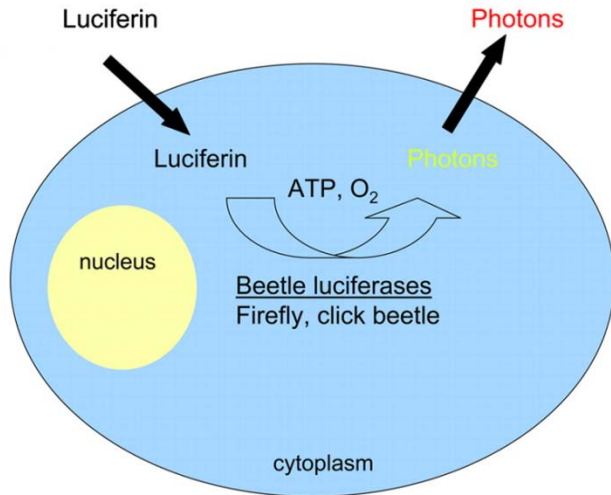
◆ **Advantage:** Mechanistic understanding of catalytic reaction is not necessary.

How sequence distribution of luciferase affects the color emission of bioluminescence

Y. Mao, “Dynamics Studies of Luciferase Using Elastic Network Model: How the Sequence Distribution of Luciferase Determines its Color”
Protein Engineering Design & Selection 2011, 24: 341-349.

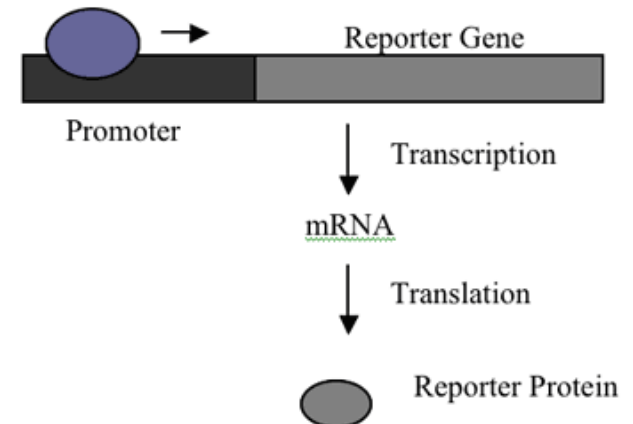
Bioluminescence

- Conversion of chemical energy into light

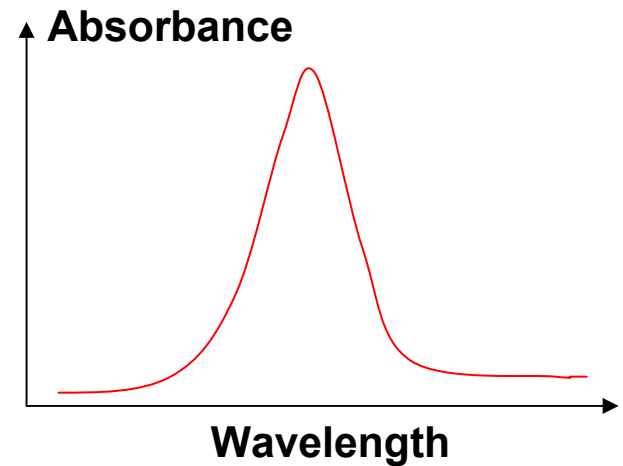
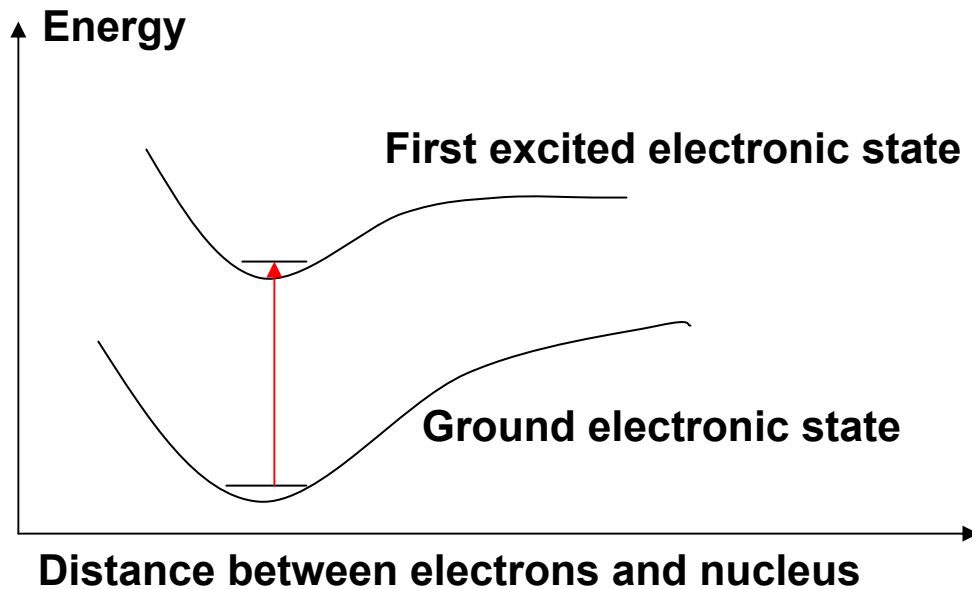
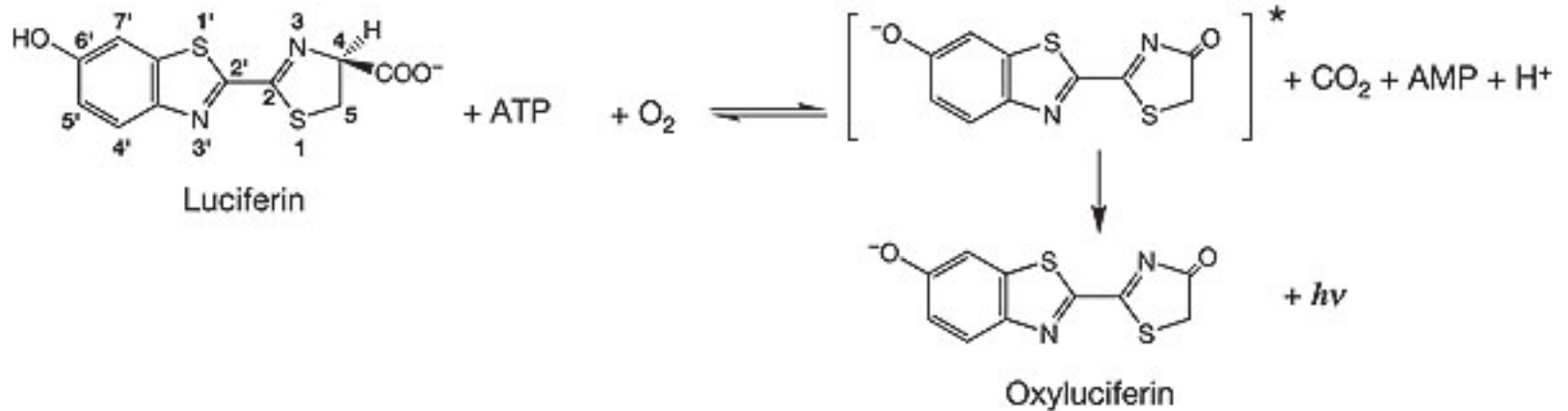


- Protein engineering challenge: create a red-emitting system?

Bioluminescence reporter gene imaging

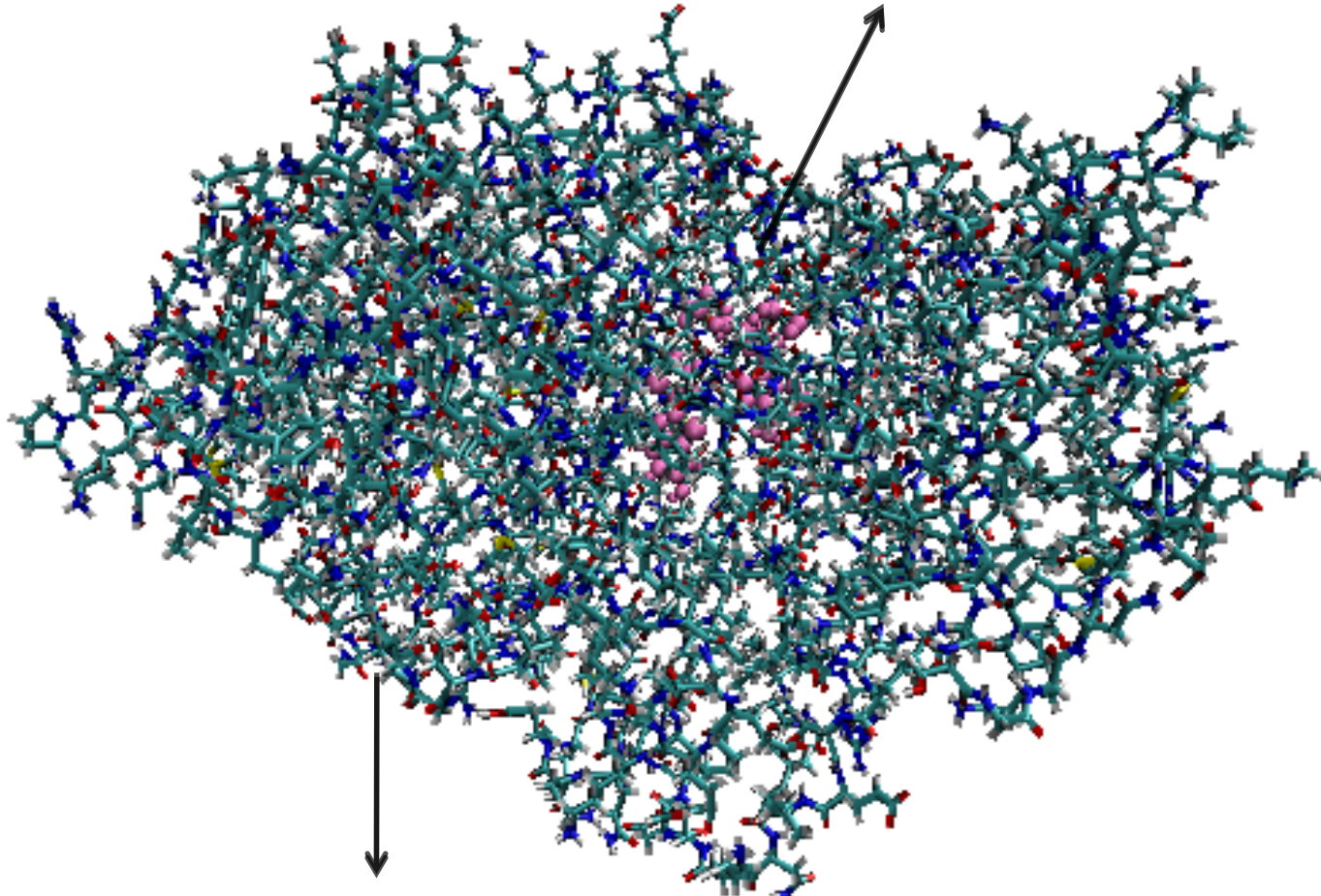


Excitation of luciferin



Atomic structure of luciferase/luciferin complex

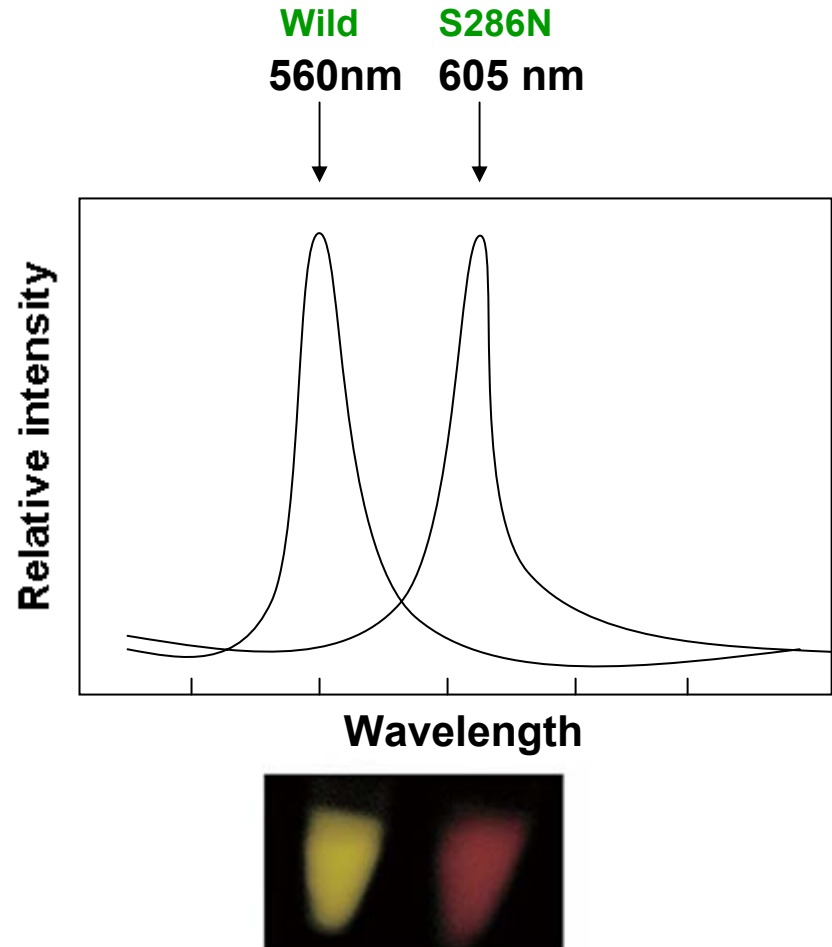
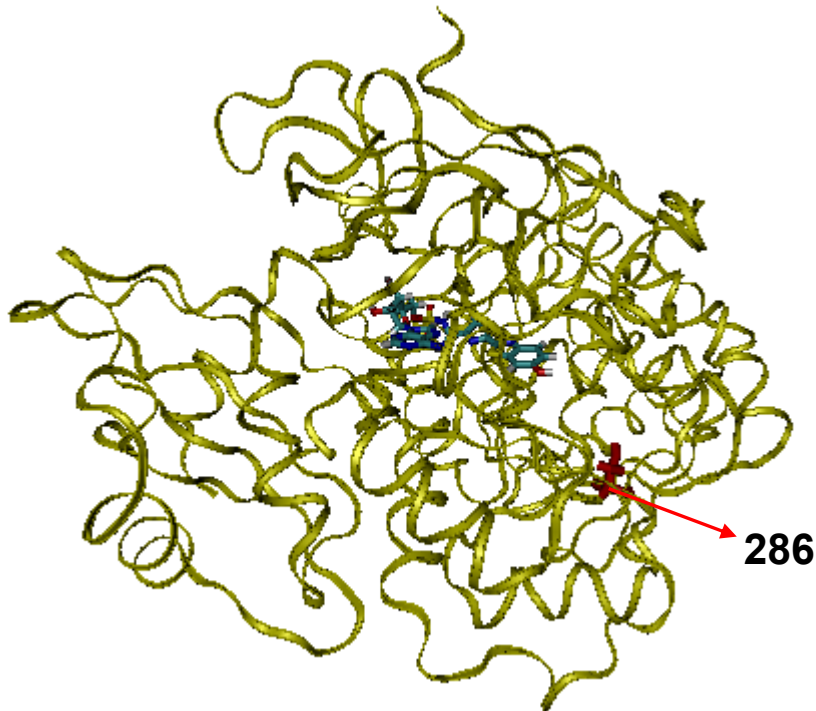
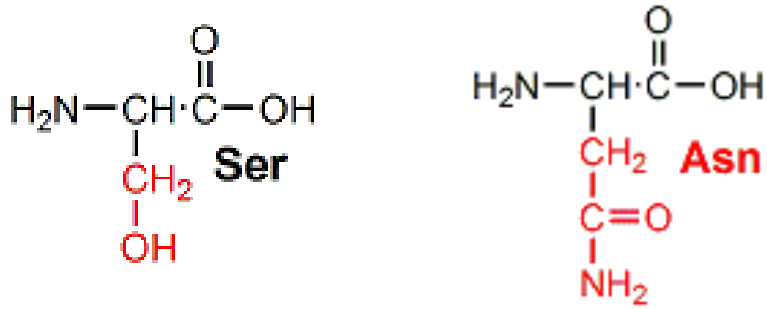
Luciferin: reactant for bioluminescence reaction



Luciferase: catalyzes the reaction and influences the outcome of reaction (frequency)

Spectral shift: luciferase-luciferin interactions

Mutation at site 286: Ser → Asn



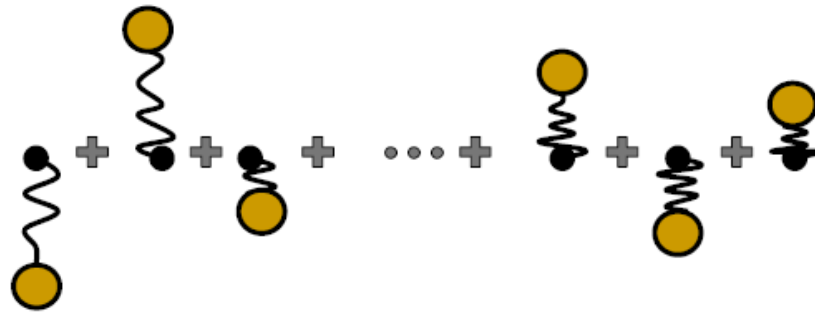
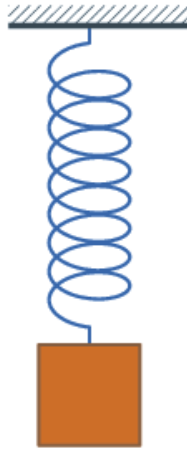
Nakatsu et al., *Nature* 440, p 372-376, 2006

Application of elastic network model to luciferase

- ◆ **Goal: to identify residues whose mutations may have the potential to change the bioluminescence frequency**
- ◆ **Approach:**
 - **validates the linkage between protein global dynamics and function**
 - **identifies the important residues**
 - **probes the nature of couplings between the important residues and the active site**

Meaning of Normal modes

- ◆ Normal mode: all parts move with the same frequency and phase
- ◆ Any motion of the system can be thought as a combination of its normal modes.

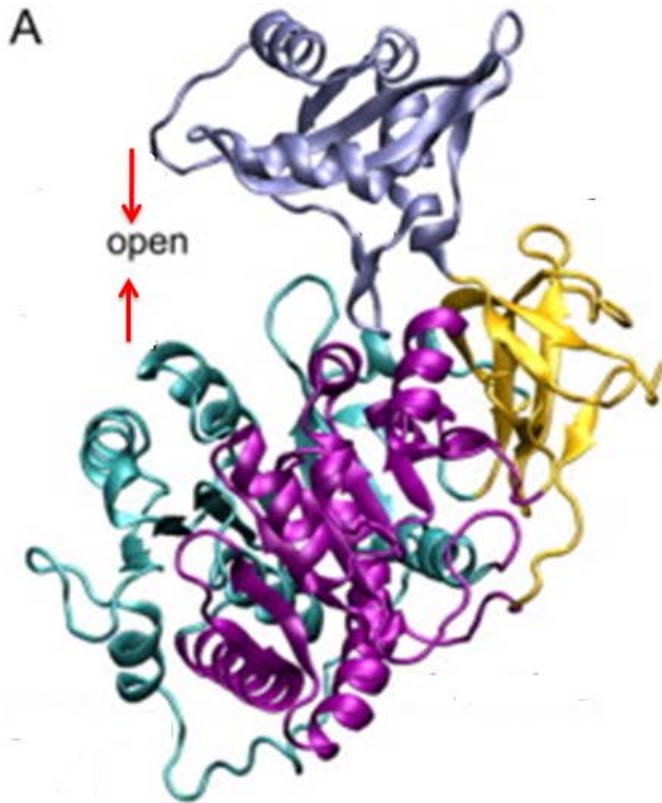


Global motions \longleftrightarrow Local motions
Low frequencies High frequencies

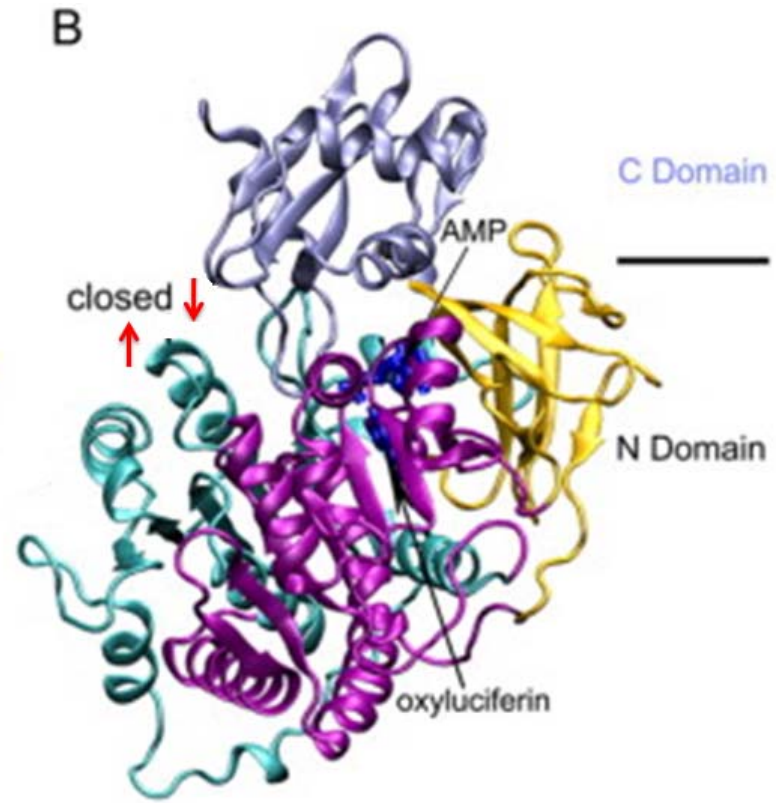
Question: what is the biological meaning of these modes?

Functionally most important motion: binding-induced change

Unbound form of Luciferase
(without substrate)



Bound form of Luciferase
(with substrate)



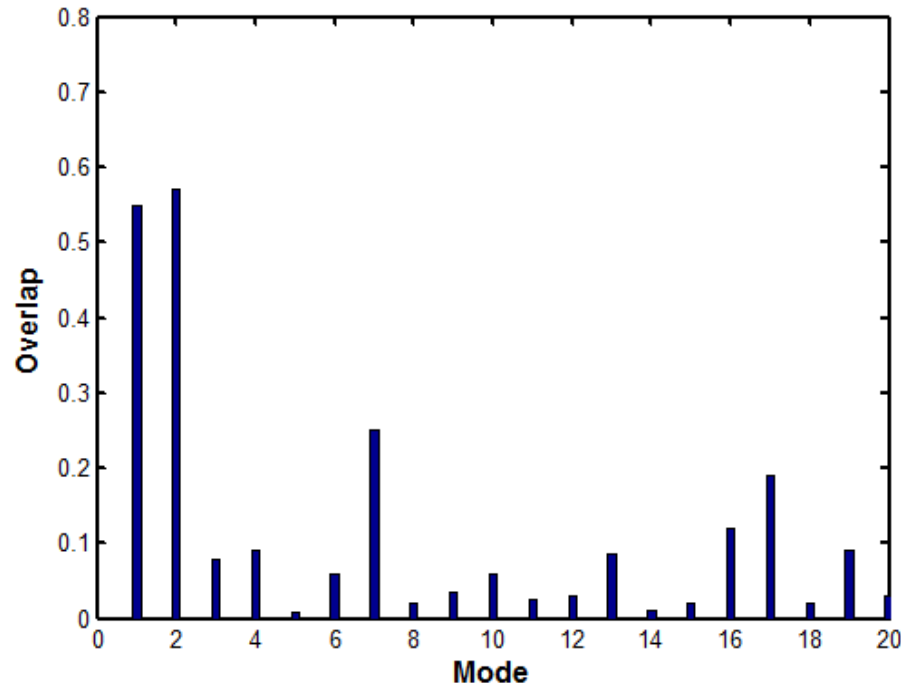
Normal mode analysis: biological meanings of normal modes

- Description of conformational change by normal modes

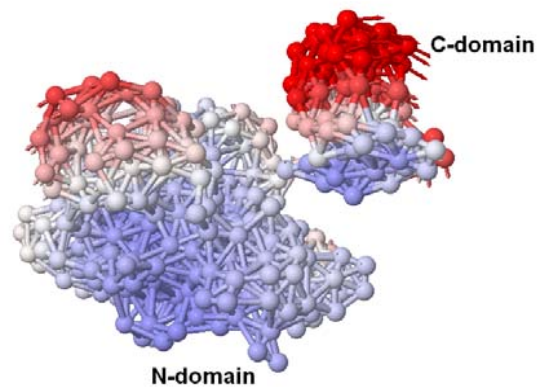
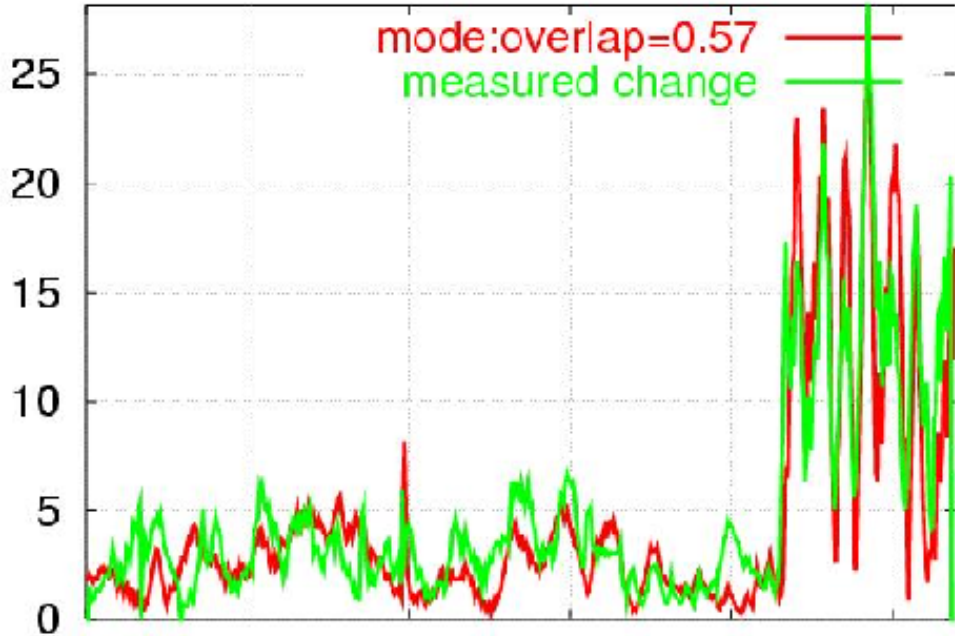
Luciferase (unbound form) \longrightarrow Luciferase + substrate (bound form)

Overlap between the modes and conformational change

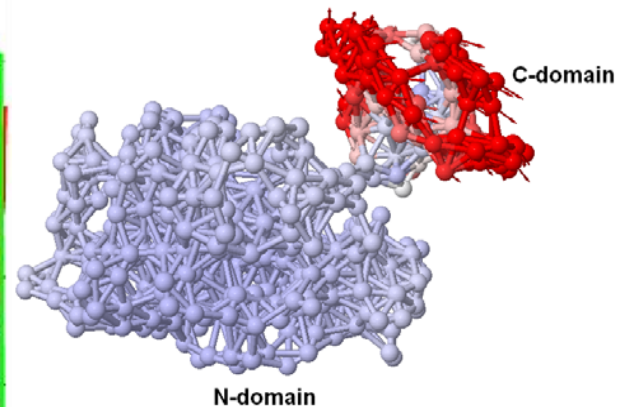
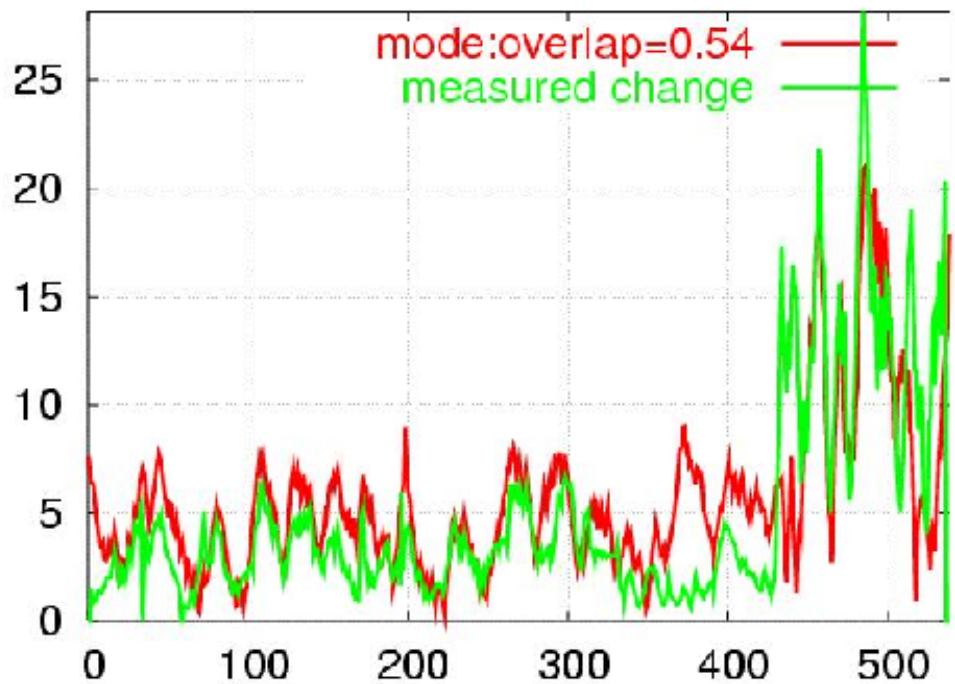
= **cosine** between two vectors



Amplitude



Amplitude




Residue

Summary from normal mode analysis

- The two lowest-frequency modes adequately account for the observed conformational changes induced by binding.
- It validates the applicability of the elastic network model to luciferase.

Perturbation analysis: identifying important residues

Change in sequence  Change in function
?

- **Experimentally
random mutagenesis**

Replace one residue type
by another at position i

- **Computationally
perturbation analysis**

Change in the
force constant of residue i

Change in the
fluctuations of the active site

Perturbation analysis

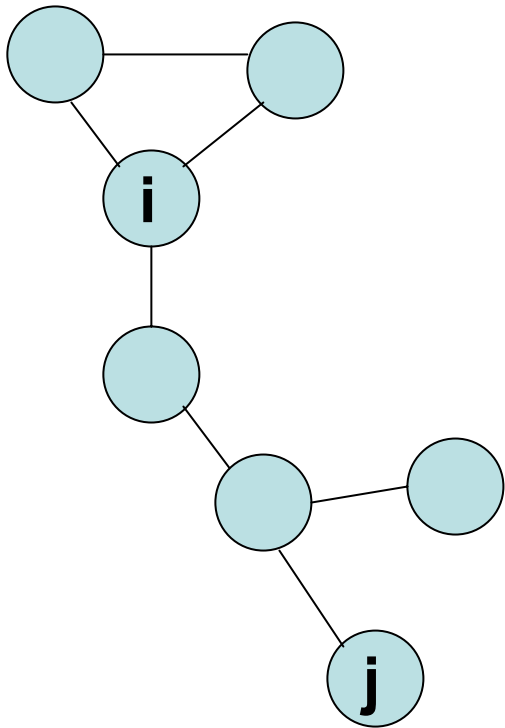
- Introduce an energy perturbation at node **i**

$$E_i' = \frac{1}{2} \sum_k C' (d_{ik} - d_{ik}^0)^2$$

- Measure the difference in the fluctuation at node **j**

the perturbation based correlation

$$\frac{R_{jj}' - R_{jj}}{R_{jj}}$$

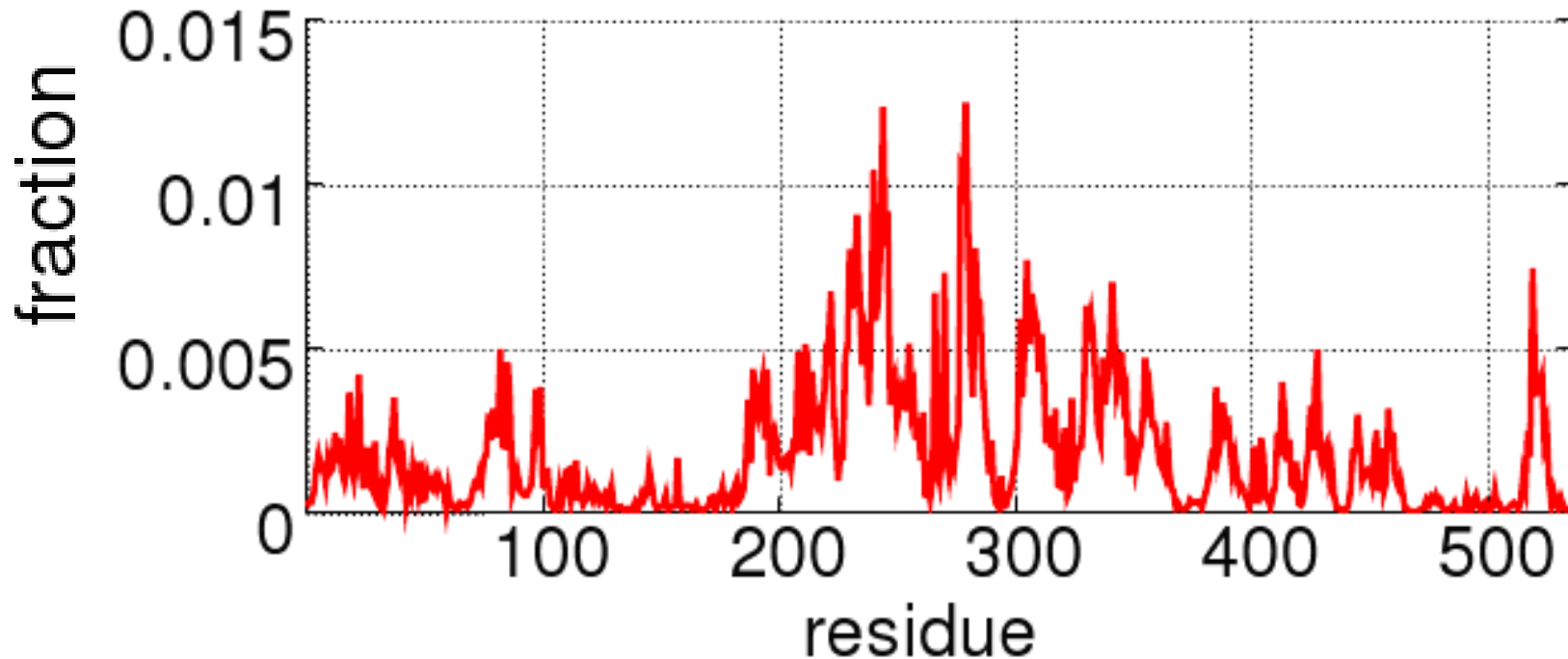
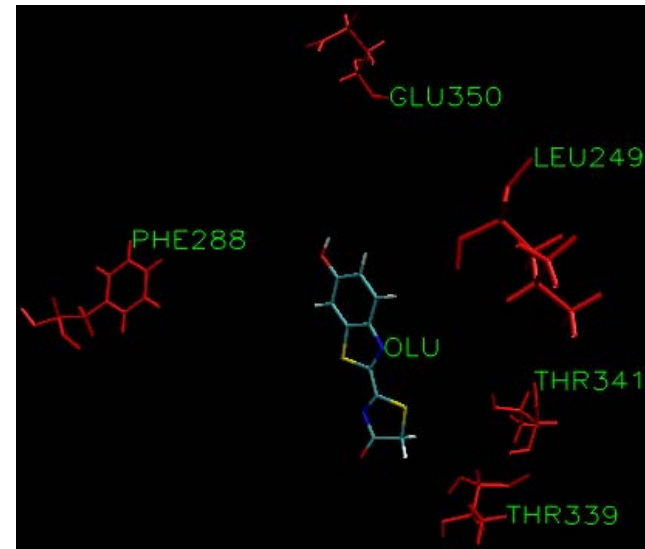


Application of perturbation analysis to luciferase

- node j = residues at the binding site (249, 288, 339, 341 and 350)

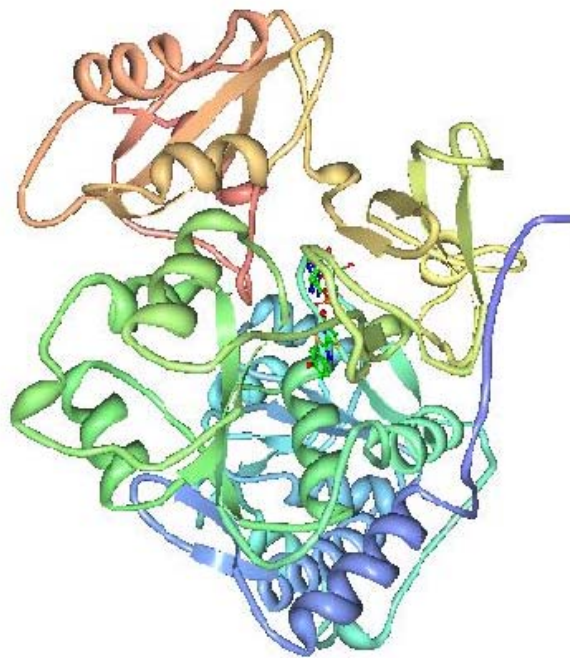
node i = all the other residues

- the fraction of change in the fluctuation of j $\frac{R_{jj}' - R_{jj}}{R_{jj}}$

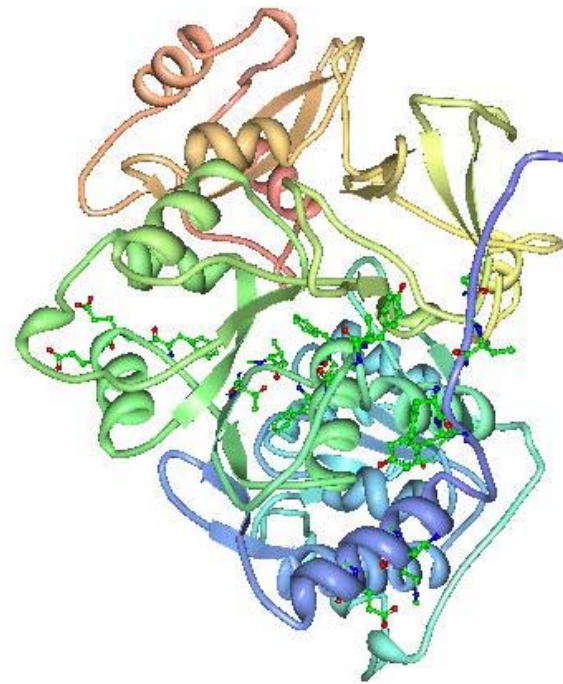


The important residues

91, 217, 220, 229, 230, 231, 237, 238,
240, 241, 242, **243**, 245, 246, 248, 250,
251, 252, 253, 254, 255, 264, 275, 279
286, 287, 289, 290, 292, 293, 294, 311,
313, 314, 315, 316, 317, 318, 320, 339,
340, 341, 342, 350, **351**, 354, 364, 437, 528



active site

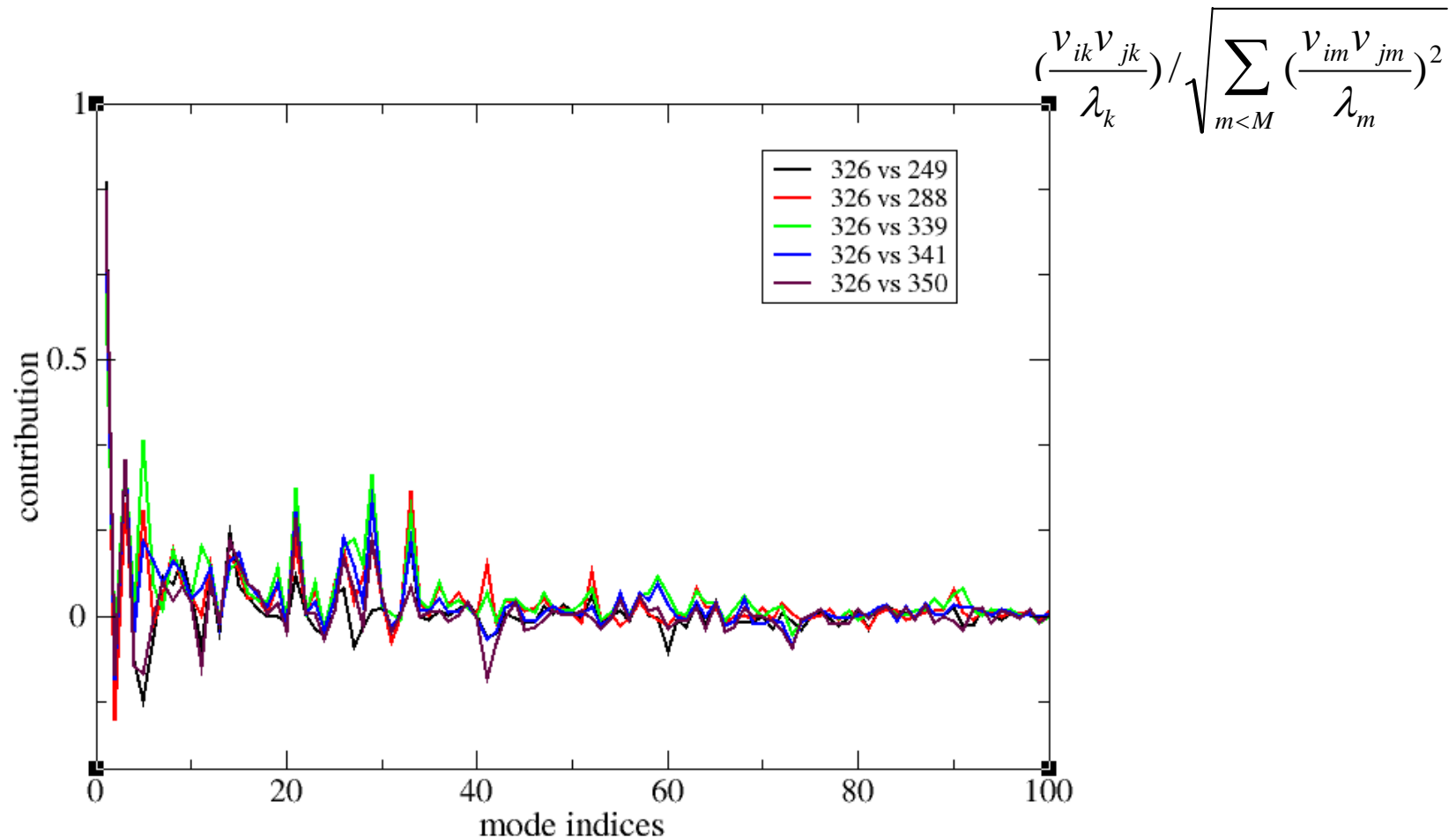


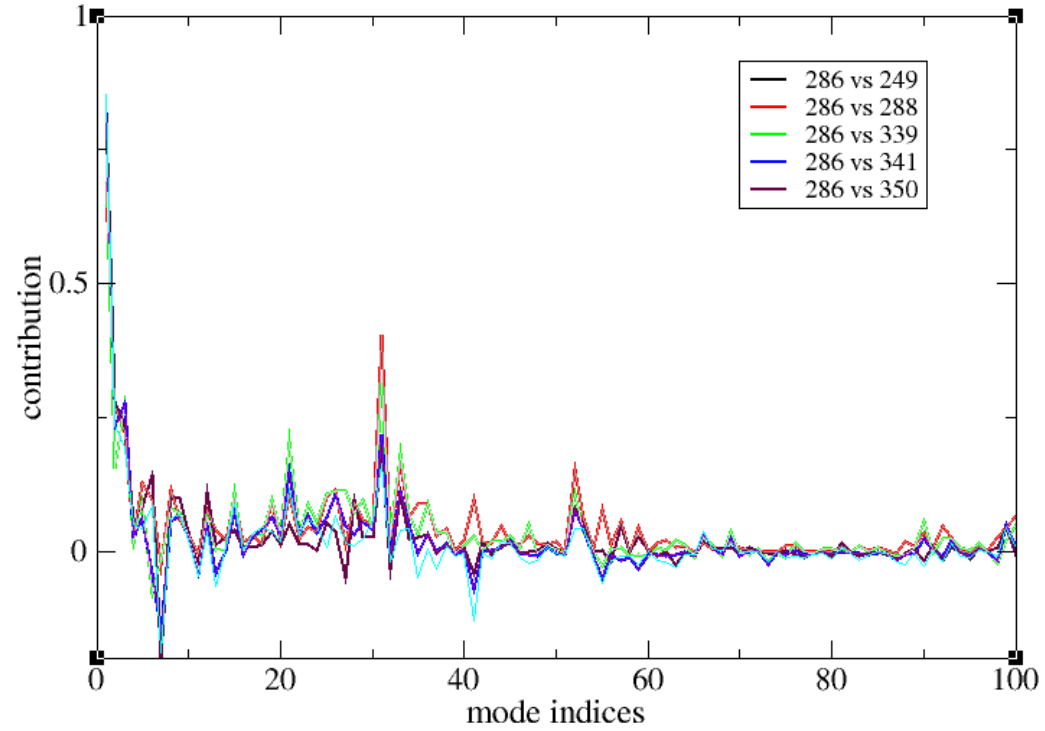
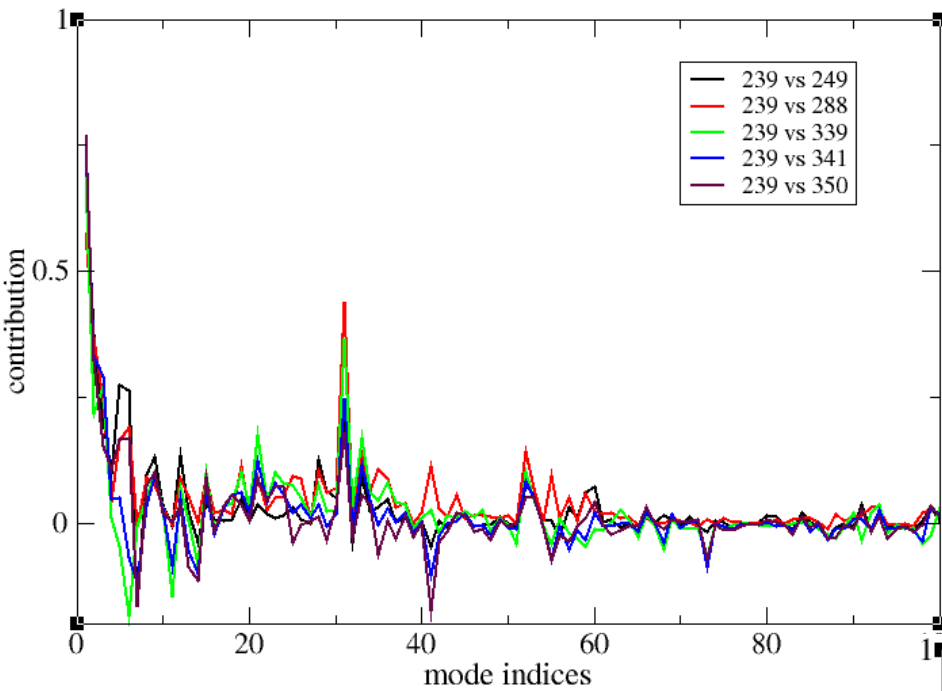
important residues

The mode decomposition analysis

$$R_{ij} = \frac{k_B T}{C} \sum_k \left[\frac{v_{ik} v_{jk}^T}{\lambda_k} \right]_{ij}$$

- Contribution of the k th mode to the correlation of residues i and j



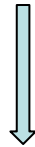


Application of elastic network model to luciferase

- **Network modeling captures the motions essential to luciferase function.**
- **Perturbation approach identifies the important residues.**
- **Lowest frequency modes are mainly responsible for the couplings between the remote important residues and the active site.**

Future plan

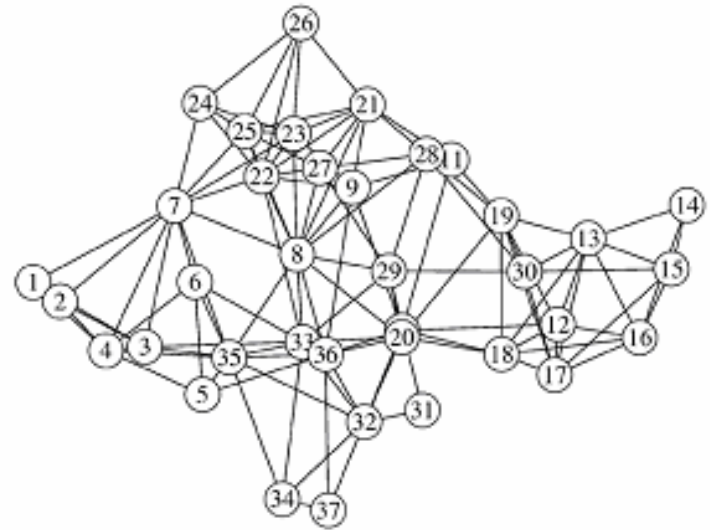
- Why do such simplified network models work?



information encoded in global topology



local stochasticity



- When do such simplified network models work?

Acknowledgements

**The work is partially supported by a
fellowship from NIMBioS.**