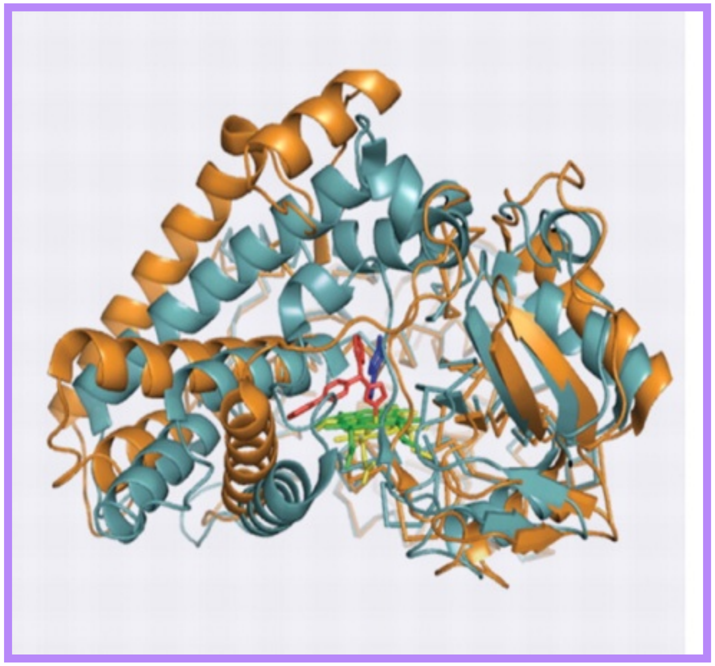
Intrinsically accessible motions enable
Optimal binding of substrate or drugs



Conformational flexibility + sequence variability mediates **substrate selectivity**

- **Two conformations of P450-CYP2B4:**

  open (orange) with a large substrate (bifonazole, red), and

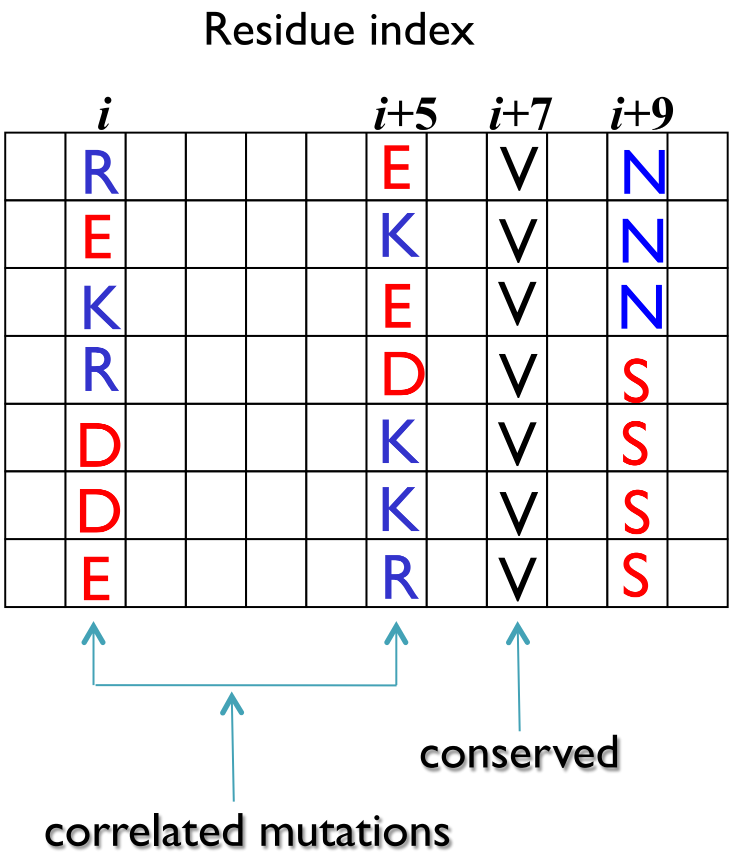
  closed (light blue) with the smaller substrate 4-(4-chlorophenyl) imidazole (blue)

See...

N. Tokuriki and D. S. **Tawfik** (2009) *Science* **324**: 203-207

Sequence evolution

an information-theoretic approach



Residue index

	i					$i+5$		$i+7$		$i+9$	
	R					E		V		N	
	E					K		V		N	
	K					E		V		N	
	R					D		V		S	
	D					K		V		S	
	D					K		V		S	
	E					R		V		S	

conserved

correlated mutations

Information entropy (Shannon, 1951)

$$S(i) = \sum_{x_i=1}^{20} P(x_i) \log \frac{1}{P(x_i)}$$

Mutual information (MI)

$$\mathrm{I}(i,j) = \sum_{x_i=1}^{20} \sum_{y_j=1}^{20} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i) P(y_j)}$$

for correlated mutations analysis (CMA)

Mutual Information

without the influence of phylogeny

MIp - to eliminate random noise and phylogenetic components

$$\mathrm{MIp}\,(i, j) = \mathrm{I}(i, j) - \mathrm{APC}$$

APC = Average product correction

$$= [\ \mathrm{I}(i, x)\ \mathrm{I}(j, x)\ ]\ \ / <\mathrm{I}(i, j)>$$

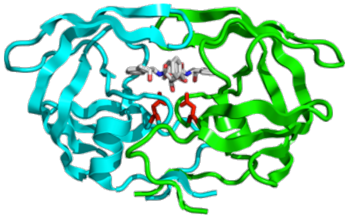
	R				E		V		N		
	E				K		V		N		
	K				E		V		N		
	R				D		V		S		
	D				K		V		S		
	D				K		V		S		
	E				R		V		S		

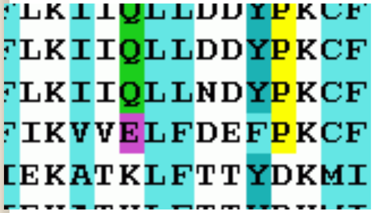
where $\mathrm{I}(i, x)$ is the mean mutual information of column i = $\Sigma_j\ \mathrm{I}(i, j)$

Dunn, Wahl and Gloor (2008) *Bioinformatics* **24**: 333-340

HIV-1 protease correlated mutation analysis (CMA)

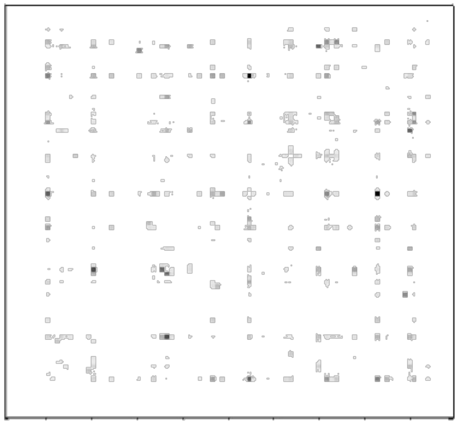
MSA of HIV-1 protease





LKIIQLLDDYPKCF
FLKIIQLLDDYPKCF
FLKIIQLLNDYPKCF
FIKVVELFDEFPKCF
IEKATKLFTTYDKMI

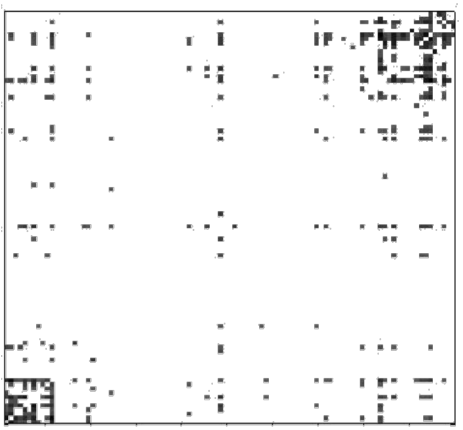
MI matrix $\mathbf{I}_{ij} = \mathrm{I}\,(i, j)$



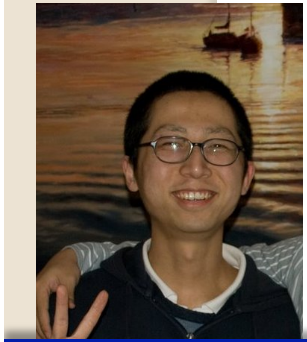
residue index

Shi and Malik (2000)

spectral clustering



reordered residue index



Dr. Ying Liu

Liu, Eyal & Bahar (2008) *Bioinformatics*

MDR mutations distinguished by CMA

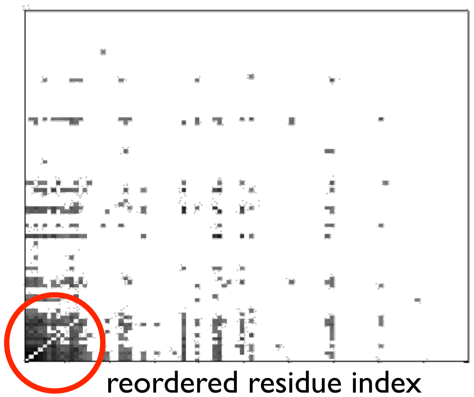
MSA of HIV-1 protease

Stanford HIV Drug Resistance Database

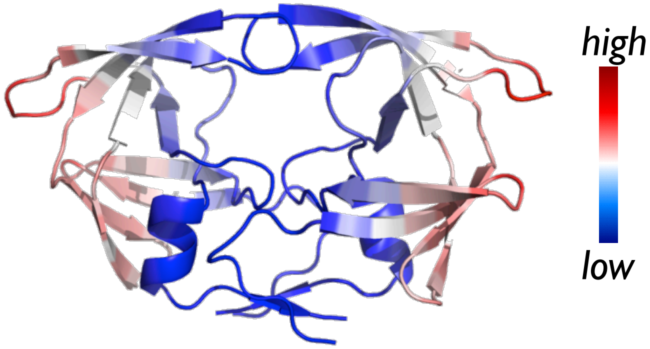
http://hivdb.stanford.edu/

```
CTLVGTAIHEMMHALGFLHEQNREDRDDWVR
CDKFGIVVHELGHVVGFWHEHTRPDREDHVV
CFRFGTVIHEFIHALGFYHAQSAYTRDDYVL
NFTVGSLIHEIGHAFGLIHEHQRPDRDDYVI
CLTYGTPIHELMHALGFFHEQNRHERDSYVR
CDKFGIVVHELGHVVGFWHEHTRPDREKHVV
CDKFGVVVHELGHVVGFWHEHTRPDRNEFVG
CAYFGTIVHEIGHAIGFHHEQSRPDRDDYIN
CVYHGIIQHELSHALGFYHEHTRSDRNKYVR
CINSGTIIHEVLHALGVHHEQARADRDGYVT
```

untreated

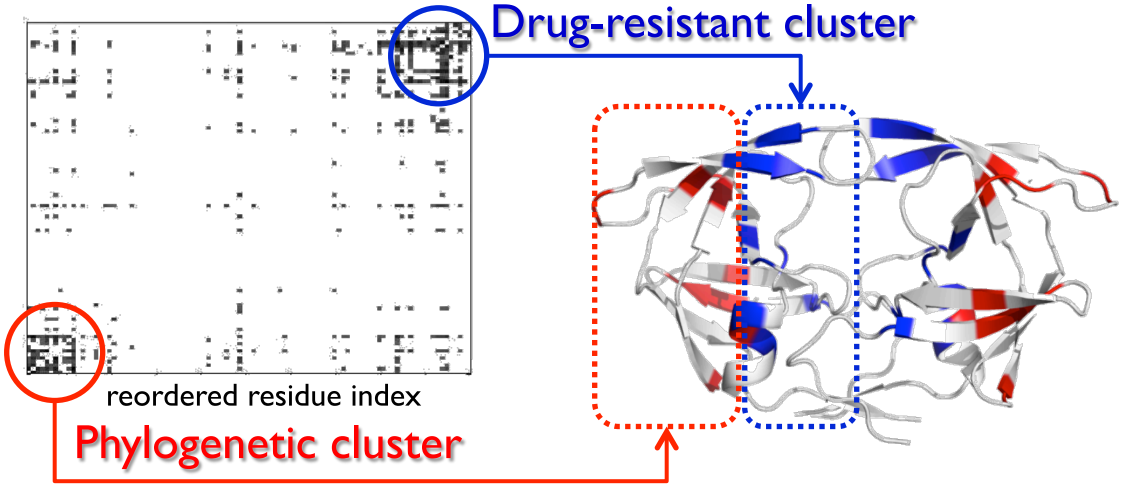


mobility profile



```
CTLVGTAIHEMMHALGFLHEQNREDRDDWVR
CDKFGIVVHELGHVVGFWHEHTRPDREDHVV
CFRFGTVIHEFIHALGFYHAQSAYTRDDYVL
NFTVGSLIHEIGHAFGLIHEHQRPDRDDYVI
CLTYGTPIHELMHALGFFHEQNRHERDSYVR
CDKFGIVVHELGHVVGFWHEHTRPDREKHVV
CDKFGVVVHELGHVVGFWHEHTRPDRNEFVG
CAYFGTIVHEIGHAIGFHHEQSRPDRDDYIN
CVYHGIIQHELSHALGFYHEHTRSDRNKYVR
CINSGTIIHEVLHALGVHHEQARADRDGYVT
```

treated by at least one drug



Summary

- two groups of correlated mutation sites

functional aspects	Structural location	structural dynamics
phylogenetic	exposed	mobile
multi-drug resistant	dimerization interface	restrained

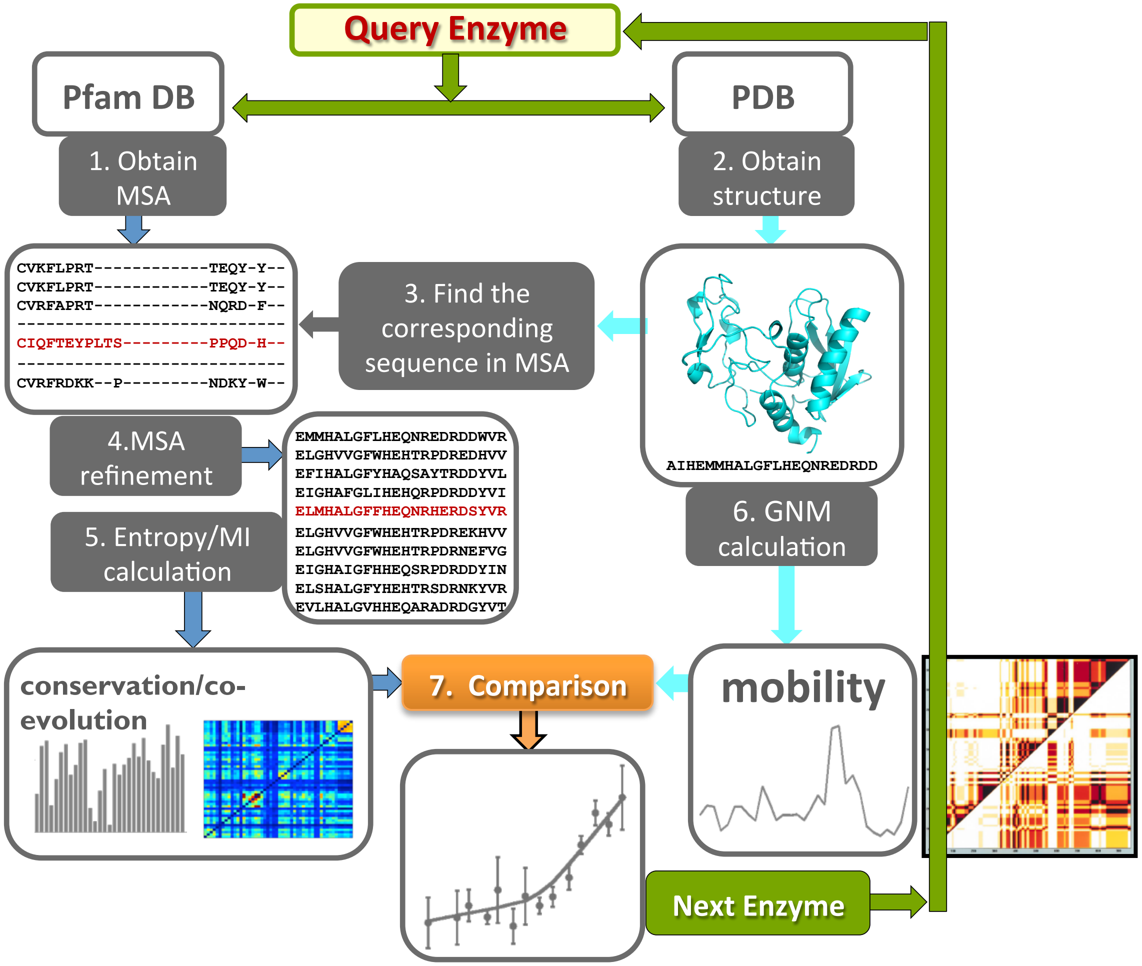
Liu, Eyal & Bahar (2008) *Bioinformatics* **15**, 1243.

Questions:

- Are key mechanical sites (e.g. hinges) conserved?
- Is there any correlation between sequence variability and structural dynamics?
- How does the structure ensure substrate specificity *and* conformational adaptability?

MOLECULAR BIOLOGY AND EVOLUTION

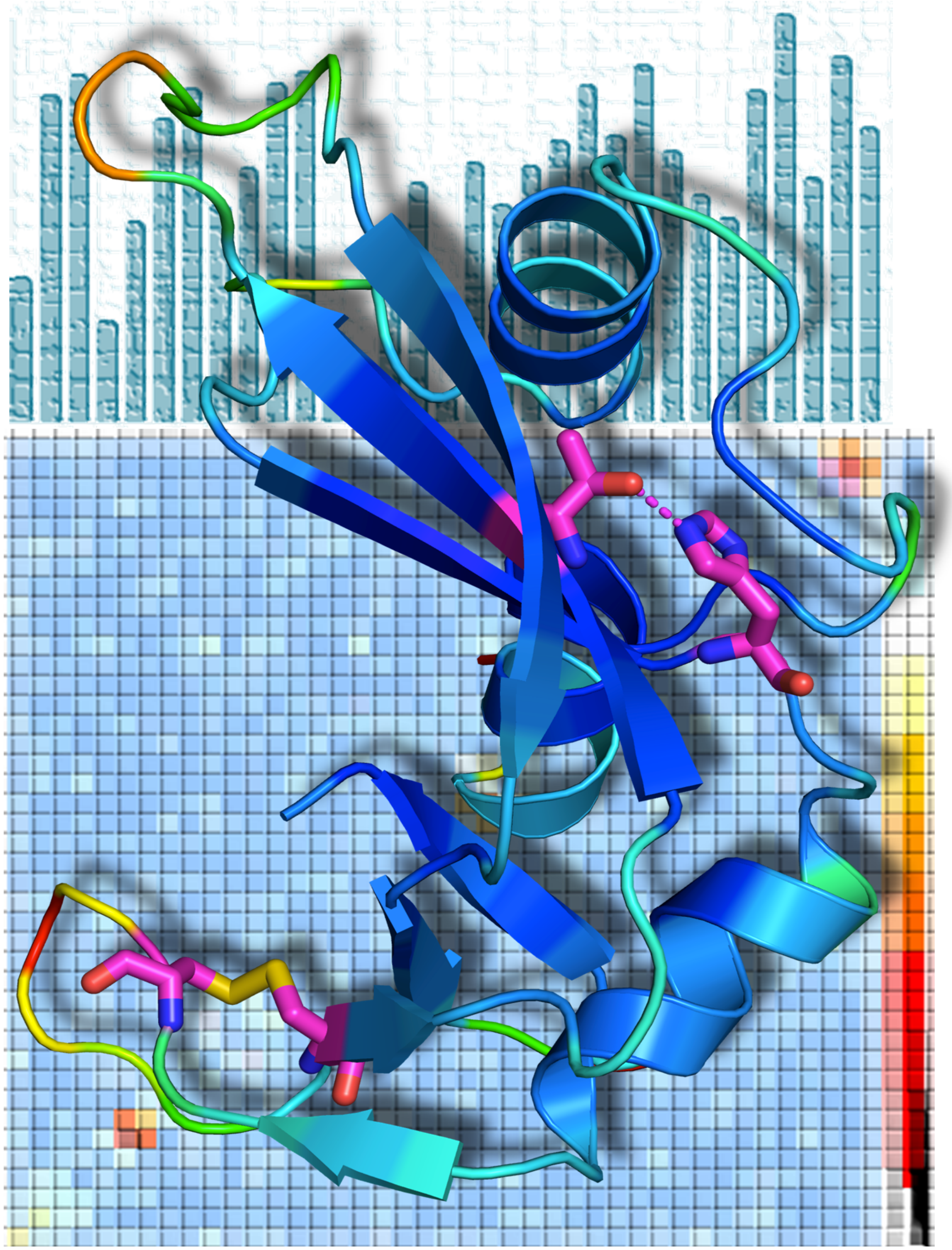
A systematic study of a set of enzymes



Liu Y, Bahar I (2012) “Sequence Evolution Correlates with Structural Dynamics” *Mol Biol Evol 9, 2253-63*

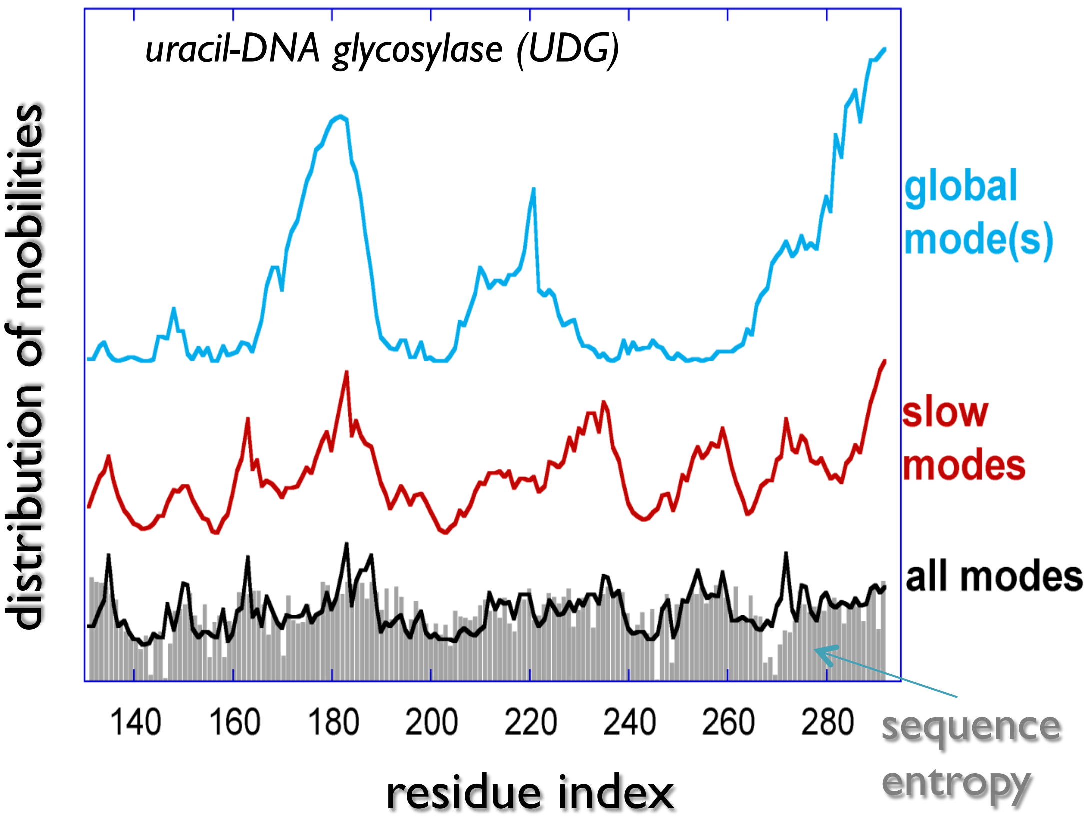
Evol

MOLECULAR BIOLOGY AND EVOLUTION

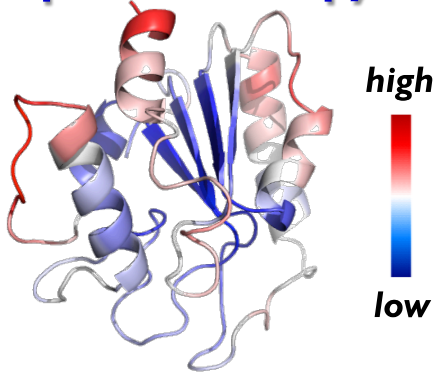


http://www.csb.pitt.edu/prody/tutorials/evol_tutorial/index.html

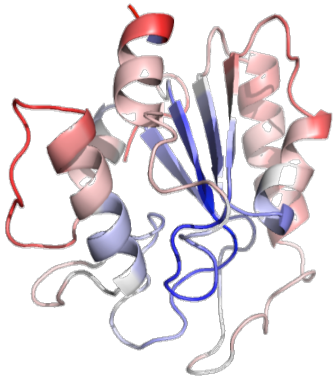
Correlation between sequence entropy & conformational mobility



sequence entropy

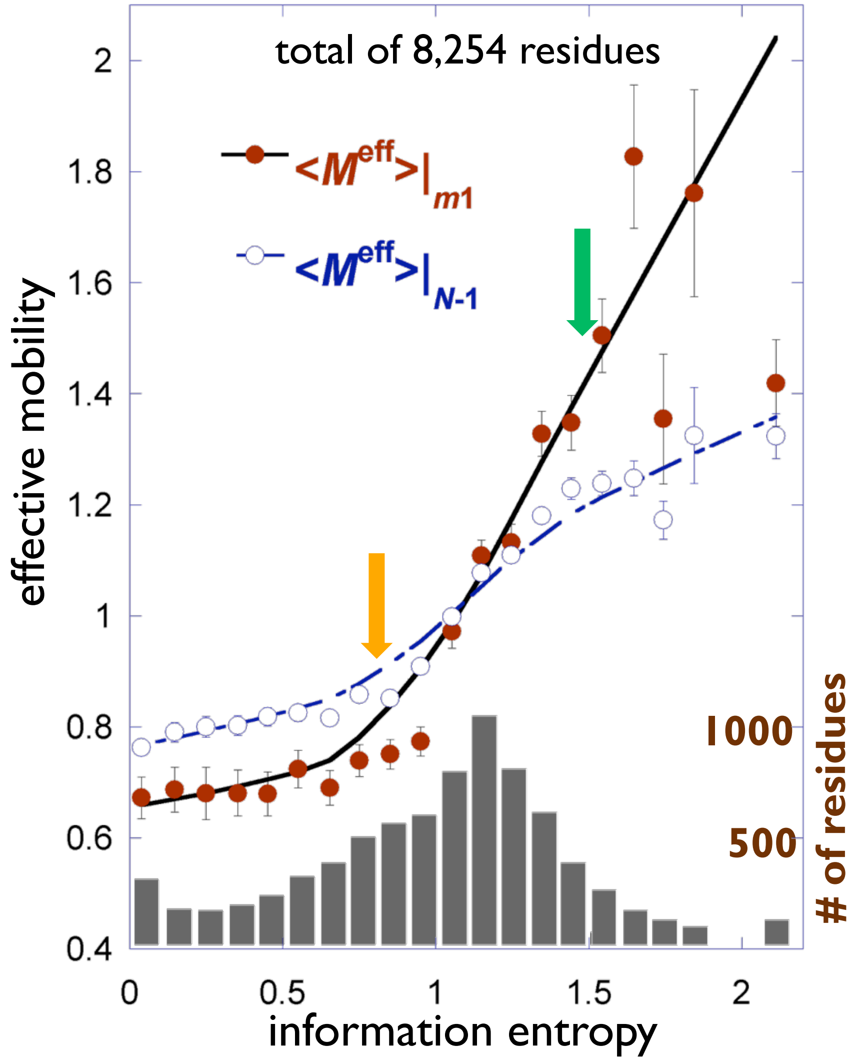


structural dynamics



Liu Y, Bahar I (2012) "Sequence Evolution Correlates with Structural Dynamics" *Mol Biol Evol 9, 2253-63*

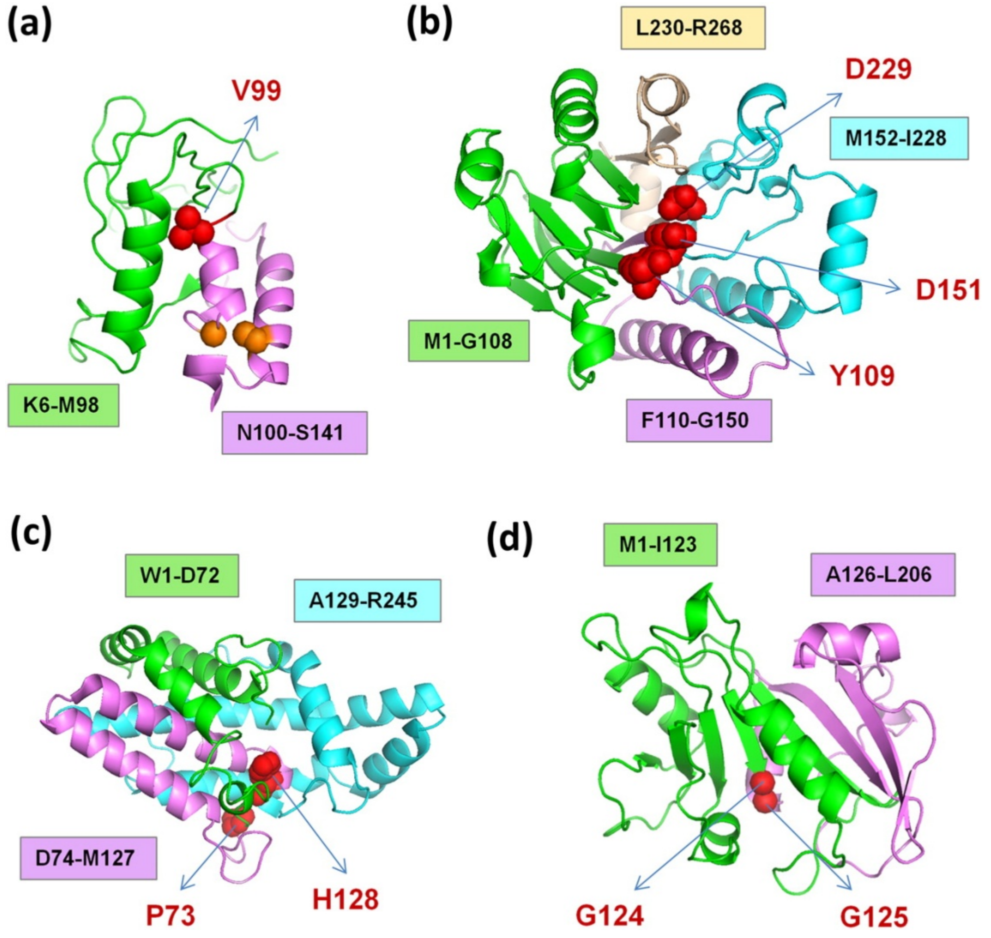
Mobility increases with sequence entropy



Liu & Bahar *Mol Biol Evol* (2012)

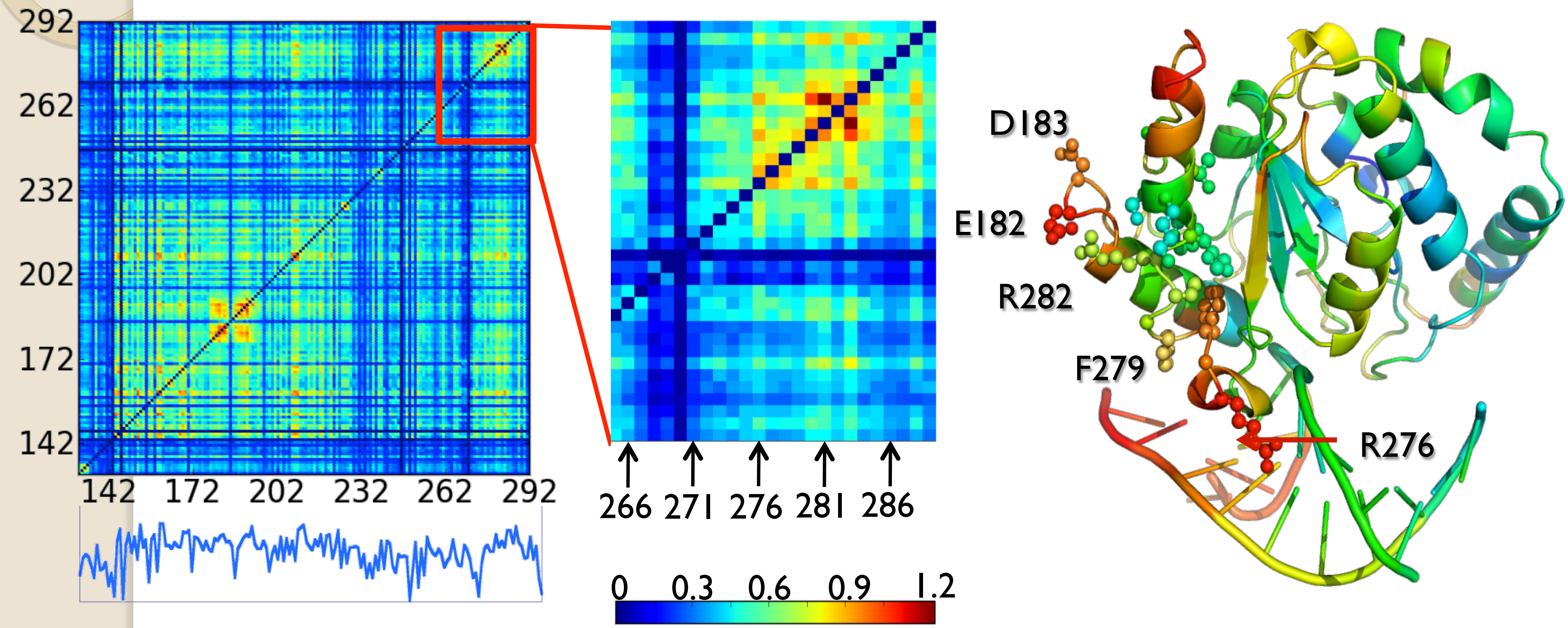
Hinge sites are evolutionarily conserved

despite their moderate-to-high exposure to environment



Liu & Bahar *Mol Biol Evol* (2012)

Amino acids involved in intermolecular recognition are distinguished by **their co-evolution propensities**

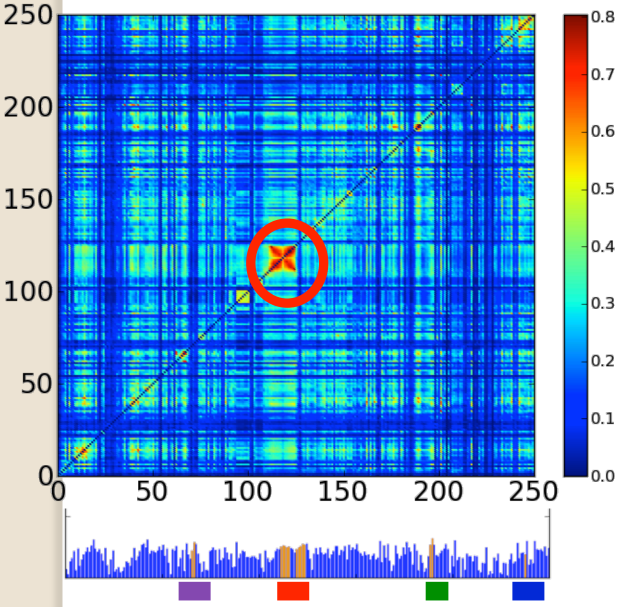


Liu Y, Bahar I (2012) “Sequence Evolution Correlates with Structural Dynamics” *Mol Biol Evol 9, 2253-63*

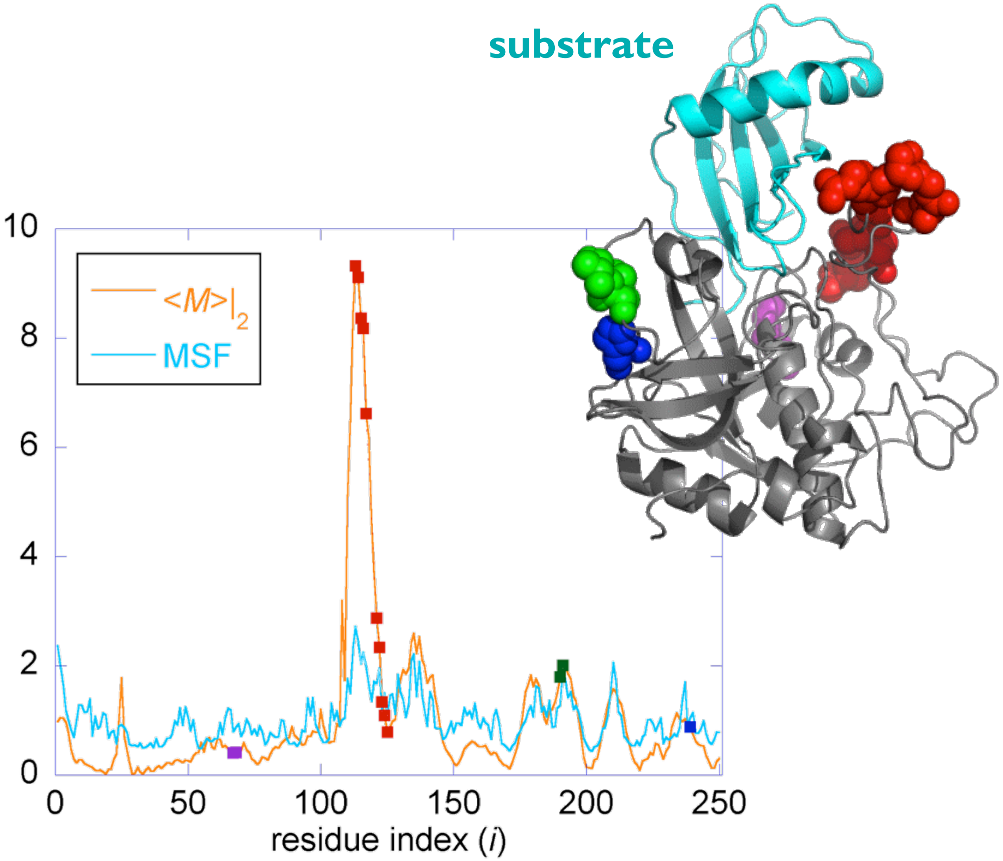
3

Amino acids involved in intermolecular recognition are distinguished by **their high global mobility**

cathepsin B



residues involved in top 0.05% of I(*i*, *j*) values



Liu Y, Bahar I (2012) "Sequence Evolution Correlates with Structural Dynamics" *Mol Biol Evol 9, 2253-63*

Summary

Four types of functional sites

Functional site	Mobility in global modes	Sequence evolution	Dominant Feature
Chemical (catalytic, ligand binding)	Minimal	Conserved	high fidelity, precision
Core	Minimal	Conserved	high stability
Hinge sites	Minimal	Conserved	rotational flexibility
Substrate recognition (specific)	High	High co-evolution propensity	adaptability

Liu & Bahar *Mol Biol Evol* (2012); Liu, Gierasch & Bahar, *PLoS Comp Bio* (2010)

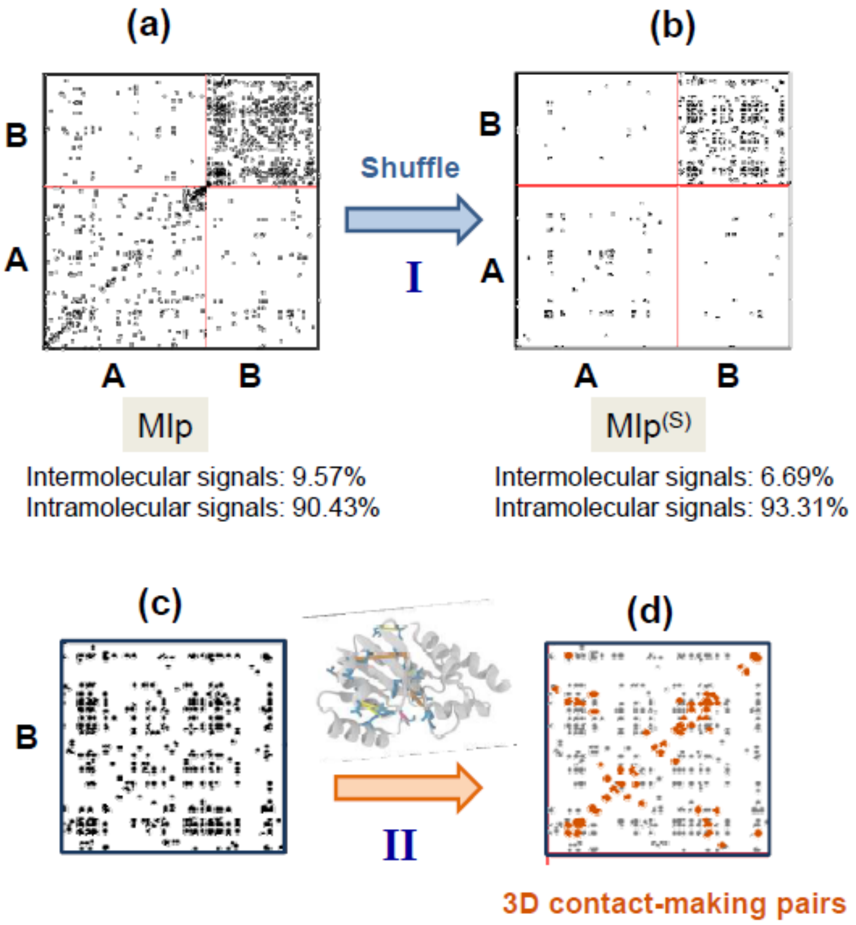
There are several methods for evaluating sequence co-evolution

Mao W, Kaya C, Dutta A, Horovitz A, Bahar I (2015) Comparative Study of the Effectiveness and Limitations of Current Methods for Detecting Sequence Coevolution *Bioinformatics* **pii**: btv103PMID: 25697822

Four possible outcomes:

- True positive (TP) – correctly predicted to be a hit
- False positive (FP); predicted but it is a miss
- True negative (TN) – correctly predicted to be a miss
- False negative (FN) – predicted as a miss, but is a hit

Two criteria for assessing the performance of different methods



MIp

Intermolecular signals: 9.57%
Intramolecular signals: 90.43%

MIp$^{(S)}$

Intermolecular signals: 6.69%
Intramolecular signals: 93.31%

3D contact-making pairs

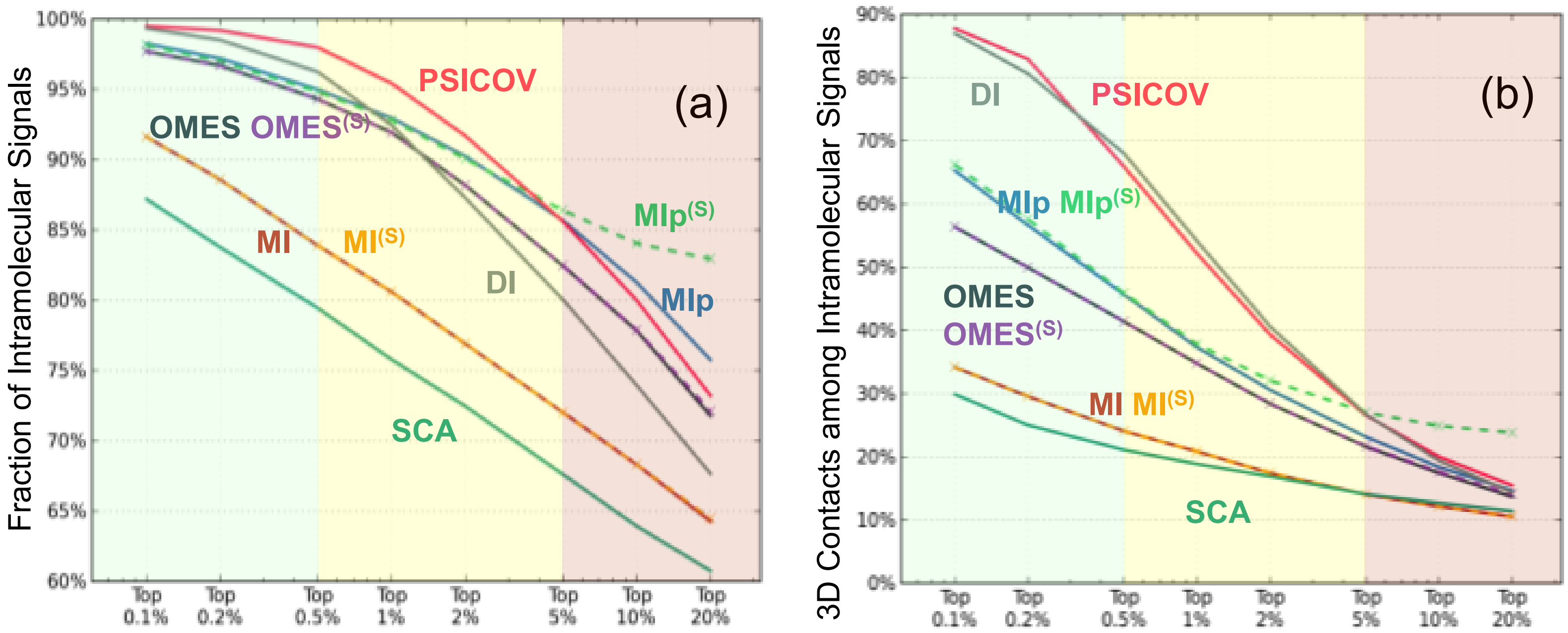
- Minimizing false positives (signals between non interacting proteins)
- Maximizing true positives (signals between contact making residues

Screening of large databases

For testing 9 methods, including

- observed-minus-expected-squared ((OMES) (Kass and Horovitz, 2002)
- statistical coupling analysis (SCA) (Halabi et al., 2009; Lockless and Ranganathan, 1999).
- Direct Coupling Analysis (DCA or DI for Direct Information) (Morcos et al., 2011; Weigt et al., 2009),
- Protein Sparse Inverse COVariance (PSICOV) (Jones et al., 2012),

PSICOV and DI are the best



Average performance of the nine methods based on two criteria, absence of intermolecular FPs (a), and fraction of 3D contact making pairs (b) among different subsets of top-ranking signals. The signals are classified to 3 groups: strong coevolution signals (0.1-0.5%), intermediate signals (0.5-5%) and relatively weak signals (5-20%), which also refer to relatively small, intermediate, and high coverage of coevolving pairs. PSICOV and DI outperform other methods in the strong coevolution region. For the intermediate signal, OMES and MIp exhibit performances similar to PSICOV and DI in panel **a**. MIp$^{(S)}$ shows the best performance in the weak signal regime. SCA and MI (and its shuffled version) have lower performance compared to all others for both criteria over the whole range.

Allosteric communication mechanisms explored by network models

- Diffusion of signal obeys a Markov process
- The structure is modeled as a network
- Network connectivity given by Γ

References

Laplacian based manifold methods (Belkin & Niyogi)

Chennubhotla & Bahar Mol Systems Biology (2006); PLoS Comp Bio (2007)

Markov Model of Network Communication

Γ = **D – A** where **A** = connectivity/affinity matrix and **D** = diagonal matrix of degrees

A discrete-time, discrete-state Markov process is defined by setting the conditional probability of signal transduction from residue *j* to *i* as

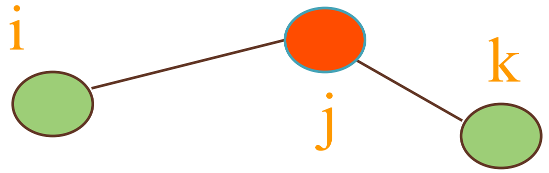
$$\boxed{m_{ij} = a_{ij} / d_j}$$

The conditional probability matrix $\mathbf{M} = \{m_{ij}\}$, also called the Markov transition matrix, is

$$\boxed{\mathbf{M} = \mathbf{A}\,\mathbf{D}^{-1}}$$

$\mathbf{M}$ completely defines the stochastics of information transfer over the network of residues.

Hitting time: a measure of communication efficiency between two endpoints



Based on all possible pathways

path	# of steps	Path Probability
$j \to i$	1	0.5
$j \to k \to j \to i$	3	0.5^2
$j \to k \to j \to k \to j \to i$	5	0.5^3

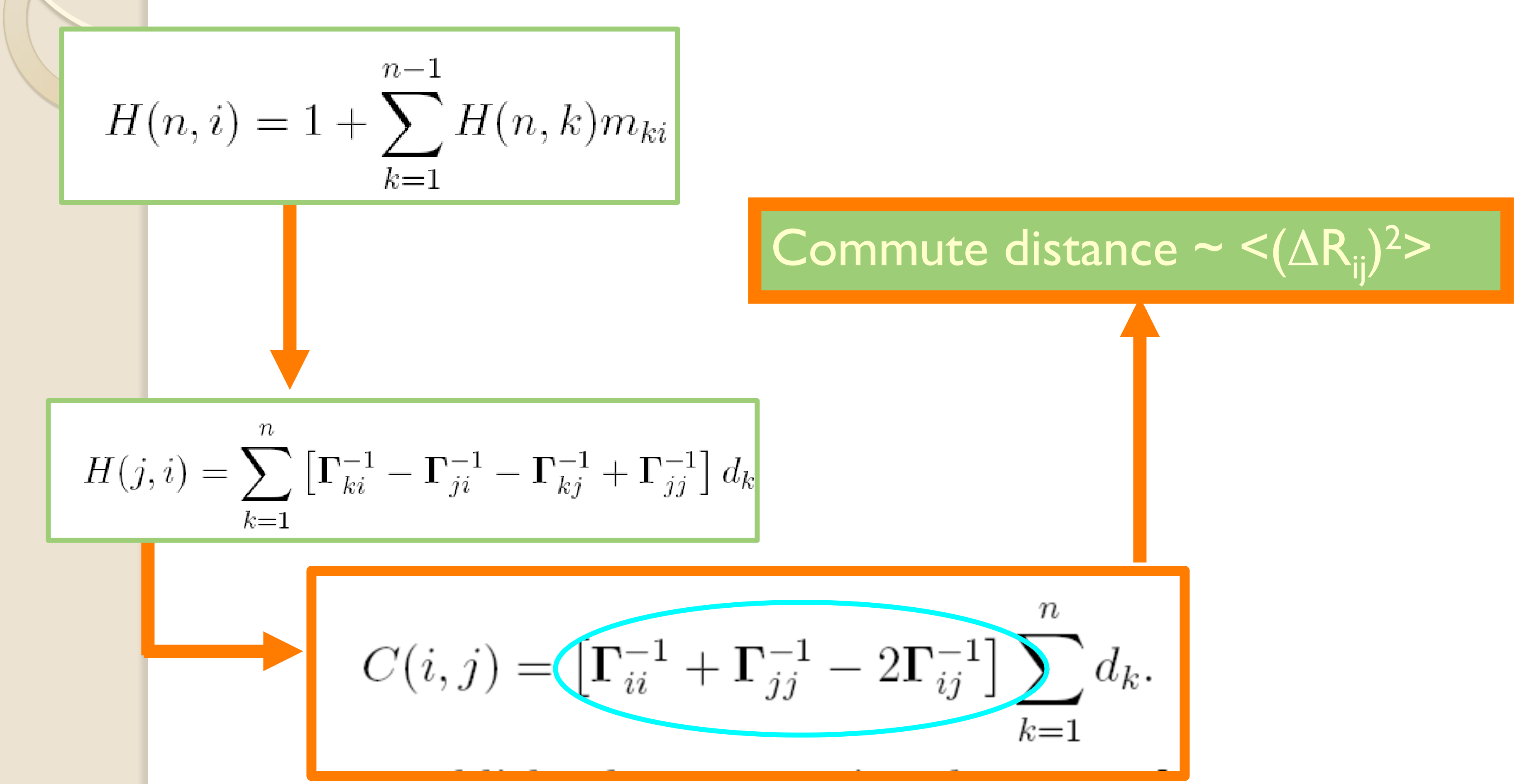
$$\boldsymbol{H}(j,i) = 1 \times 0.5 + 3 \times 0.5^2 + \cdots = \sum_{j=1}^{\infty}(2j-1) \times 0.5^j, = 3.$$

path	# of steps	Path Probability
$i \to j \to k$	2	0.5
$i \to j \to i \to j \to k$	4	0.5^2
$i \to j \to i \to j \to i \to j \to k$	6	0.5^3

$$\boldsymbol{H}(k,i) = 2 \times 0.5 + 4 \times 0.5^2 + \cdots = 2\sum_{j=1}^{\infty} j \times 0.5^j = 4.$$

P(t) = M P(0), where M = AD^{-1} is the conditional prob matrix for signal diffusion

Fluctuations as determinant of communication



$$H(n,i) = 1 + \sum_{k=1}^{n-1} H(n,k) m_{ki}$$

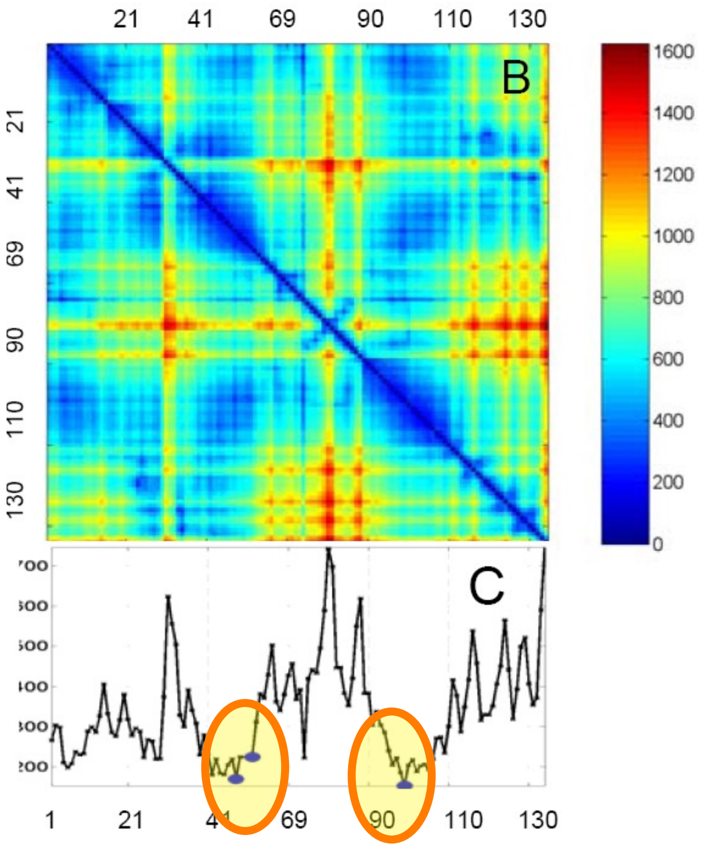
$$H(j,i) = \sum_{k=1}^{n} \left[\mathbf{\Gamma}^{-1}_{ki} - \mathbf{\Gamma}^{-1}_{ji} - \mathbf{\Gamma}^{-1}_{kj} + \mathbf{\Gamma}^{-1}_{jj} \right] d_k$$

$$C(i,j) = \left[\mathbf{\Gamma}^{-1}_{ii} + \mathbf{\Gamma}^{-1}_{jj} - 2\mathbf{\Gamma}^{-1}_{ij} \right] \sum_{k=1}^{n} d_k.$$

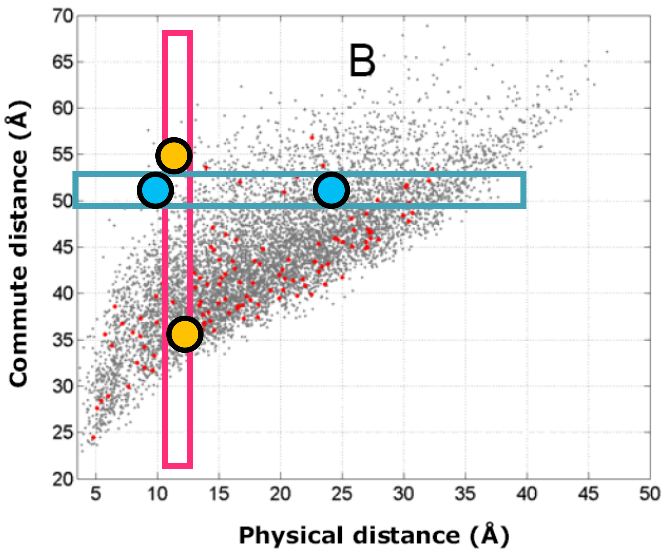
Commute distance ~ $<(\Delta R_{ij})^2>$

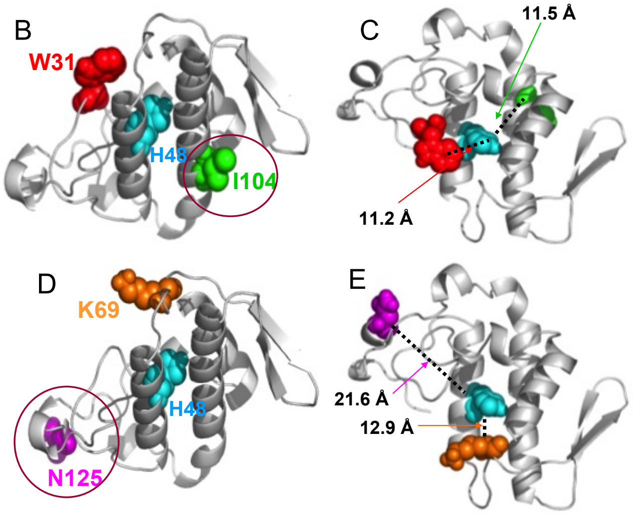
Chennubhotla & Bahar (2007) *PLoS Comp Bio*

Communication times



Distribution of Commute Times for Phospholipase A2 (1bk9)



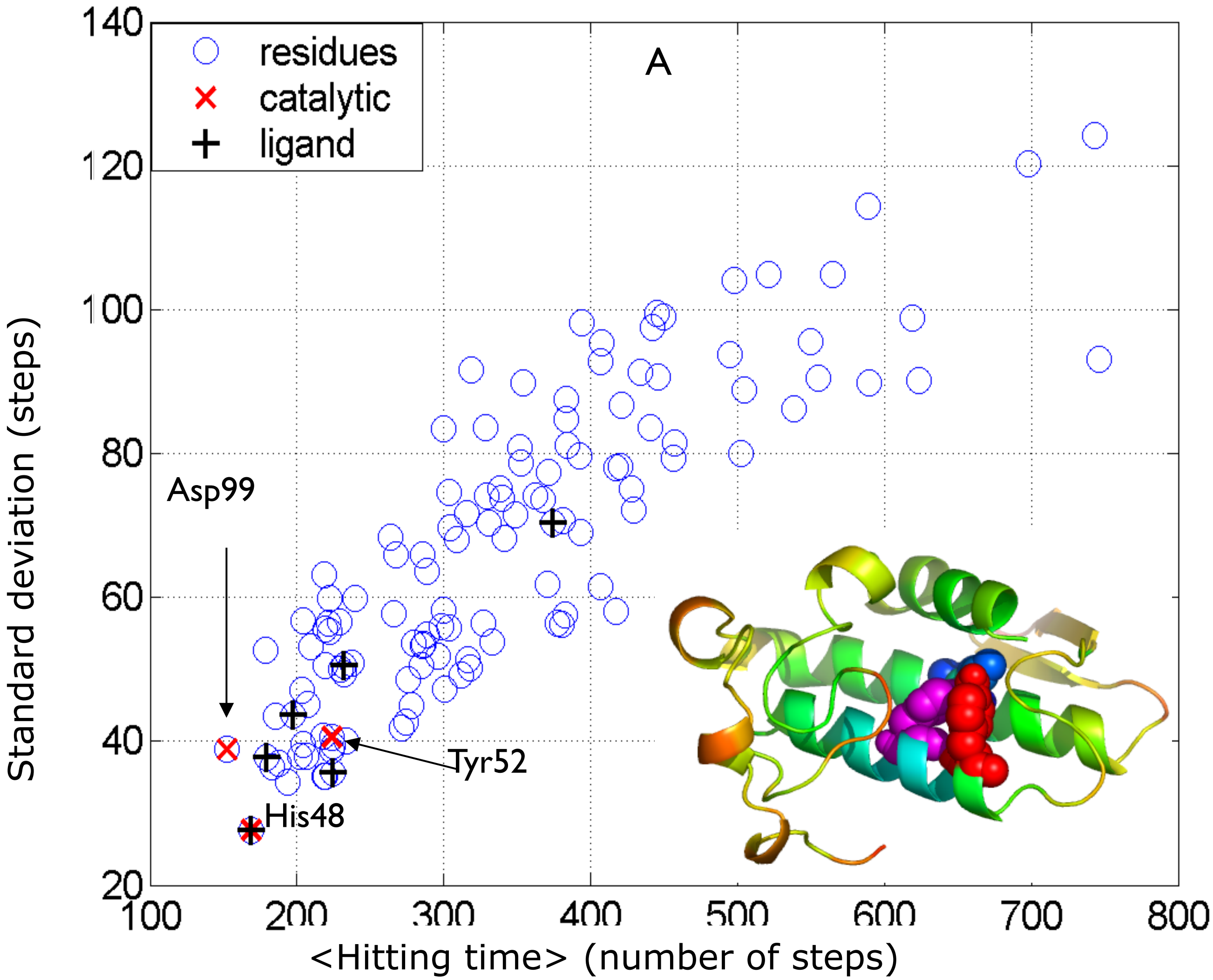


See also

His48 ,Tyr52, Asp99 – catalytic residues

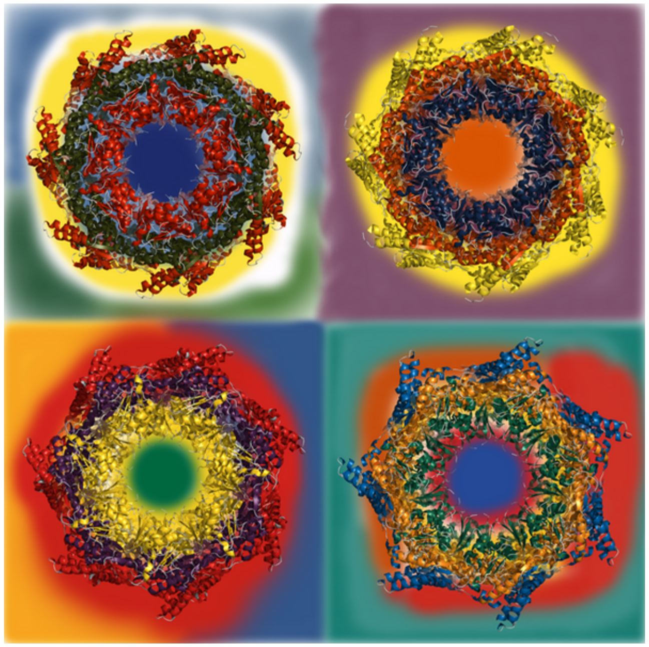
Nadler, Lafon, Kevrekidis & Coifman (2005) Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators, NIPS 18; Coifman et al (2005) PNAS 102, 7426.

Active sites are distinguished by effective communication properties



Chennubhotla & Bahar (2007) *PLoS Comp Bio*

CONCLUSION



- Proteins are designed to favor functional changes in their structure. Pre-existing soft modes facilitate substrate binding.

- Collective mechanics/allosteric dynamics are mediated by conserved residues

- The intrinsic motions confer enhanced flexibility at substrate recognition sites

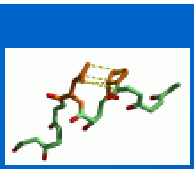
- Correlated mutations at recognition sites enable substrate specificity while conferring conformational adaptability

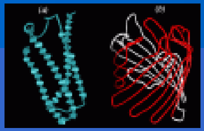
- Accurate modeling of protein dynamics is essential to assessing target druggability

Mechanics vs chemistry?

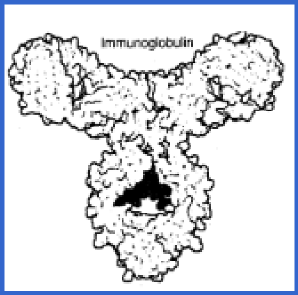
How does complexity scale with size of the system?

Increasing specificity/chemistry)

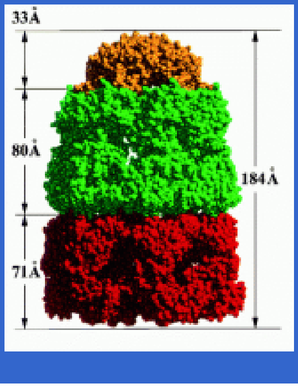


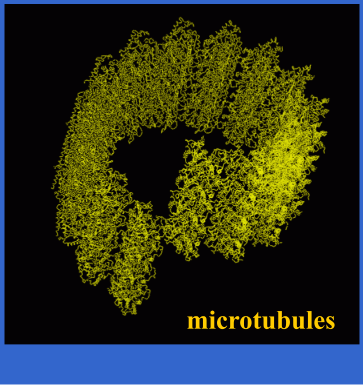


Cytochrome c



Tryptophan Synthase





Residues	Proteins
10	-
10^2	1
10^3	2-10
10^4	10-100

Dominance of molecular machinery

DISCUSSION

- Different tools for different time/length windows: MD cannot explore long-time processes for multimeric systems; ANM does not incorporate detailed atomic forces

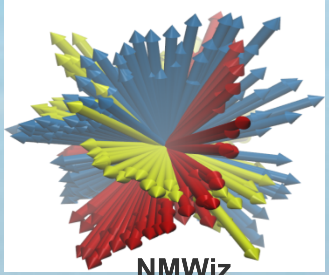
- Not all evolutionarily correlated sites refer to structural or dynamic correlations

- Accurate modeling of protein dynamics is essential to computer-aided drug discovery, but not sufficient for quantitative evaluation of binding affinity

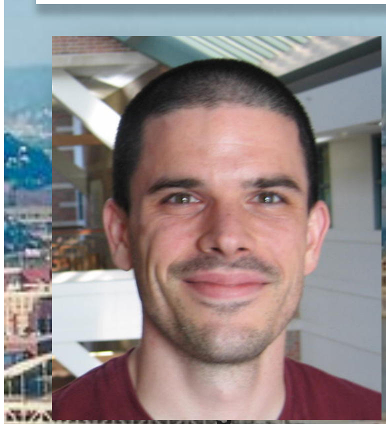
- Druggability simulations identify druggable sites, but not the type of drugs that optimally bind those sites

ProDy

Protein Dynamics Analysis in Python



Dr. Indira Shrivastava
Assist Prof, DCSB, Pitt



Markus Dittrich, PhD
NRBSC Group Leader
Pittt Supercomputing Center

Dr. Timothy R Lezon
Assistant Prof, DCSB, Pitt

Drs. Ahmet Bakan and Anindita Dutta

Dr. Chakra Chennubhotla
Assist Prof, DCSB, Pitt

Acknowledgment: NIH - 5 P41 GM10371202