Computing structural and dynamic properties of biological systems at multiscale

1. Protein geometry, evolution, and function

Jie Liang 梁杰

Dept. of Bioengineering
University of Illinois at Chicago

Outline

1. Structures and geometry of proteins

2. Model of molecular evolution

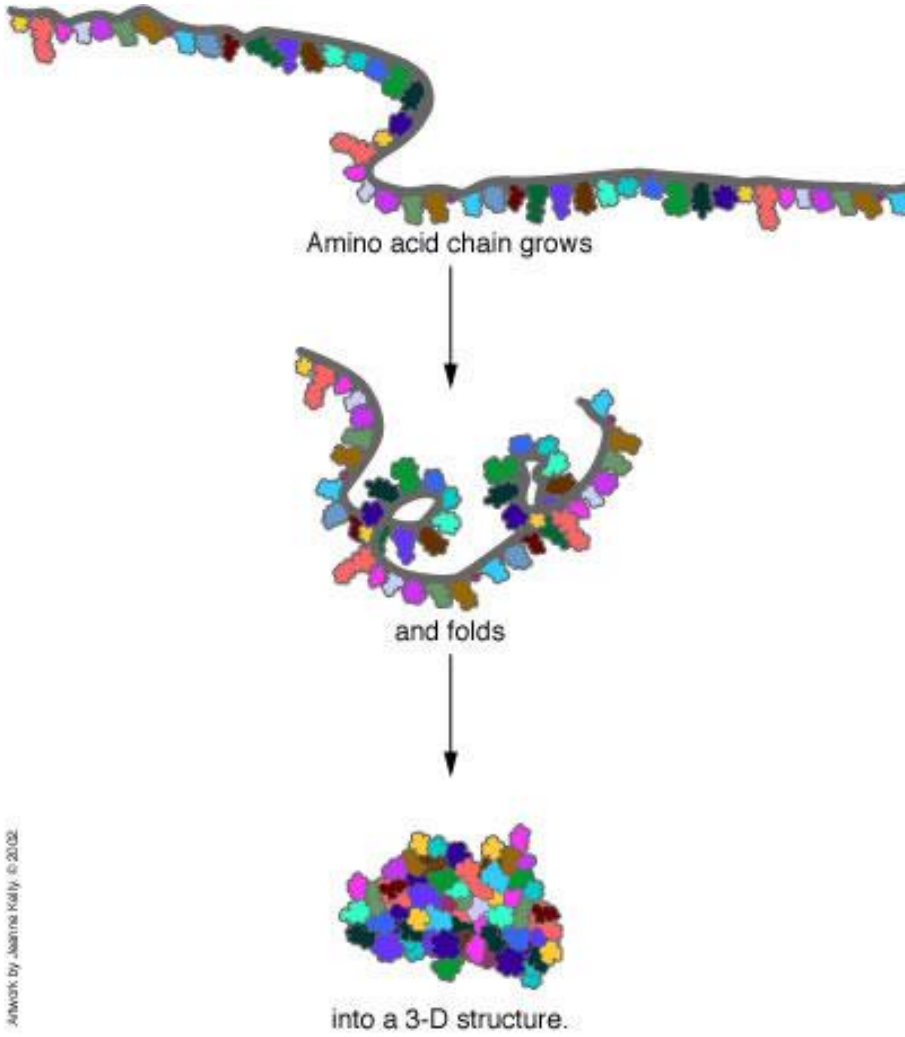
3. Markov chain Monte Carlo for parameter estimation
   - Bayesian Monte Carlo

4. Application in protein function prediction and evolution of metabolic pathway

1. Structures and geometry of proteins

Protein Sequence & Structure

- Proteins as linear heteropolymers:
  - Fixed **number** and **composition** of monomers,
  - Monomers are amino acid residues of 20 types.
  - Size: from a few tens to approximately a thousand residues.
  - Linear chains fold into specific three-dimensional conformations.
  - Only a tiny fraction of sequences will fold

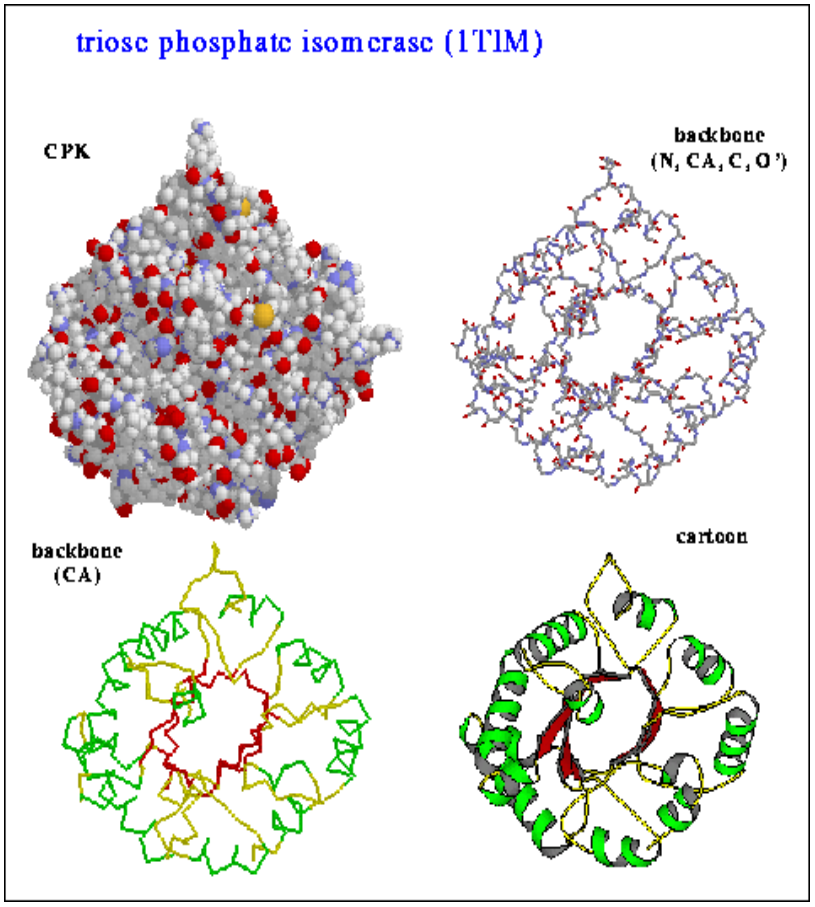


Genetic blue prints

Working molecules of cell

70,000 Structures

Different structural models of proteins

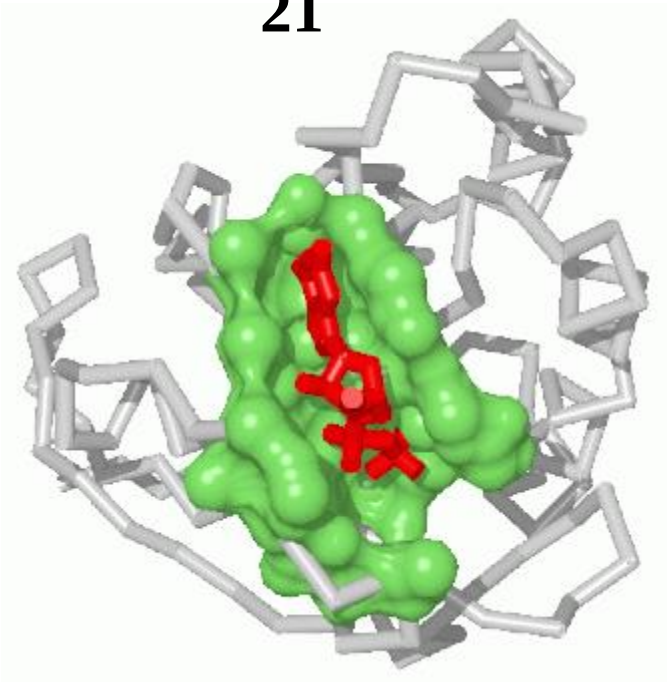


Volumetric and surface models

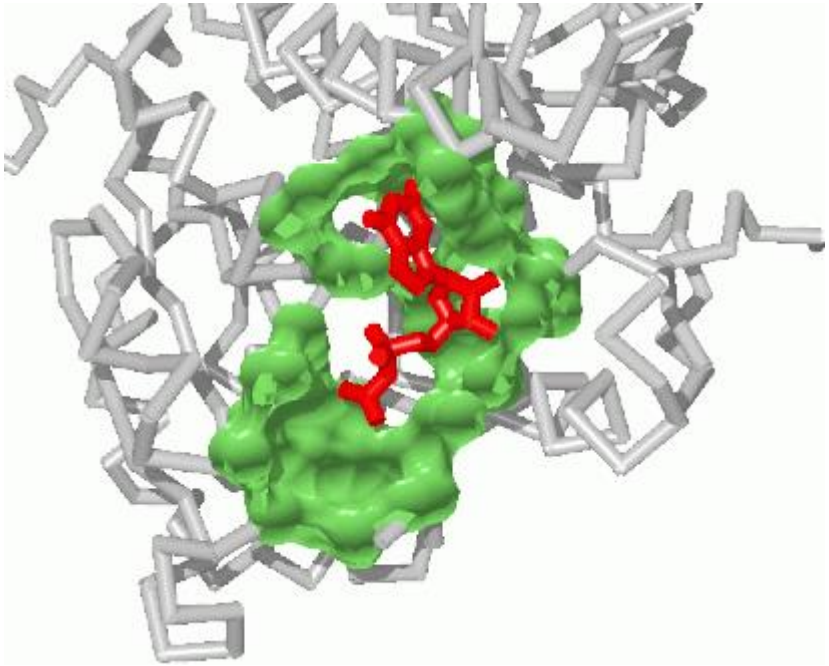
- Backbone centric view
  - Secondary structure, tertiary fold, side chain packing
- But ligand and substrate sees differently!
  - We are interested in things like binding surfaces
- Volumetric and surface models
  - Much more complicated, as there could be 10,000 atoms.

Functional Voids and Pockets

Ras
21



Fts Z



GDP Binding Pockets

Space-filling Model of Protein

- The shape of a protein is complex
  - Properties determined by distribution of electron charge density,
  - Chemical bonds transfer charges from one atom to another
    - Isosurface of electron density depend on locations of atoms and interactions
  - X-ray scattering pattern are due to these distributions.

- Space Filling model: Idealized model
  - Atom approximated by balls, difference between bonded and nonbonded regions ignored.

    "interlocking sphere model", "fused ball model"
  - Amenable for modeling and fast computation
  - Ball radius: many choices, eg. van der Waals radii

(B. Lee, F. M. Richards, 1971 ; F. M. Richards, 1985)

Mathematical Model: Union of balls

- For a molecule M of n atoms, the i-th atom is a ball b_i, with center at $\mathbf{z}_i \in R^3$

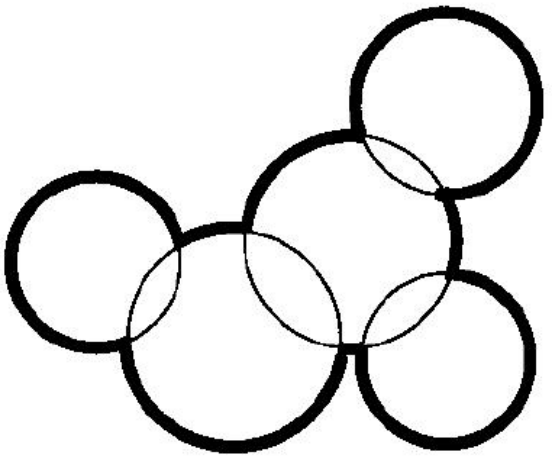
$$b_i \equiv \{\mathbf{x} \mid \mathbf{x} \in R^3, |\mathbf{x}-\mathbf{z}_i| <= r_i\},$$

parameterized by $(\mathbf{z}_i, r_i)$.

- Molecule M is formed by the union of a finite number n of such balls defining the set **B:**

$$\mathbf{M} = \bigcup \mathbf{B} = \bigcup_{i=1}^{n} \{\mathbf{b}_i\}$$

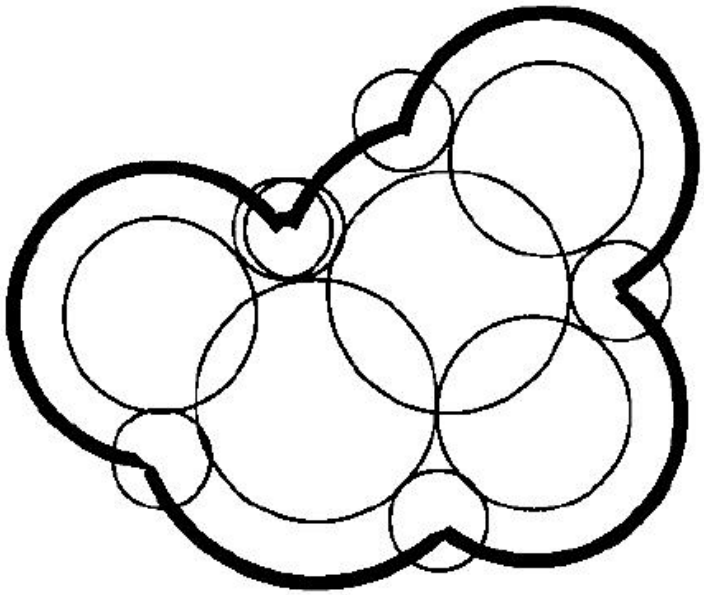
  - Creates a space-filling body corresponding to the volume of the union of the excluded volume
  - When taken vdw radii, the boundary $\partial \bigcup \{B\}$ is the **van der Waals surface**.



(Edelsbrunner, 1995; see also Liang et al, 1998)

Solvent Accessible Surface Model

- Solvent accessible surface (SA model):
  - Solvent: modeled as a ball
  - The surface generated by rolling a solvent ball along the van der Waals atoms.
  - Same as the vdw model, but with inflated radii by that of the solvent radius

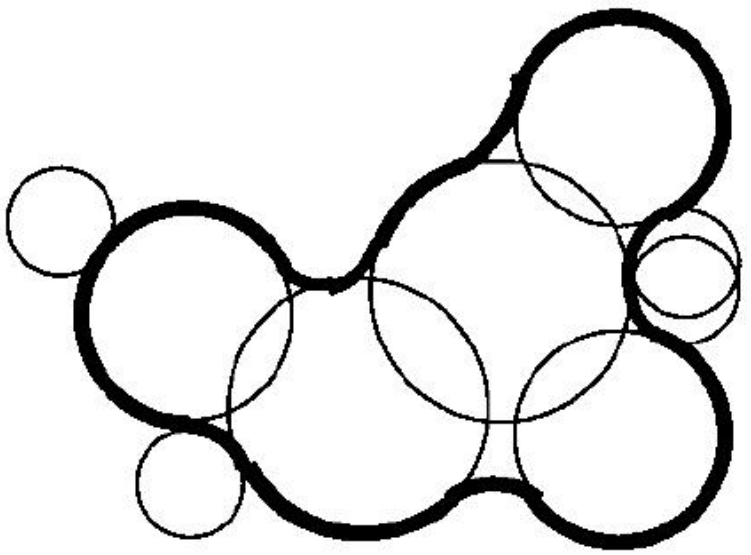


(B. Lee, F. M. Richards, 1971)

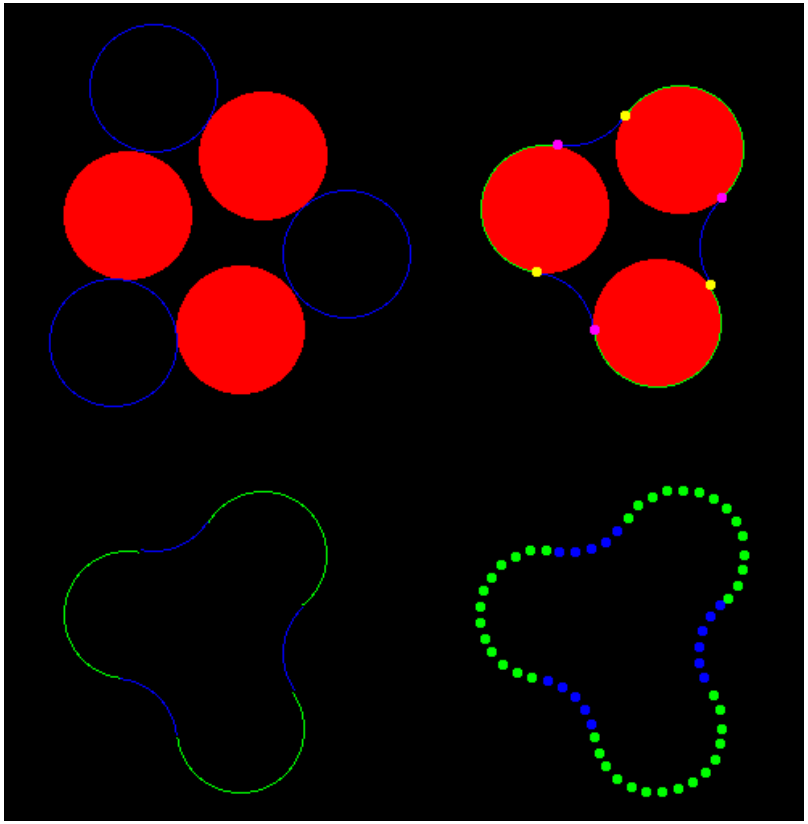
Molecular Surface Model

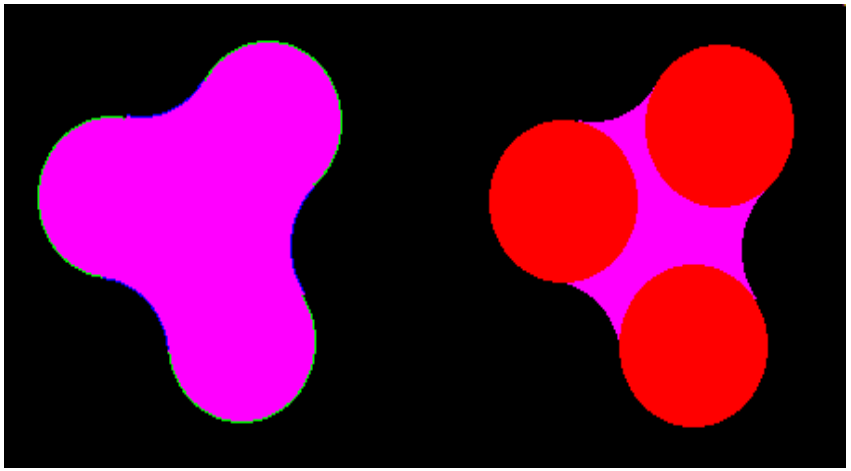
- Molecular Surface Model (MS):

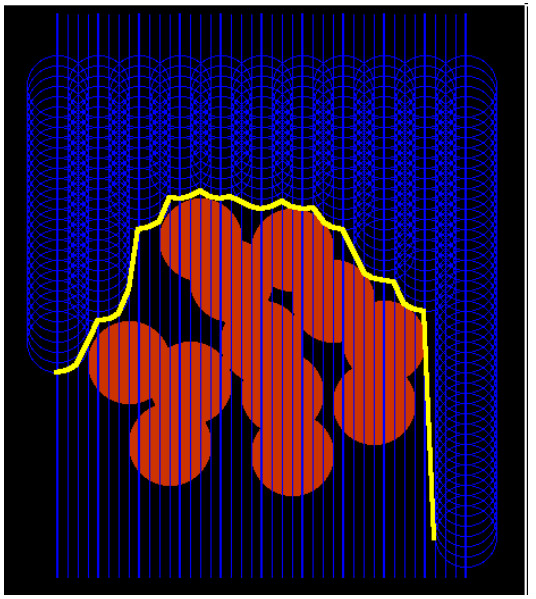
  The surface rolled out by the front of the solvent ball.
  - Also called Connolly's surface.



More on molecular surface model



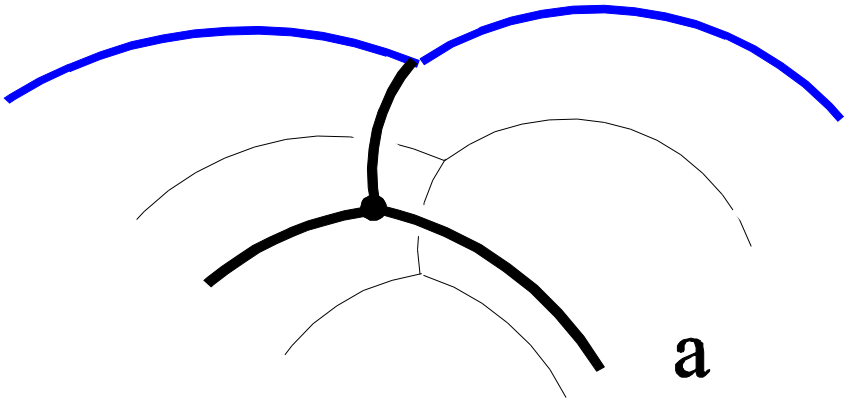




(Michael Connolly, http://www.netsci.org/Science/Compchem/feature14e.html)

Elementary Surface Pieces: SA

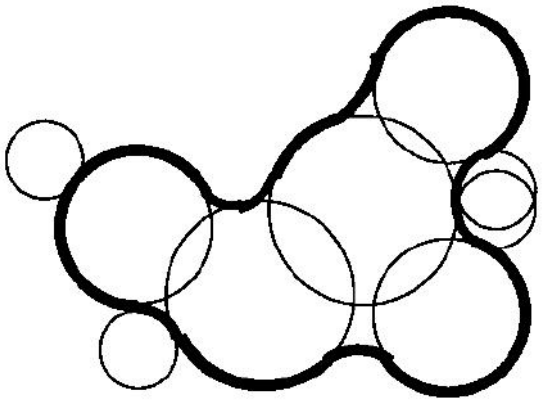
- SA: the boundary surface is formed by three elementary pieces:
  - Convex sphereical surface pieces,
  - arcs or curved line segments
    - Formed by two intersecting spheres
  - Vertex
    - Intersection point of three spheres
- The whole surface: stitching of these three elementary pieces.

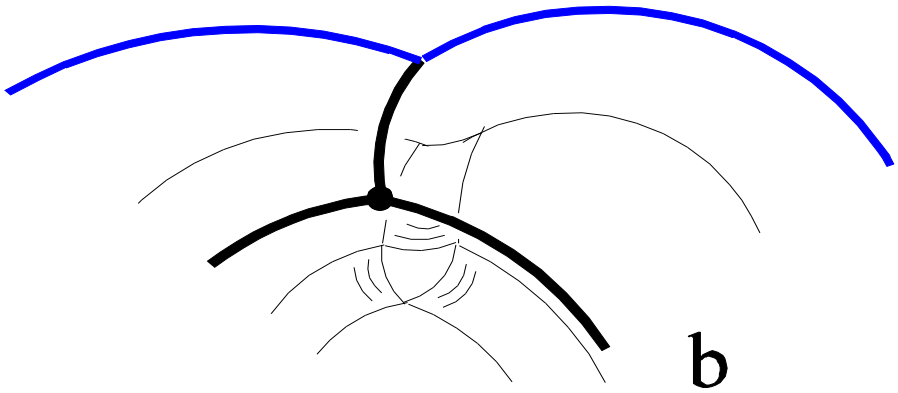


Vdw surface: Shrunken version of SA surface by 1.4 A

Elementary Surface Pieces: MS

- MS: three different elementary pieces:
  - Convex spherical surface pieces,
  - Concave toroidal surface pieces
  - Concave spheric surface
    - The latter two are also called “Re-entrant surface”
- The whole surface: stitching of these three elementary pieces.

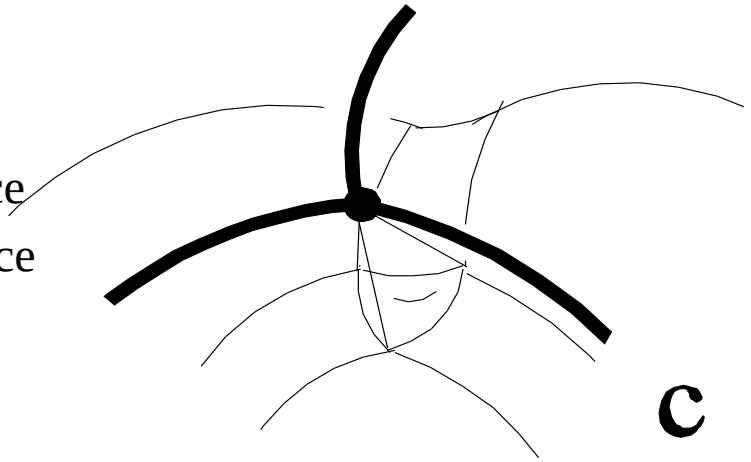




Relationship between different surface models

- vdW and SA surfaces.
- SA and MS surfaces:
  - Shrink or expand atoms.

SA	MS
Vertex	concave spheric surface piece
Arcs	concave toroidal surface piece
Conv. surfade	Smaller conv surface



- SA and MS:
  - Combinatorically equivalent
  - Homotopy equivalent

- But, different metric properties!
  - SA: void of 0-volume ---- MS: void of $4\pi r^3/3$

Computing protein geometry

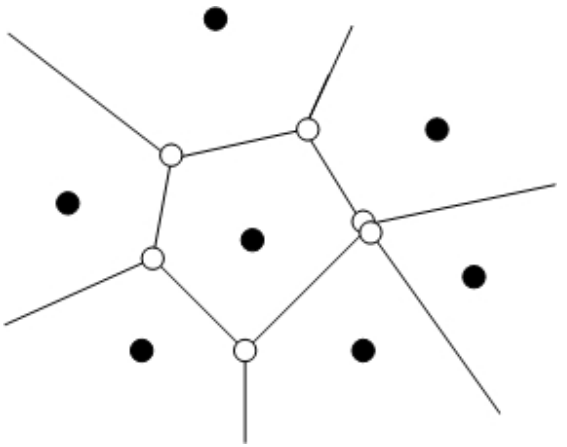
- It is easy to conceptualize different surface models
- But how to compute them?
  - Topological properties
  - Metric properties (size measure)
- Need:
  - Geometric constructs
  - Mathematical structure
  - Algorithms

Geometric Constructs: Voronoi Diagram

- A point set **S** of atom centers in $\mathbb{R}^3$
- The Voronoi region / Voronoi cell of an atom b_i with center $z_i \in \mathbb{R}^3$

$$V_i = \{ x \in \mathbb{R}^3 \mid |x - z_i| \leq |x - z_j|, z_j \in S \}$$

  - All points that are closer to (or as close as to) b_i than any other balls b_j

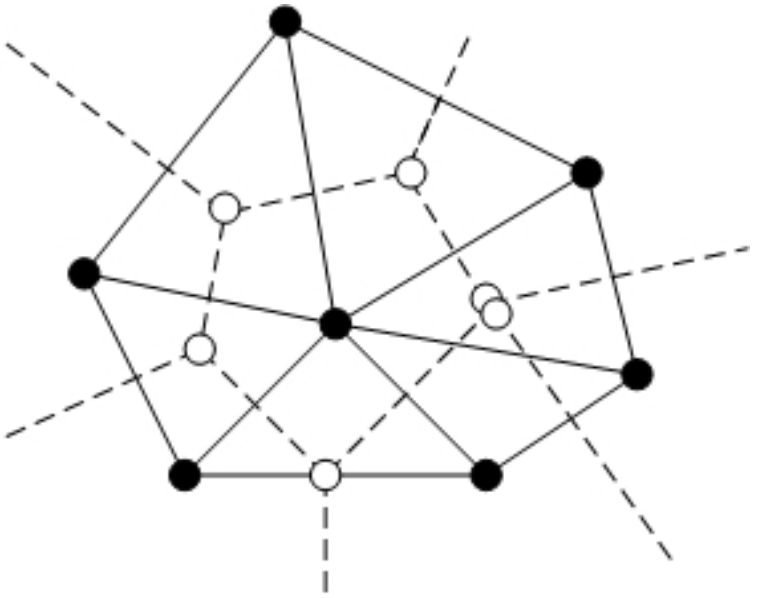


- Alternative view:
  - Bisector plane has equal distance to both atoms, and forms a half space for b_i.
  - Half space of b_i with each of the other balls b_j
  - Intersection of the half spaces forms the Voronoi cell, and is a convex region

(M. Gerstein, F. M. Richards, 1999)
(A. poupon, 2004)

Delaunay Triangulation

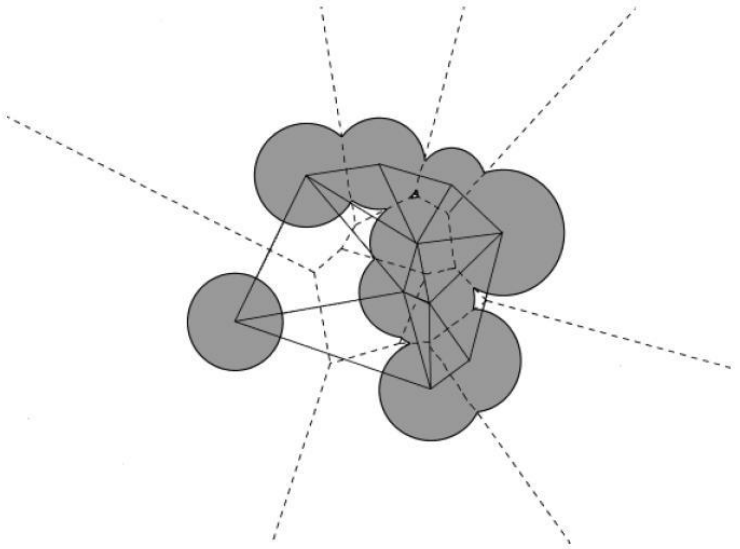
- Convex hull of point set S:
  - The smallest convex space contain all points of S.
  - It is formed by intersection of halfplanes, and is a convex polytope.



- Delaunay triangulation:
  - uniquely tesselate/tile up the space of the convex hull of a point set with tetrahedra, together with their triangles, edges, and vertices
  - (triangles instead of tetrahedra in 2D)

Dual relationship between Voronoi and Delaunay

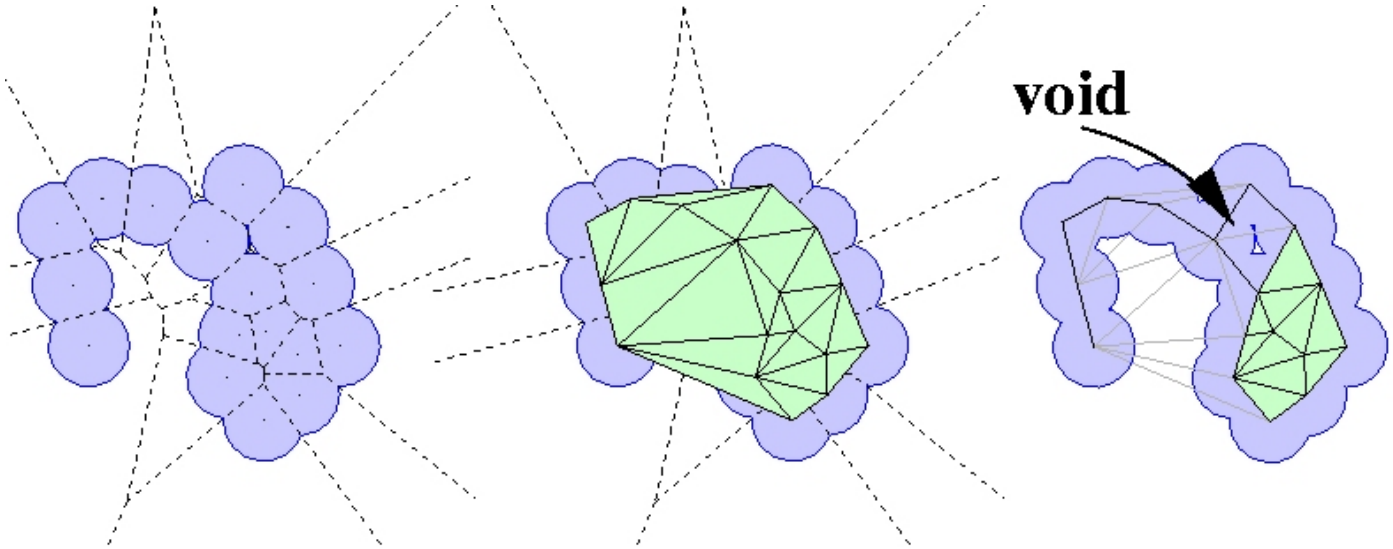
- These two geometric constructs look very different!
- In fact, they are dual to each other
  - Reflect the same combinatorial structures



(Edelsbrunner, 1995; Liang et al, 1998a; Liang et al, 1998b)

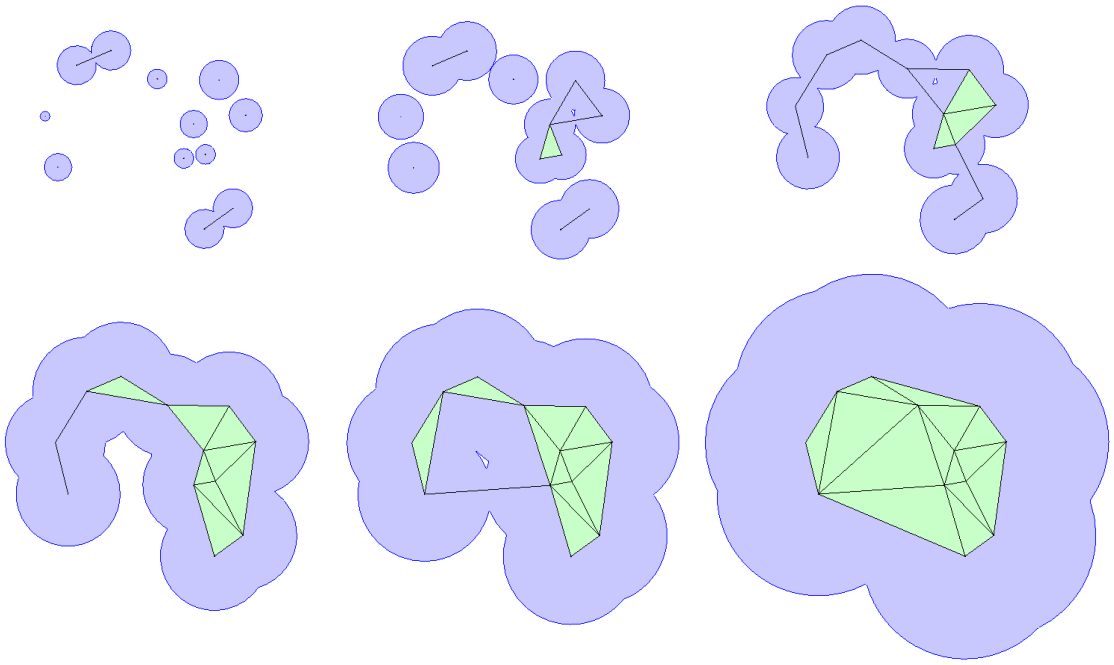
Dual Relationship and Void

- Geometric structure from alpha shape

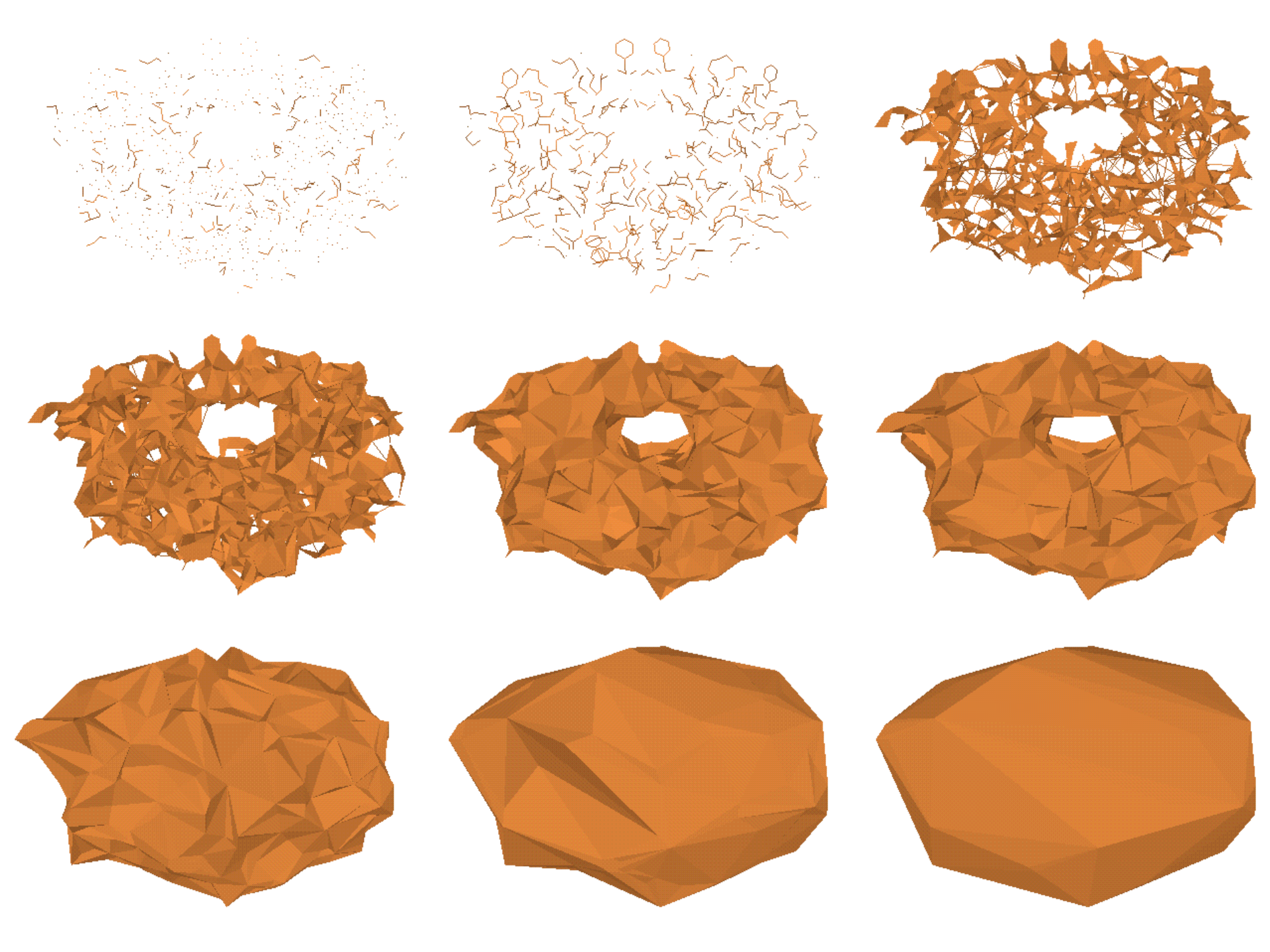


Mucke & Edelsbrunner, 1994, *ACM Tran Graph*
Edelsbrunner, Facello, Liang, 1998, *Disc Appl Math*
Liang, Edelsbrunner, Woodward, 1998, *Protein Sci*

A series of 2D simplicial complexes (alpha shapes).

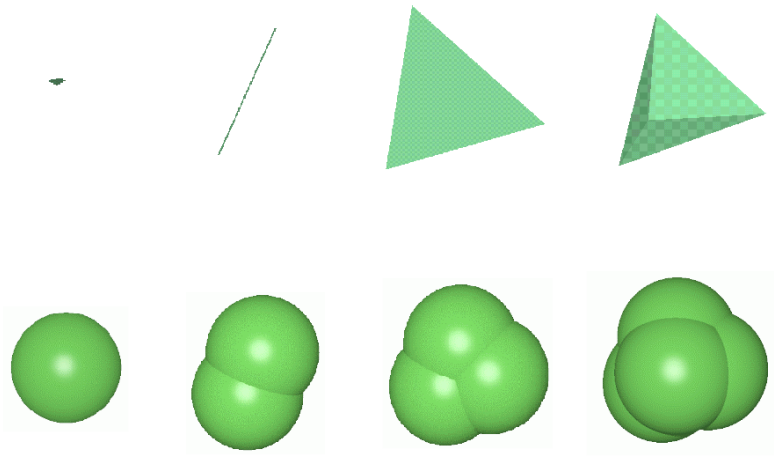


Each faithfully represents the geometric and topological property of the protein molecule at a particular resolution parametrized by the α value



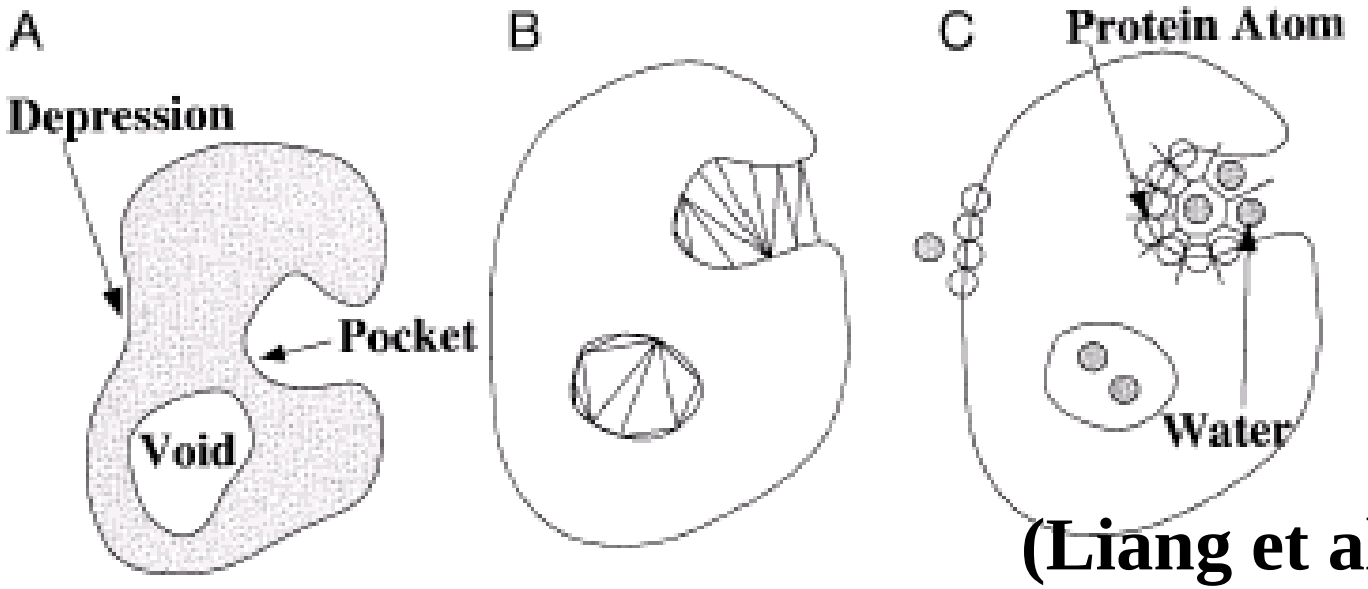
A Guide Map for Computing Geometric Properties

- Combinatorial Information



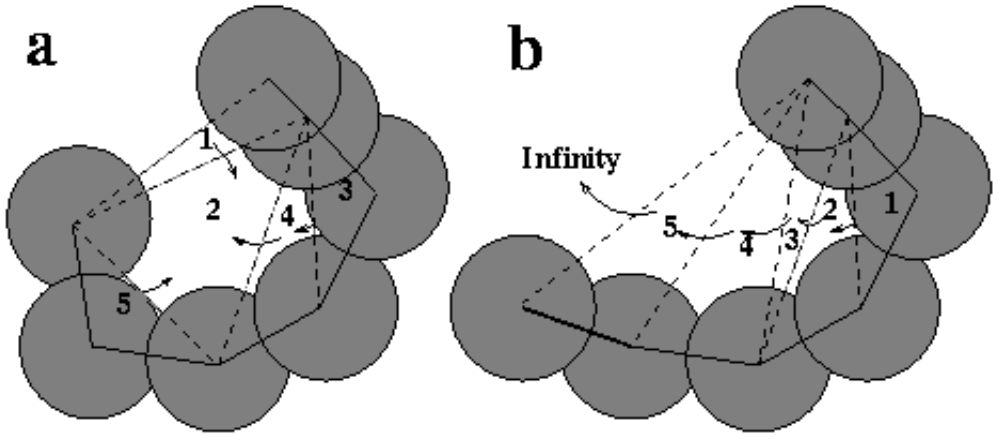
Voids and pockets in proteins

- Concave regions on protein surfaces
- Shape complimentarity important for molecular recognition
  - Binding frequently occurs in p0ckets and voids
  - Eg. enzymes



(Liang et al, 1998a)

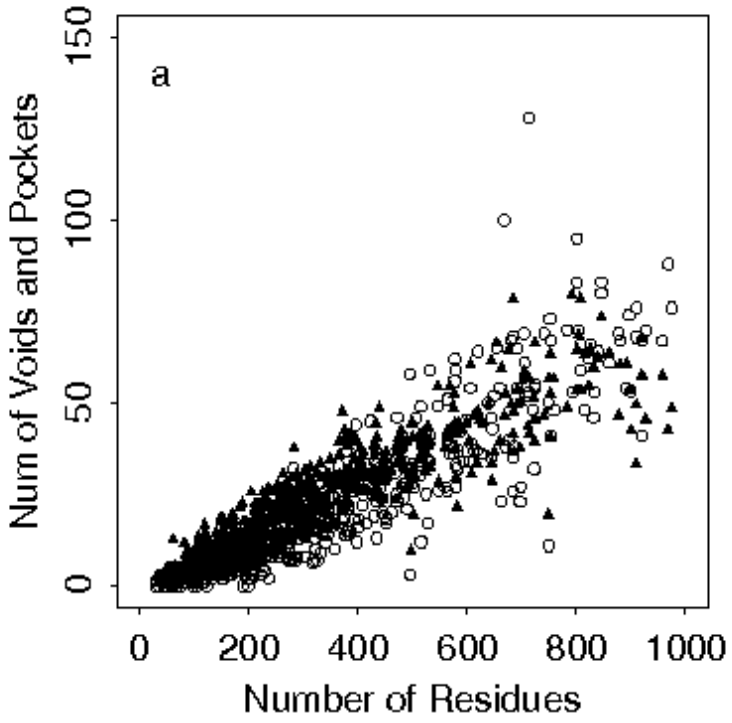
Computing pockets in proteins



Edelsbrunner et al, 1998, *Disc Appl Math*
Liang, Edelsbrunner, Woodward, 1998, *Protein Sci*

Voids and Pockets in Soluble Proteins

- "Protein interior is solid-like, tightly packed like a jig-saw puzzle"
  - High packing density (Richards, 1977)
  - Low compressibility (Gavish, Gratoon, and Harvey, 1983)
- Many voids and pockets.
  - At least 1 water molecule; 15/100 residues.



(Liang & Dill, 2001, Bioph J)

Origin of Voids and Pockets in Proteins

- Do *compact random chain polymers* pack like proteins?
  - Off-lattice models.

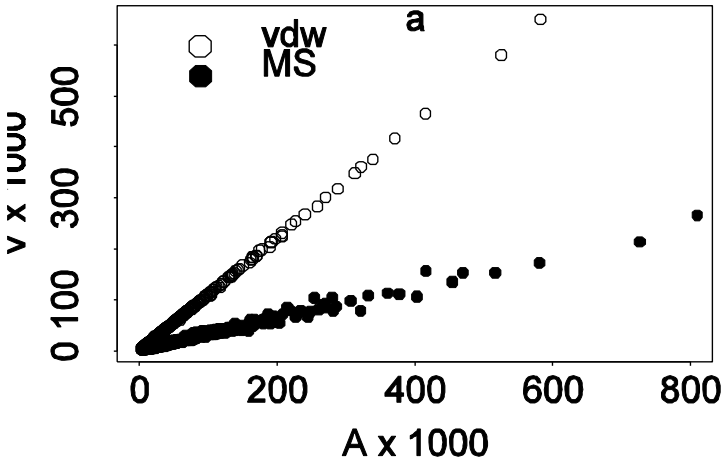
Scaling relationship

- Volume and area scaling:

  $V = 4 \pi r^3/3$ and $A = 4 \pi r^2$, therefore we should have

  $$V \approx A^{3/2}$$

- Protein has linear scaling:
  - Clustered random sphere with mixed radii (Lorenz et al, 1993).
  - Lattice models of simple clusters (Stauffer, 1985)

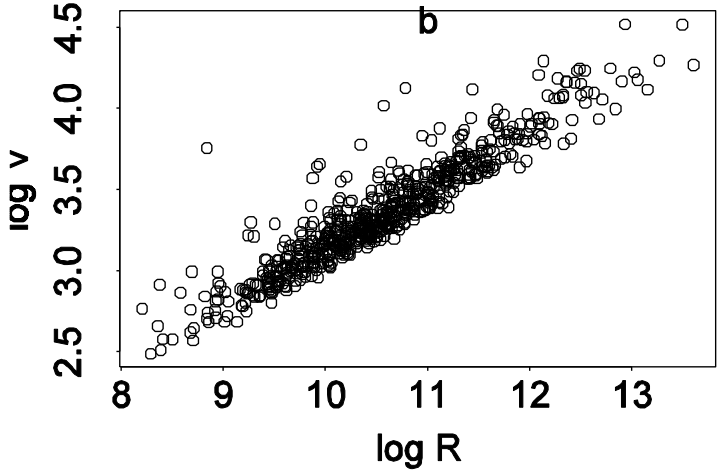


(Liang & Dill, 2001, Bioph J)

Scaling relationship of proteins

- At percolation threshold, V and R of a cluster of random spheres:
  - $V \propto R^D$, where D = 2.5 (Stauffer, 1983; Lorenz et al 1993)

$$R = \sum_j^d (x_{j,\max} - x_{j,\min})/2d$$

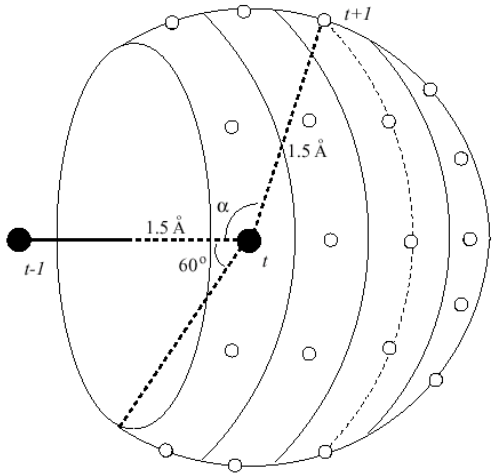


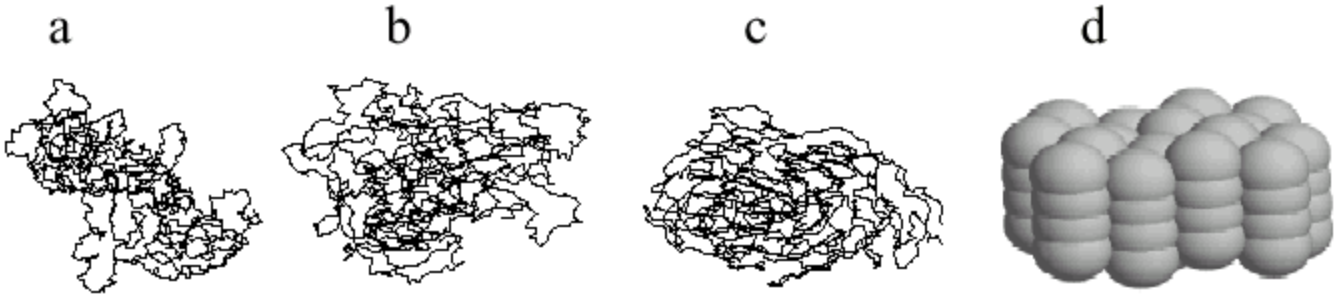
- Proteins:
  - ln V » ln R, D = 2.47 ±0.04 (by nonlinear curve fitting).
  - Similar to random spheres near percolation threshold.

By volume-area and volume-size scaling, proteins are packed more like random spheres than solids.

Simulating Protein Packing with Off-Lattice Chain Polymers

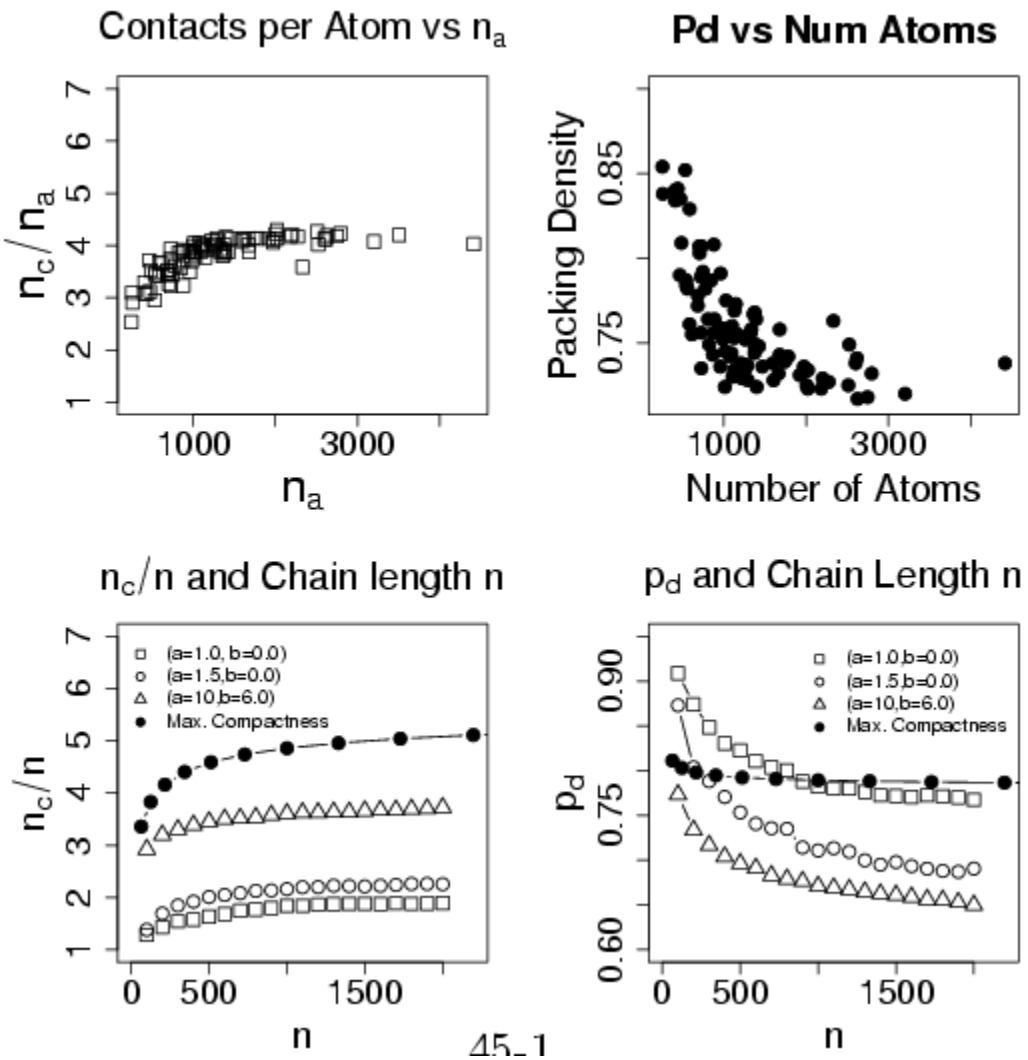
- 32-state off-lattice discrete model
- Sequential Monte Carlo and resampling:
  - 1,000+ of conformations of N = 2,000





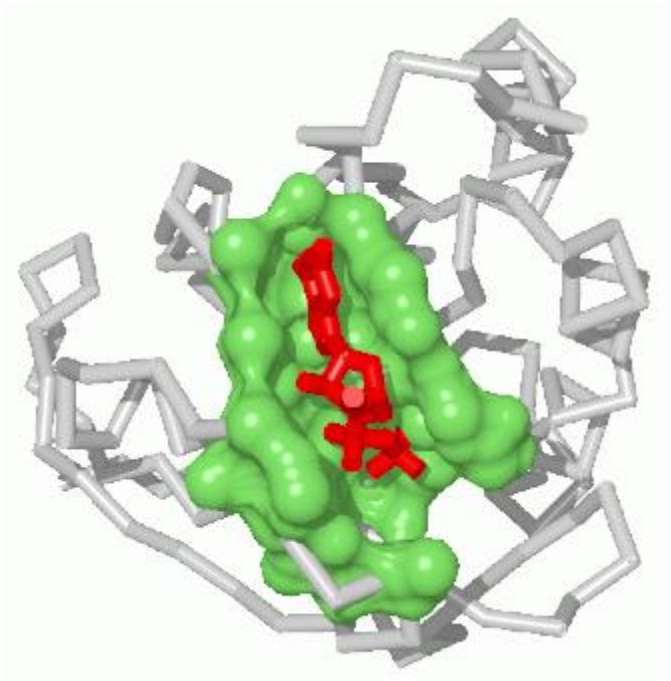
(Zhang, Chen, Tang and Liang, 2003, *J. Chem. Phys*. **118(13):6102-6109**)

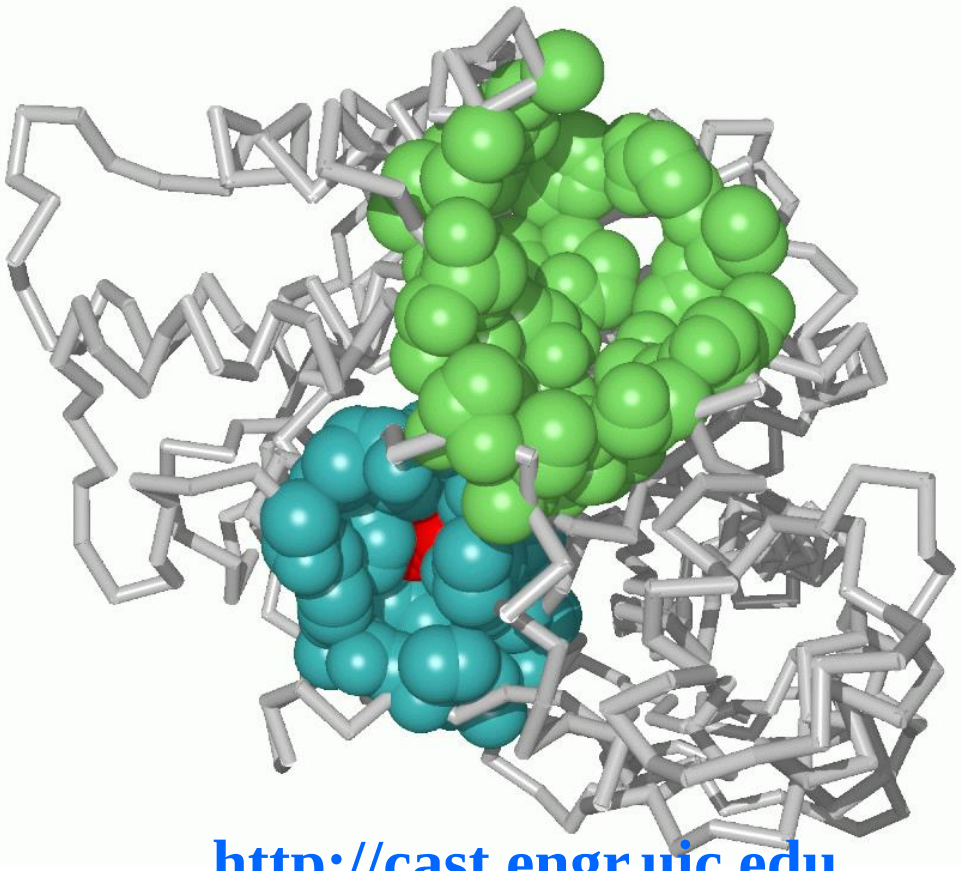
- Proteins are not optimized by evolution to eliminate voids.
  - Protein dictated by generic compactness constraint related to n_c.



(Zhang, Chen, Tang and Liang, 2003, *J. Chem. Phys*. **118(13):6102-6109**)

Surfaces with unknown functional roles



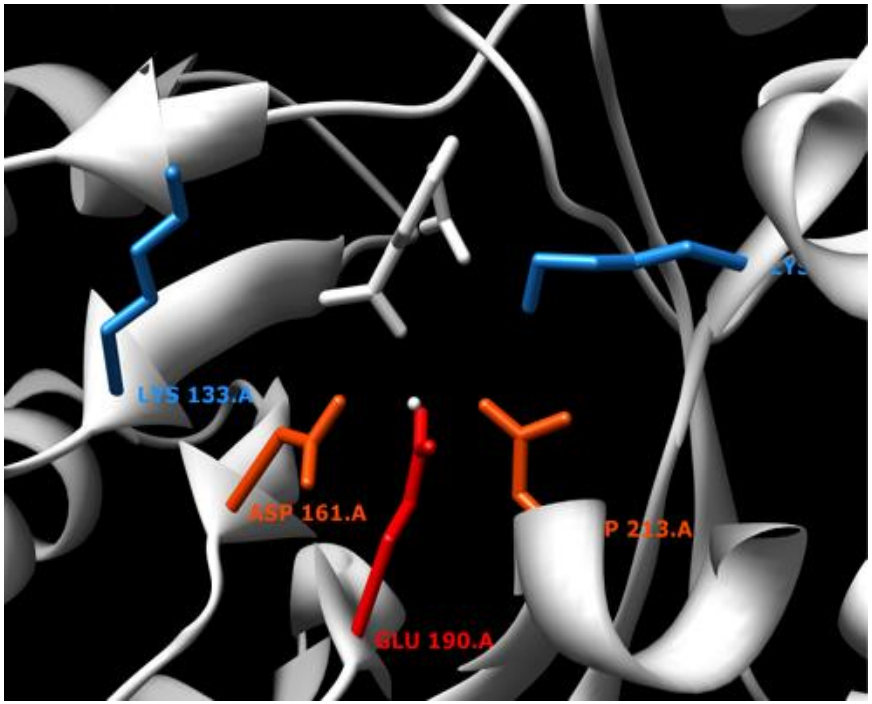


http://cast.engr.uic.edu

(Dundas et al, 2006)

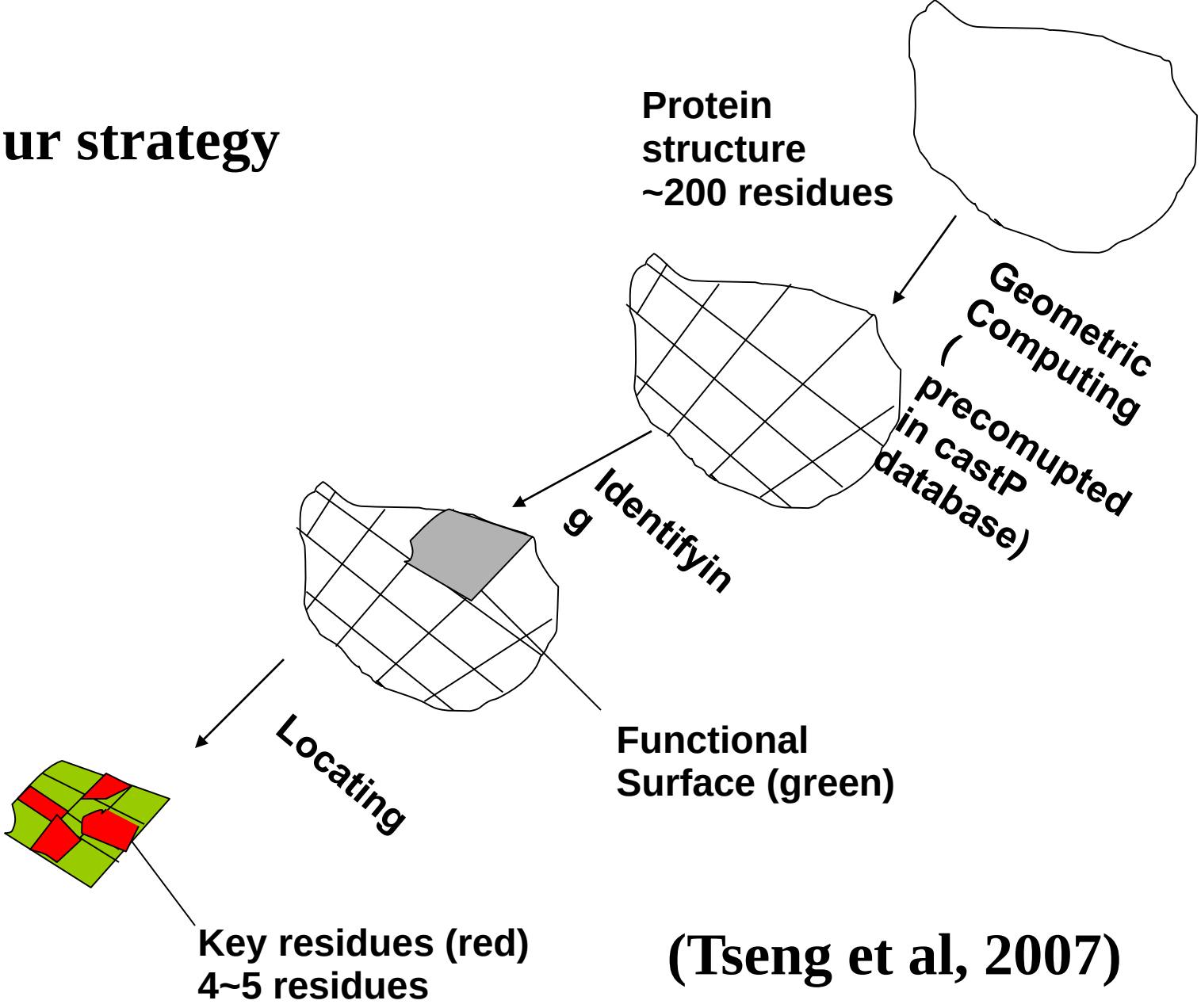
Enzyme Functional Site Prediction

- Where are the functional surfaces located ?
- What are the key residues (active site residues) in the functional surfaces ?
  - Which mutations to make?



(Tseng et al, A*nnals of Biomedical Engineering, 2007*, 35(6):1037-1042)

Our strategy



(Tseng et al, 2007)

Structure is the only input!!

Validation study: large-scale prediction of functional surface

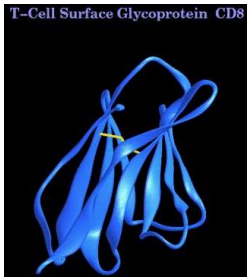
- Data Set (~700 structures).
- 10-fold cross validation.
- Average accuracy of predicting functional surfaces of proteins is 91.2%.

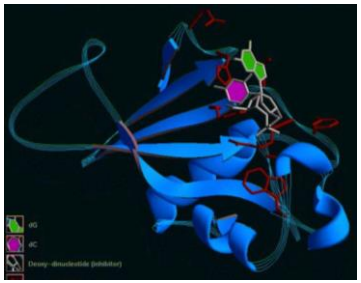
(Tseng et al, 2007)

The Universe of Protein Structures

- Human genome: 3 billion nucleotides;
  - Number of genes: 20,000 – 25,000; Protein families: 10,000; Number of folds: 1,000s
- Currently in PDB: about 1,000 folds
  - Comparative modeling: needs a structural template with sequence identities > 30-35%
    - eg. ~50% of ORFs and ~18% of residues of *S. cerevisiae* genome

All β **α/β**





(from SCOP)

- Structural Genomics: populating each fold with 4-5 structures
  - One for each superfamily at 30-35% sequence identities.
  - Fold of a novel gene can be identified
    - Its structure can then be interpolated by comparative modeling.

(A. Heger and L. Holm, 2000)

Predicting and characterizing protein functions

- Important, but challenging tasks:
  - Needs > 60-70% sequence identity.
  - Fold prediction: >20-30% sequence identity.

(Rost, 02, JMB; Tian & Skolnick, 03, JMB)

- Proteins from structural genomics often are of unknown functions.
  - Sequence homologs are often hypothetical proteins.

Another Way: How to identify biologically important pockets and voids from random ones?

By Homology:

Assessing Local Sequence and Shape Similarity

(Binkowski, Adamian, Liang, 2003, *JMB*, 332:505-526)

Binding Site Pocket: Sparse Residues, Long Gaps

ATP Binding: cAMP Dependet Protein Kinase (1cdk) and Tyr Protein Kinase c-src (2src)

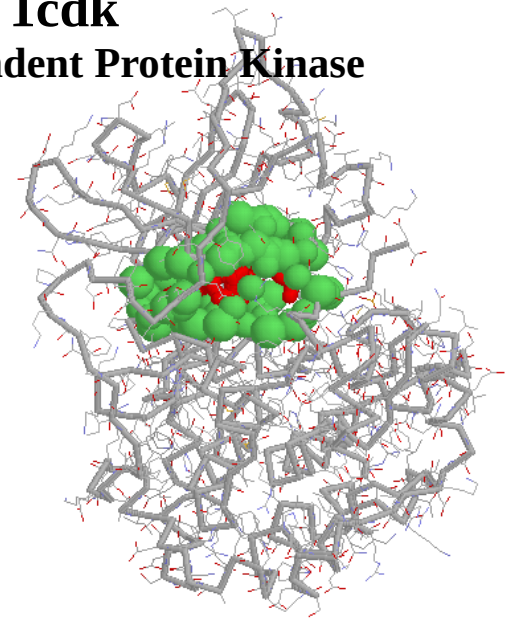
```
1cdk.A
   49LGTGSFGRVMLVKHKETGNHFAMKILDKQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEYSFKDNSNL
   YMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFG
   FAKRVKGRTWTLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFADQPIQIYEKIVSGKVR
   FPSHFSSDLKDLLRNLLQVDLTKRFGNLKDGVNDIKNHKWFATTDWIAIYQRKVEAPFIPKFKGPGDTSN
   F327
1cdk.A_p
   49LGTGSFGRV          A K                          V
      MEYV   E                                   K EN L          TD
   F


2src.m
   273LGQGCFGEVWMGTWNGTTRVAIKTLKPGTMSPEAFLQEAQVMKKLRHEKLVQLYAVVSEEPIYIV
   TEYMSKGSLLDFLKGETGKYLRLPQLVDMAAQIASGMAYVERMNYVHRDLRAANILVGENLVCKVAD404
2src.m_p
   273LGQGCFGEV          A K                          V
   TEYM  GS   D                                  D R AN L          AD
```

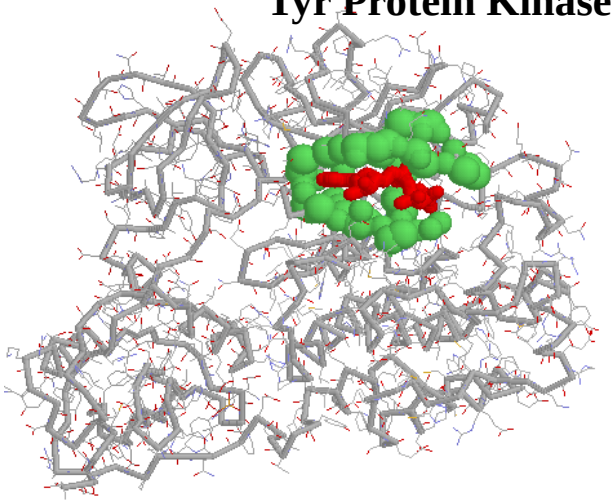
Low overall sequence identity: 13 %

High Sequence Similarity of Pocket Residues

1cdk
cAMP Dependent Protein Kinase



2src
Tyr Protein Kinase c-src



```
1cdk.A LGTGSFGRVAKVMEYV---EKENLTDF 24
2src.m LGQGCFGEVAKVTEYMGSDDRANLAD- 26
       ** *.**.**** **:    :: **:*
```

High sequence identity: 51 %

Sequence Similarity of Surface Pockets

- Similarity detection:
  - Dynamic programming SSEARCH (Pearson, 1998)
  - Order Dependent Sequence Pattern.

- Statistics of Null Model:

  → ***Statistical Significance !***

  - Gapless local alignment: Extreme Value Distribution (Altschul & Karlin, 90)
  - Alignment with gaps: (Altschul, Bundschuh, Olsen & Hwa, 01)

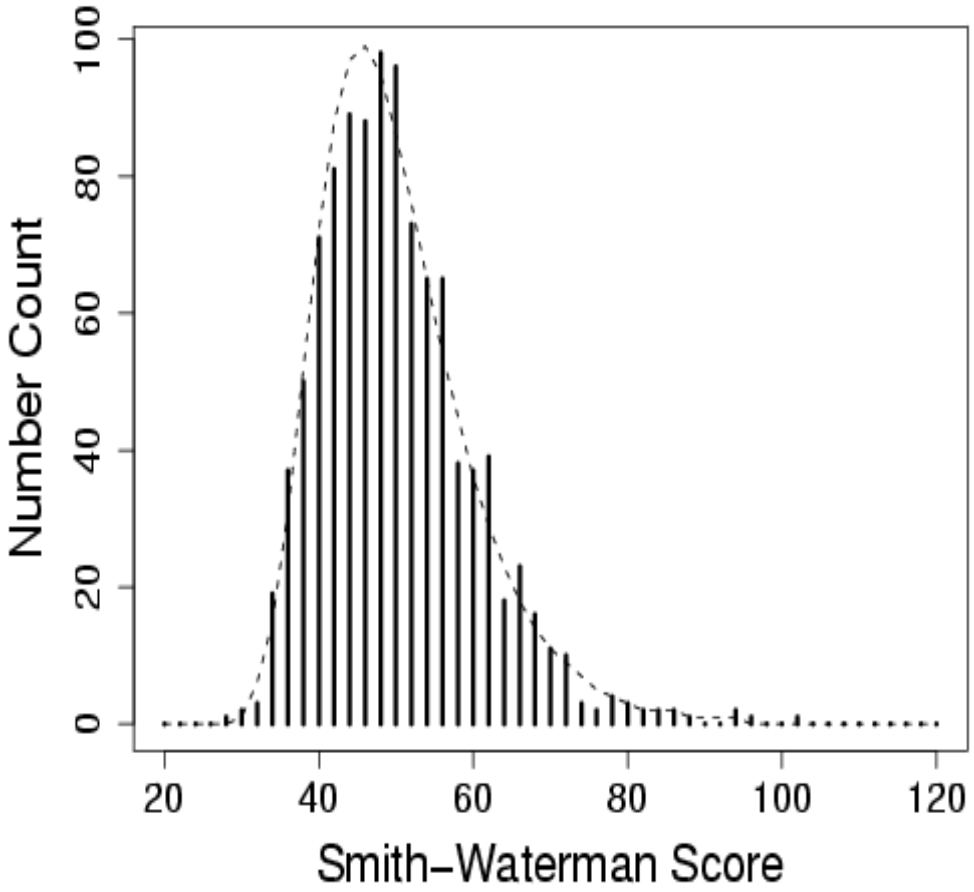
Approximation with EVD distribution *(Pearson, 1998, JMB)*

- Kolmogorov-Smirnov Test:
  - Estimate K and λ parameters.

$$S = \ln K + \ln mn$$

$$P(S \geq x) = 1 - \exp(-e^{-\lambda x})$$

- Estimation of E-value:
  - Estimate p value of observed Smith-Waterman score by EVD.



(Binkowski, Adamian, Liang, 2003, JMB, 332:505-526)

$$E = p(N_{all}) p_a$$

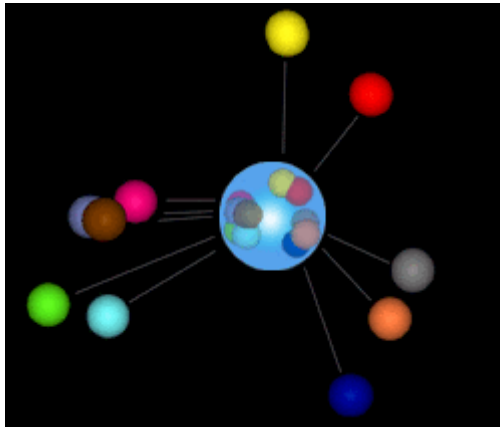
(Pearson, 1998, JMB)

Shape Similarity Measure

- cRMSD (coordinate root mean square distance)
- oRMSD (Orientational RMSD):
  - Place a unit sphere S^2 at center of mass $\mathbf{x}_0 \in R^3$
  - Map each residue $\mathbf{x} \in R^3$ to a unit vector on S^2 :

$$f : \mathbf{x} = (x, y, z)^T \mapsto \mathbf{u} = (\mathbf{x} - \mathbf{x}_0) / \| \mathbf{x} - \mathbf{x}_0 \|$$

  - Measuring RMSD between two sets of unit vectors.



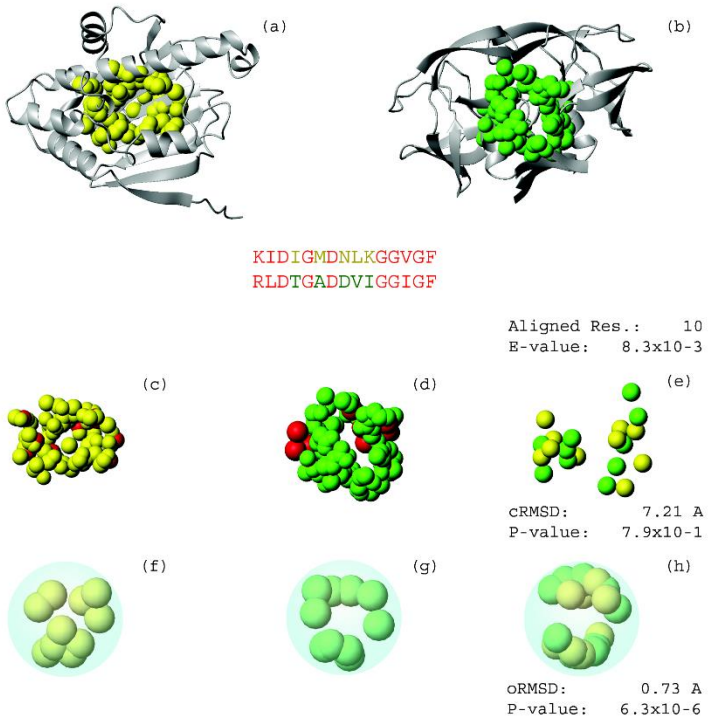
(cf. uRMSD by Kedem and Chew, 2002)

Surprising Surface Similarity

HIV-1 Protease (5hvp)		
CATH	Class	All β
	Fold	Acid proteases
	Family	Retroviral protease
Pocket	Binds poly-peptide substrate acetyl-pepstatin	

Heat Shock Protein 90 (1yes)		
CATH	Class	α+β
	Fold	α/β sandwhich
	Family	Hsp90
Pocket	Binds protein segment geldanamycin	

(Binkowski et al, 2003)



- **Conserved residues both important in polypeptide binding**
- **Both pockets undergo conformational changes upon binding**

2. Model of Evolution

How to capture evolutionary signals due to biological functions?

(Tseng and Liang, 2006, *Mol Biol Evo*, 23:421; Tseng et al, 2009, *J. Mol.Biol*)

- Strong selection pressures that increase the overall fitness of proteins.

- But may not be related with biological function:
  - Structural constraints
  - Protein stability
  - Folding kinetics

- Isolating selection pressure due to biological function:
  - Unsolved problem!

Deriving Scoring Matrix from Evolutionary History

- A scoring matrix is critical:
  - Determines similarity between residues and hence statistical significance.
  - Derived from evolutionary history of proteins sharing the same function.
- Existing approach:
  - PAM and BLOSUM heuristics.
  - Position specific weight matrix.
  - Entropy/relative entropy for full proteins or domains.
- Our approach:
  - Evolution: Explicit phylogenetic tree.
  - Model: Continuous time Markov process.
  - Geometry: Evolution of only residues located in the binding region.
  - Bayesian Markov chain Monte Carlo.

Evolutionary Model

- Assuming no insertion and deletion
- Relationship between proteins (species) can be described by a phylogenetic tree
  - Binary tree:
    - No multifurcation
  - Ignore horizontal transfer of genes
- Residue substitution follows a Markovian process
- Assuming time reversibility

R Nielsen, Z Yang - Genetics, 1998

Model: Continuous time Markov process for substitution

20 × 20 rate matrix Q for the instantaneous substitution rates of 20 amino acid residues

- Model parameters: Q

$$Q=\{q_{ij}\}=\begin{pmatrix} - & q_{12} & \cdots & q_{120} \\ q_{21} & - & \cdots & q_{220} \\ & & \ddots & \\ q_{200} & q_{202} & \cdots & - \end{pmatrix},$$

- Transition probability matrix can be derived from Q :

$$P(t)=\{p_{ij}(t)\}=\exp\{Qt\}=U\exp\{\Lambda t\}U^{-1}$$

U : right eigenvectors,
U^{-1} : left eigenvectors,
Λ : diagonal matrix.

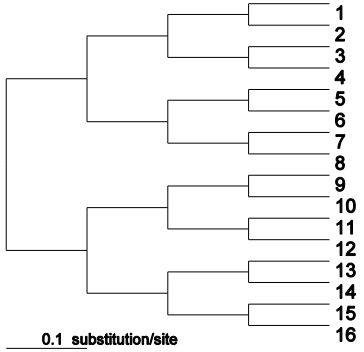
(Felsenstein, 1983; Yang 1994; Whelan and Goldman, 2000; Tseng and Liang, 2004)

Likelihood function of a given phylogeny

- Given a set of multiple-aligned sequences $S = (x_1, x_2, \ldots, x_s)$ and a phylogenetic tree $T = (V, E)$,

A column x_h at poisition h is represented as:

$$x_h = (x_{1,h}, x_{2,h}, \ldots, x_{s,h})$$



- The Likelihood function of observing these sequences is:

One column

$$P(x_h \mid T, Q) = \pi_{x_k} \sum_{i \in A} \prod_{(i,j) \in E} P_{x_i x_j}(t_{ij})$$

Whole sequence

$$P(S \mid T, Q) = P(x_1, \ldots, x_s \mid T, Q) = \prod_{h=1}^{s} P(x_h \mid T, Q)$$

Tseng and Liang, 2006

3. Markov chain Monte Carlo for parameter estimation

Bayesian Model

• Posterior probability distribution of rate matrix given the sequences and tree:

$$\pi(Q|S,T) \propto \int P(S|T,Q) \cdot \pi(Q) dQ$$

where

$\pi(Q)$: prior distribution,

$P(S|T,Q)$: likelihood distribution

$\pi(Q|S,T)$: posterior distribution.

• Bayesian estimation of posterior mean of rates in Q :

$$\mathbf{E}_{\pi}(Q) = \int Q \cdot \pi(Q \mid S, T)\, dQ,$$

• Estimated by Markov chain Monte Carlo.

Markov chain Monte Carlo method for parameter estimation

- Target distribution π :
  - posterior probability function
- Can evaluate this function π,
  - but direct sampling from it is impossible!
- Generate (correlated) samples from the target distribution π
  - Run a Markov chain with π as its stationary distribution

Markov chain Monte Carlo

- Proposal function:

Yan Yuan Tseng and Jie Liang, *Mol Biol Evo. 2006*

- Detailed balance: samples target distribution after convergency.

- Metropolis-Hastings Algorithm:

- Collect data from *m* acceptant samples

$$\mathbf{E}_{\pi}(Q) \approx \sum_{i=1}^{m} Q_i / m \approx \int Q \, \pi(Q \mid S, T) \, dQ.$$

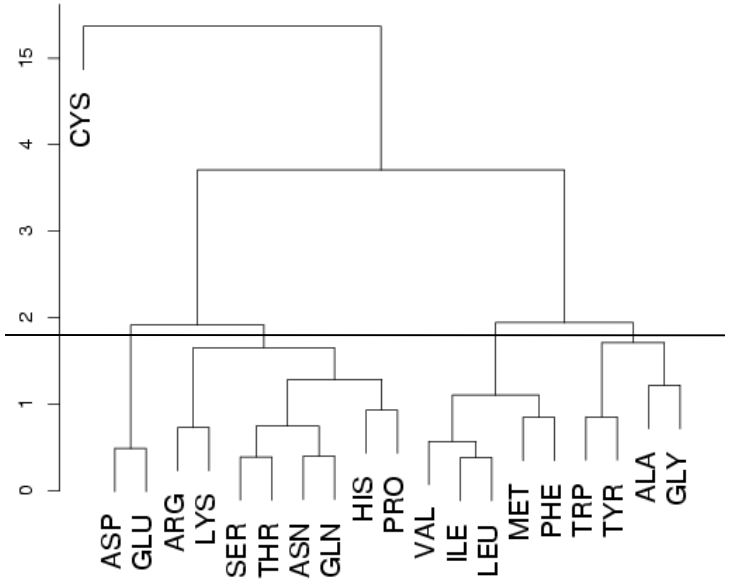
Move Set

- **Two types of moves : s_1, s_2**
- **Individual moves : s_1**

$$q_{jt+1} = \alpha_1 q_{jt+1},$$
$$q_{jt+1} = \alpha_2 q_{jt+1},$$
$$\text{where } \alpha_1 \neq 0, \alpha_2 = 1.1$$

- **Block moves: s_2**

Select different dihedral angles
G, A, P, S, V, I,
S, C, M, Q, L, D, K, R, Y
Achieve with biased sampling
where $\alpha_1 \neq 0, \alpha_2 = 1.1$



- **Transition matrix between two types of moves:**

$$\begin{pmatrix} s_1 \to s_1 & s_1 \to s_2 \\ s_2 \to s_1 & s_2 \to s_2 \end{pmatrix} = \begin{pmatrix} 0.9 & 0.1 \\ 0.9 & 0.1 \end{pmatrix}$$

Acceptance ratio:

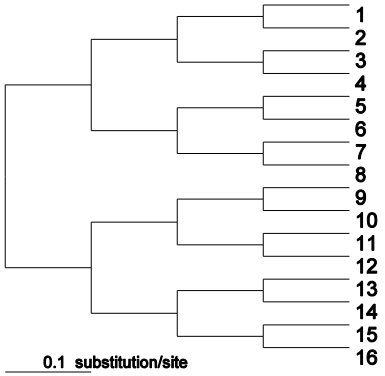
Individual moves : 50%-66%

Block moves: <10%

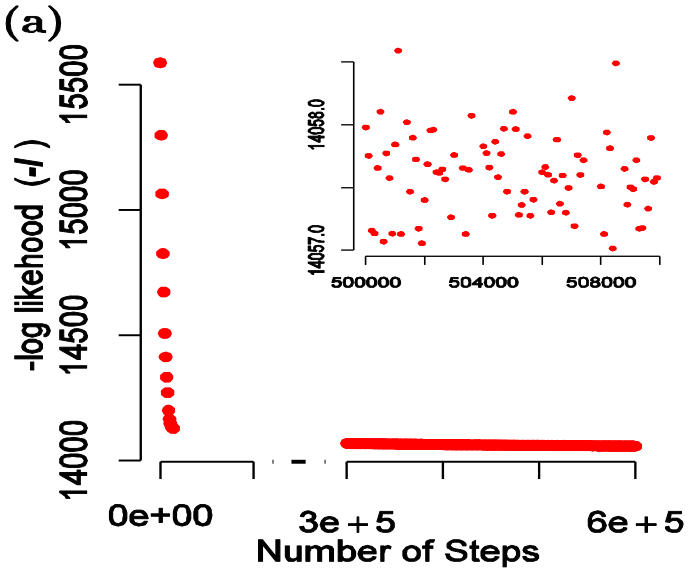
Yan Yuan Tseng and Jie Liang, *Mol Biol Evo. 2006,* 23:421-436.

Validation by simulation

- Generate 16 artificial sequences from a known tree and known rates (JTT model)
  - Carboxypeptidase A2 precursor as ancestor, length = 147
- Goal: recovering the substitution rates



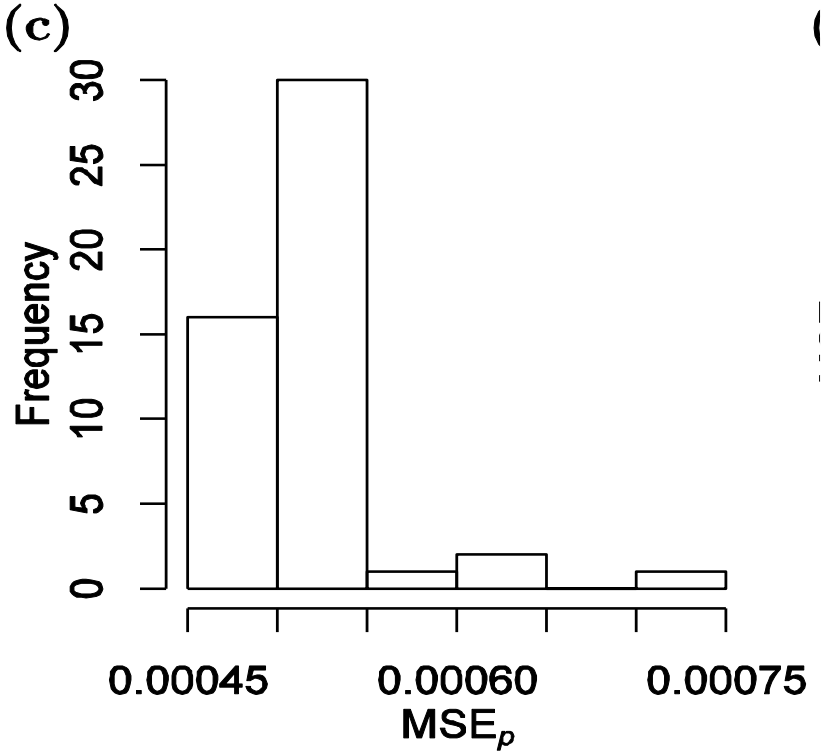
Phylogenetic tree used to generate 16 sequences

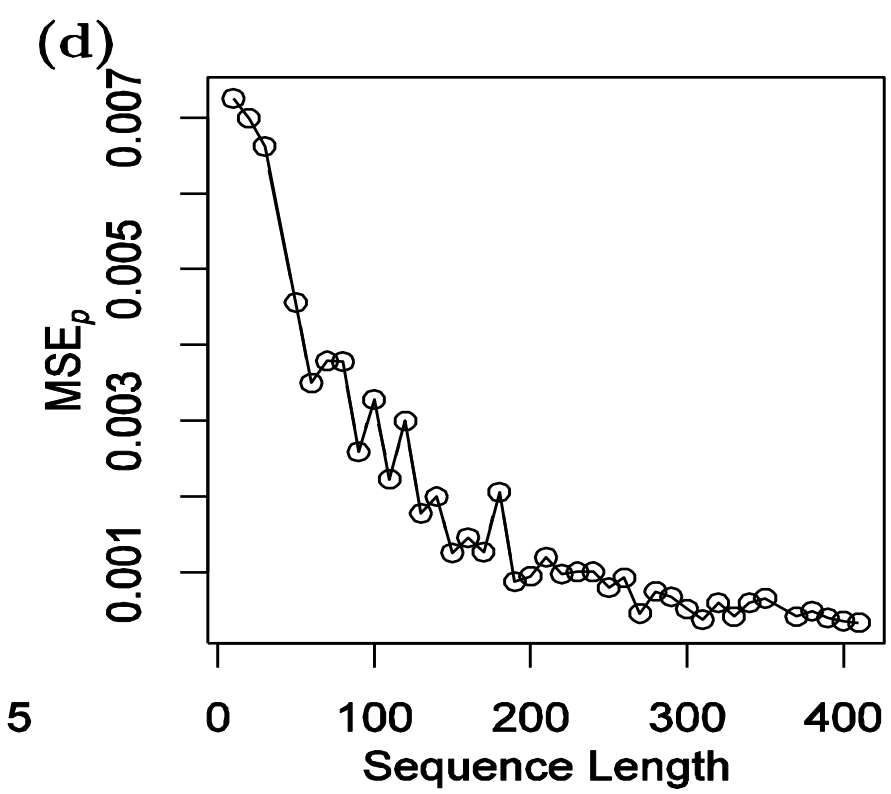


Convergence of the Markov chain

Yan Yuan Tseng and Jie Liang, *Mol Biol Evo. 2006,* 23:421-436.

Accurate Estimation with > 20 residues and random initial values





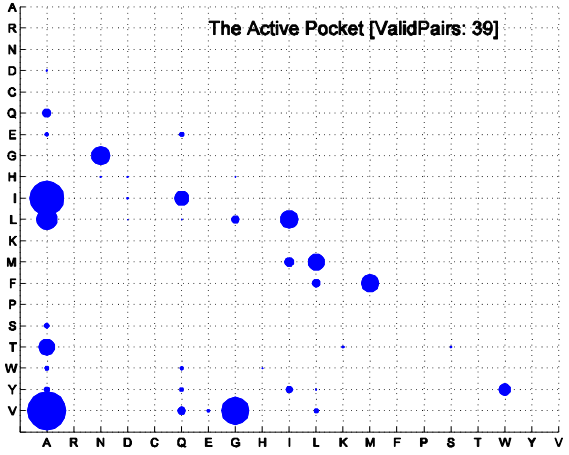
Distribution of MSE of estimated rates starting from 50 sets of random initial values.

All MSE < 0.00075.

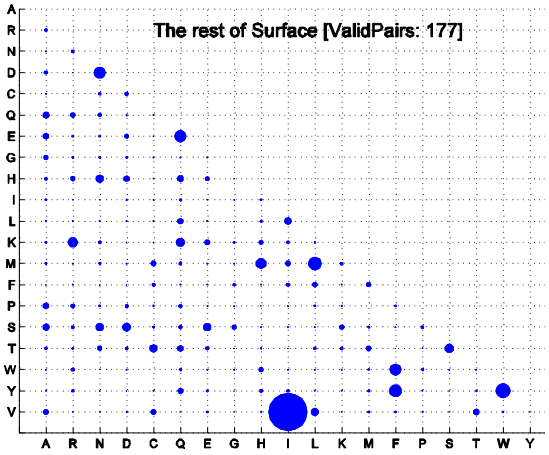
Accurate when > 20 residues in length.

Yan Yuan Tseng and Jie Liang, *Mol Biol Evo. 2006,* 23:421-436.

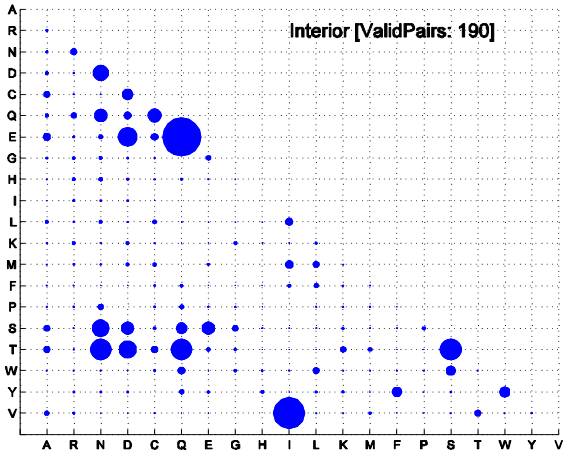
Evolutionary rates of binding sites and other regions are different



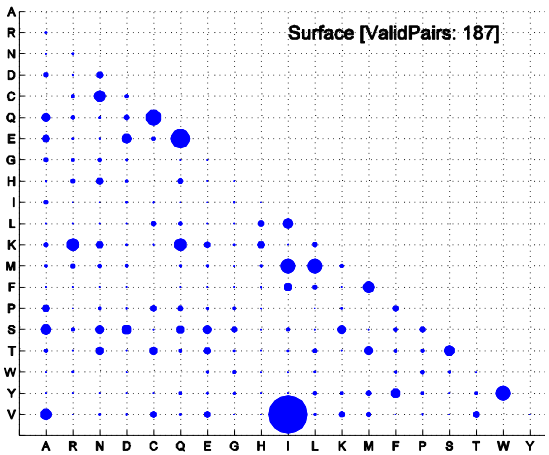
(a)



(b)



(c)



(d)

Residues on protein functional surface experience different selection pressure.

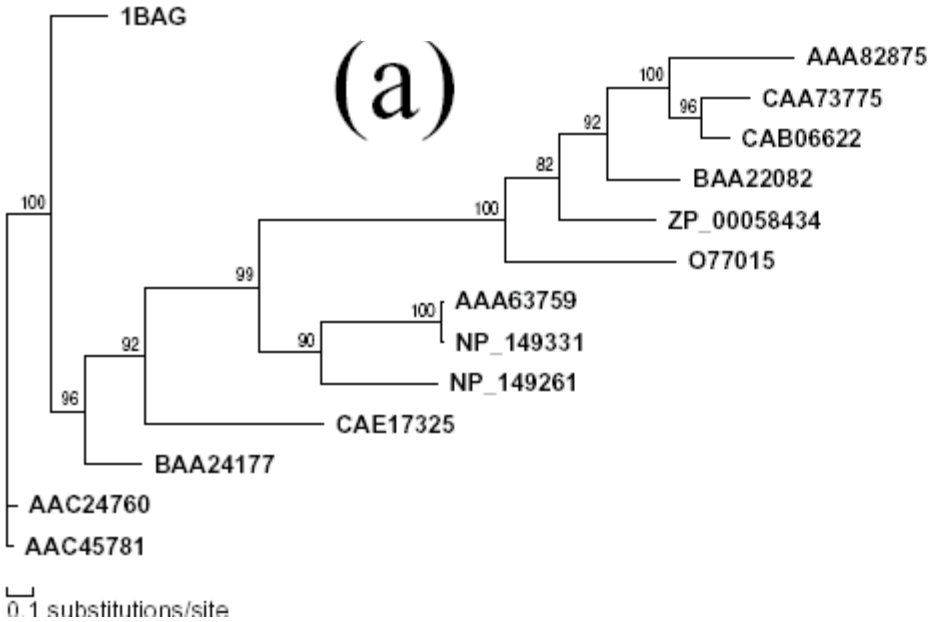
Estimated substitution rate matrices of amylase:

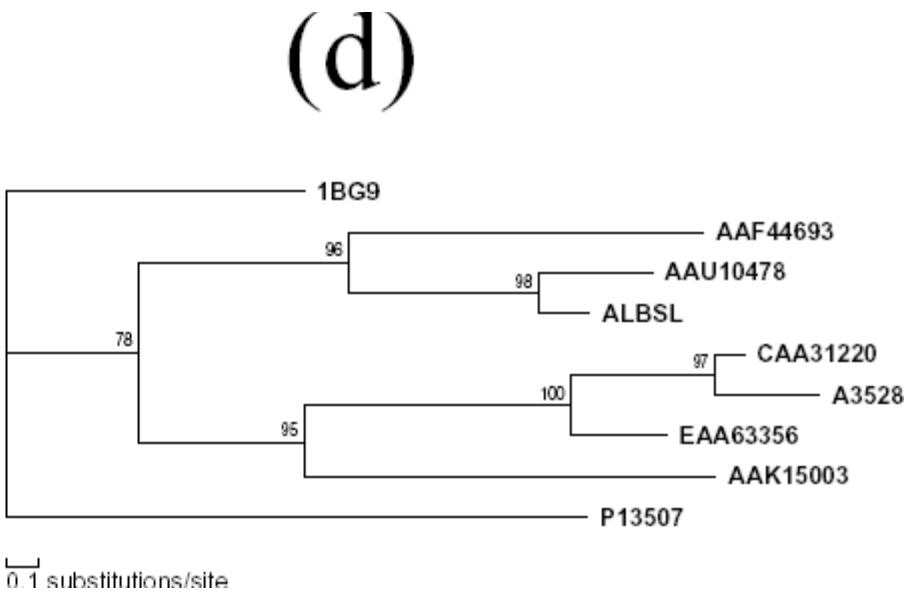
- Functional surface residues.
- The remaining surface,
- The interior residues
- All surface residues.

Example 1: Finding alpha amylase by matching pocket surfaces

Challenging:

– amylases often have low overall sequence identity (<25%).





–1bag, pocket 60; *B. subtilis*

–14 sequences, none with structures, 2 are hypothetical

–1bg9; *Barley*

–9 sequences, none with structures.

Criteria for declaring similar functional surface to a matched surface

- Search >2million surfaces with a template surface.
- Shapes have to be very similar:
  - *p*-value for cRMSD: $< 10^{-3}$.
- Customized scoring matrices of 300 different time intervals.
  - The most similar surface has n_{max} of matrices capable of finding this homologous surface.
  - Declare a hit if >1/3 n_{max} of matrices give positive results.

Yan Yuan Tseng and Jie Liang, *Mol Biol Evo. 2006,* 23:421-436.

4. Application in protein function prediction

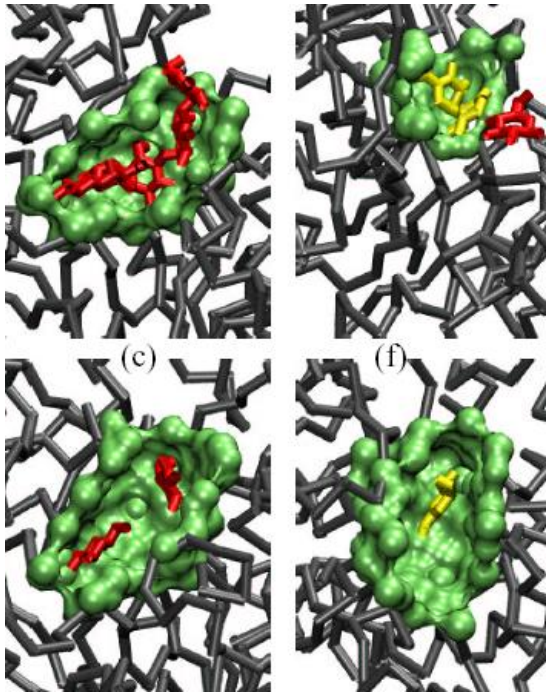
Results for Amylase

- `1bag`: found 58 PDB structures.
- `1bg9`: found 48 PDB structures.

- Altogether: 69
  - All belong to amylase (EC 3.2.1.1)

Query: *B. subtilis* Barley

1bag 1bg9

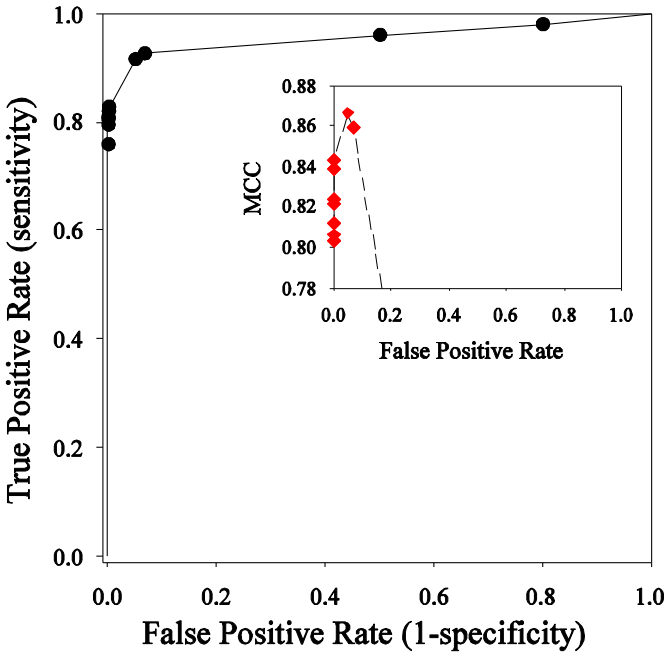


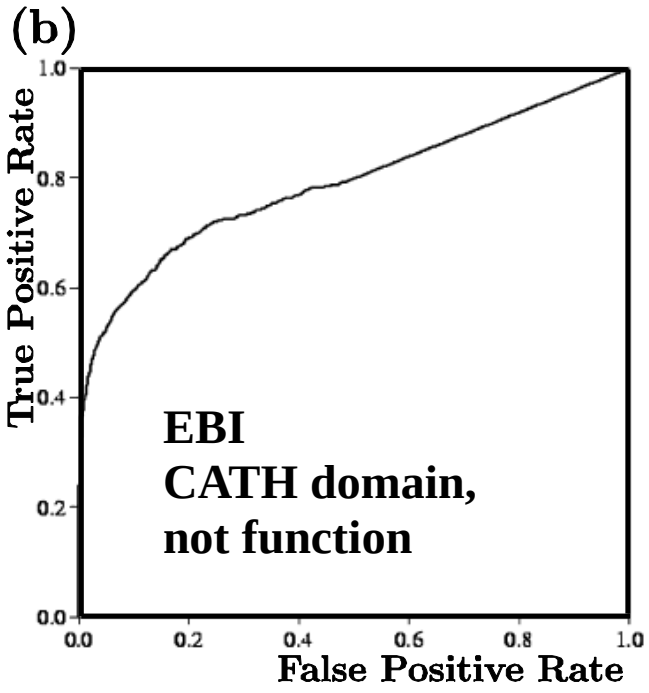
Hits: human

1b2y 1u2y

22% 23%

Large Scale Prediction of Protein Functions





Laskowski, Thornton, JMB, 05, 351:614-26

False positive rate ≡

$$1 - \frac{TN}{TN + FP} = \frac{FP}{TN + FP}$$

True positive rate ≡

$$\frac{TP}{TP + FN}$$

- 110 protein families
- Each points on the curve corresponds to *p*-values of various cRMSD cutoffs
- Accuracy ~92% (EBI: 75%)

Helmer-Citterich, M et al (BMC Bioinformat. 2005)
Russell RB. (JMB 2003)
Sternberg MJ
Skolnick, J
Lichtarg, O (JMB2003)
Ben-Tal, N and Pupko, T (ConSurf)

(Tseng, Dundas, and JL, J Mol Biol, 2009, **387(2)** ***451-464.*** ; Liang et al, Adv Protein Chem, 2008)

Orphan protein structures without functional annotation

Orphan proteins from Structural Genomics.

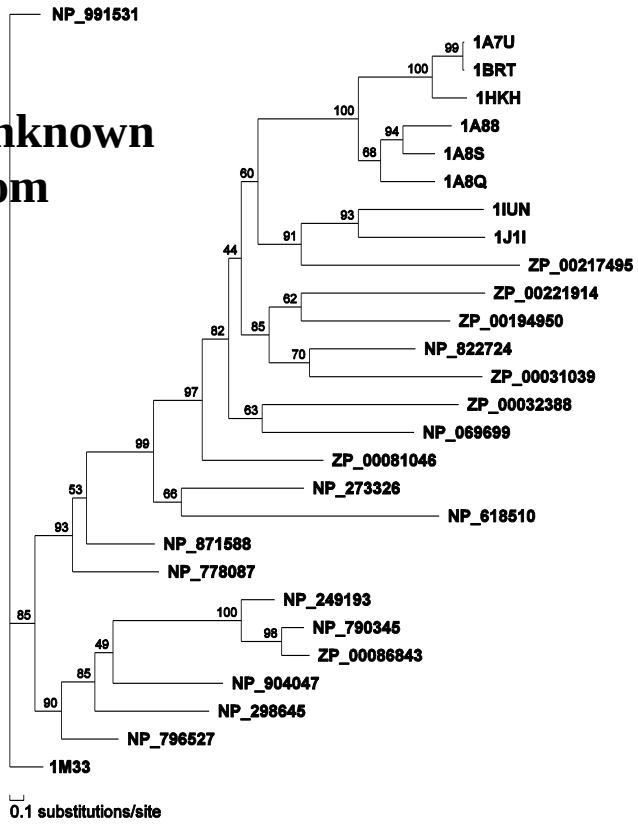
- No known functions.
- Often sequences homologs are hypothetical proteins.

Our tasks:

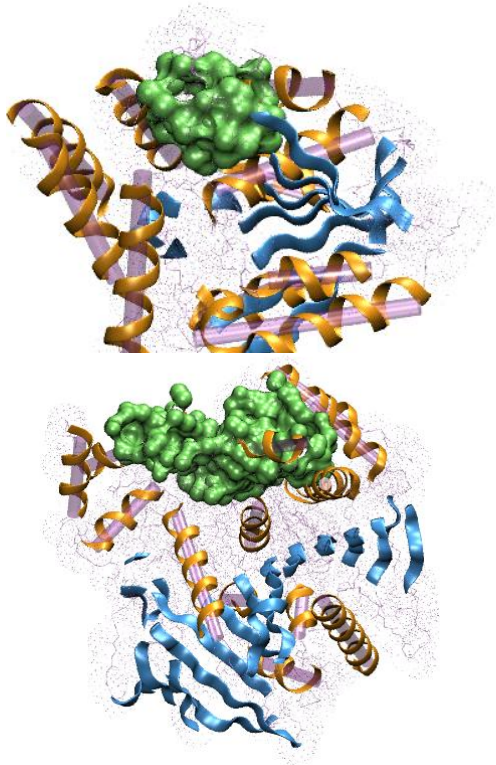
- Identify the functional pocket for the structure.
- Predict protein function.

Inferring biological functions of protein BioH

Protein of unknown functions from structural genomics



The phylogenetic tree of 28 sequences related to BioH. Many are hypothetical genes.



The candidate binding pocket (CASTp id=35) of BioH (1m33) and a similar functional surface detected from (1p0p, E.C. 3.1.1.8)

A Probablistic Model of Enzyme Function

- Enzymes have diverse reactivites:
  - Some are very specific
  - Some can react with a range of substrates
  - An E.C. number alone is inadequate

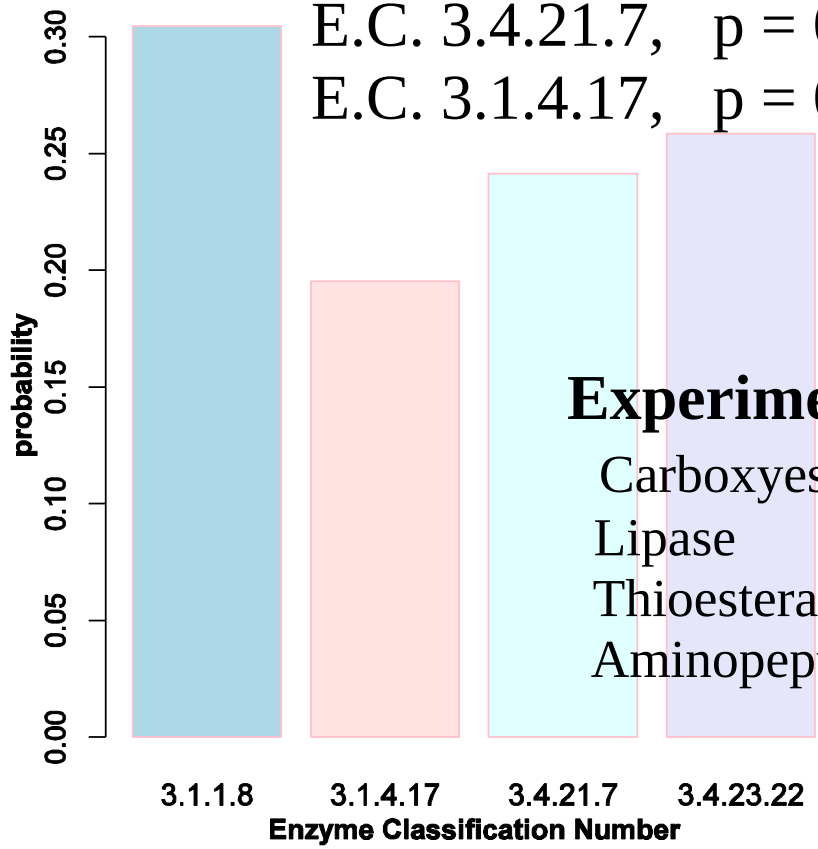
- Our approach: a probabilistic model

(Tseng, Dundas, and JL, J Mol Biol, 2009, **387(2)** ***451-464.*** ; Liang et al, Adv Protein Chem, 2008)

BioH

Significant hits predicted for BioH:

E.C. 3.1.1.8, p = 0.31, cholinesterase
E.C. 3.4.23.22, p = 0.26, aspartic endopeptidase
E.C. 3.4.21.7, p = 0.24, serine endopeptidase
E.C. 3.1.4.17, p = 0.20, phosphodiesterase

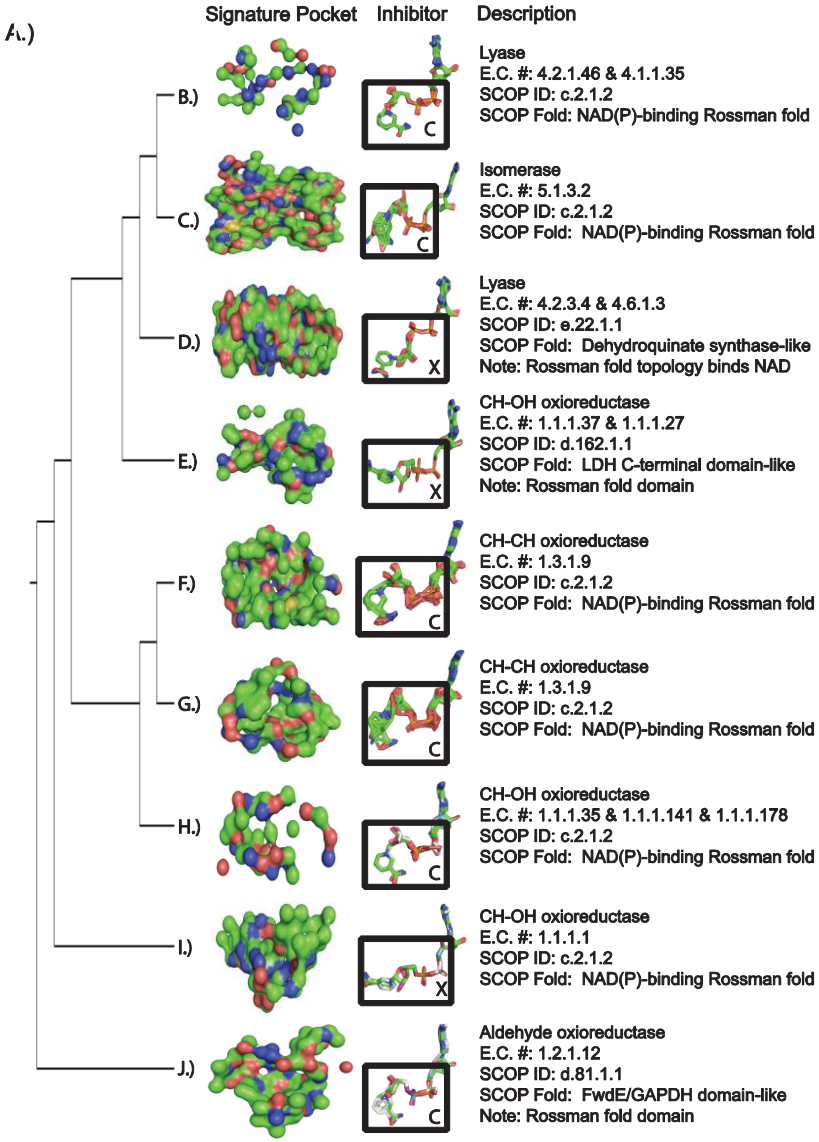


Experimental Results:

Carboxyesterase	E.C. 3.1.1.1	high
Lipase	E.C. 3.1.1.3	low
Thioesterase	E.C. 3.1.1.5	low
Aminopeptidease	E.C. 3.4.11.5	low

(Tseng, Dunda, Liang, 2009, JMB)

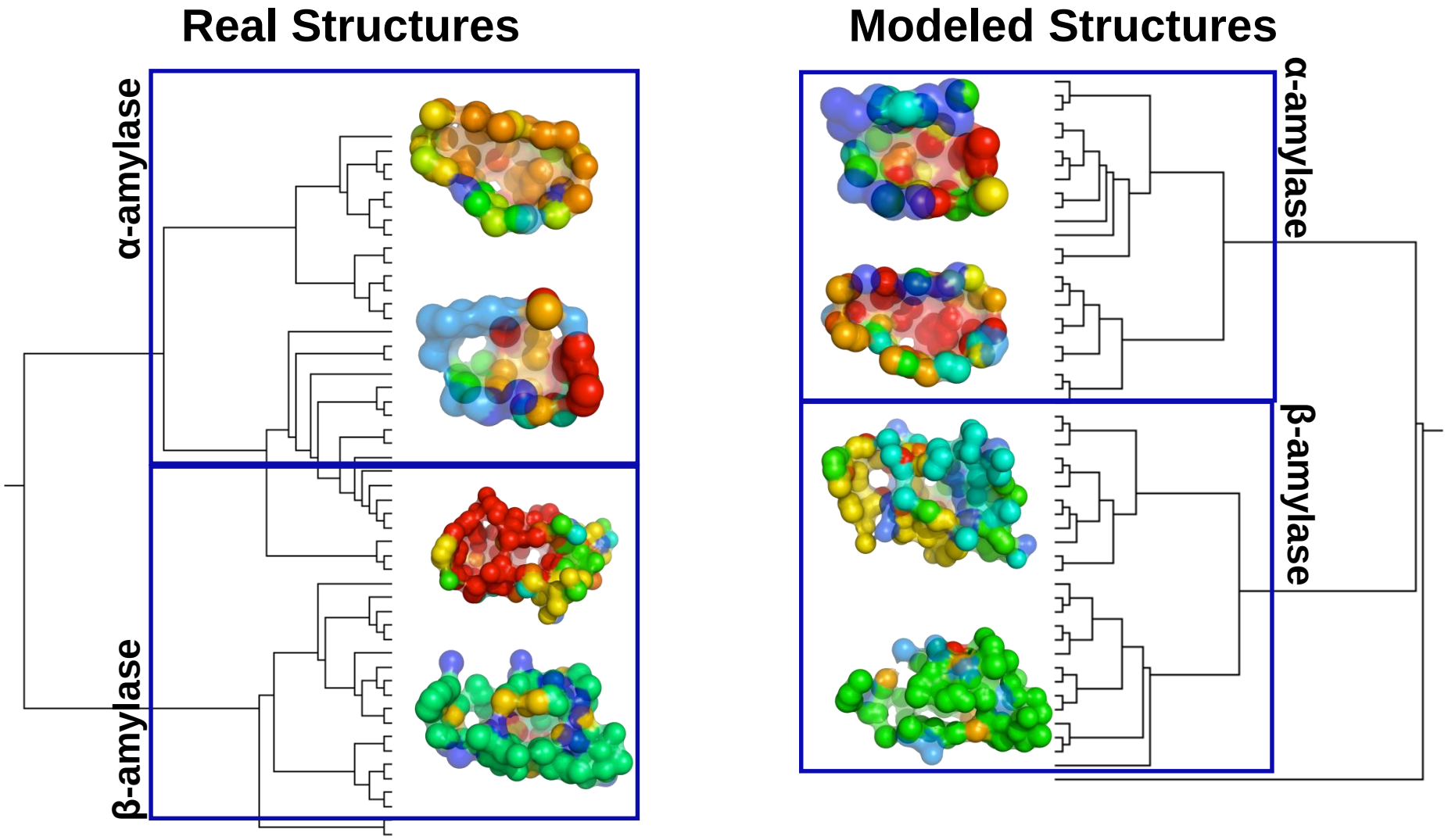
Signature and Basis Set of Binding Surfaces: NAD Sites



- Canonical spatial surface conformations for binding NAD
  - Explains all known NAD binding sites
- Reveal structural basis of binding
- Can predict NAD binding surfaces

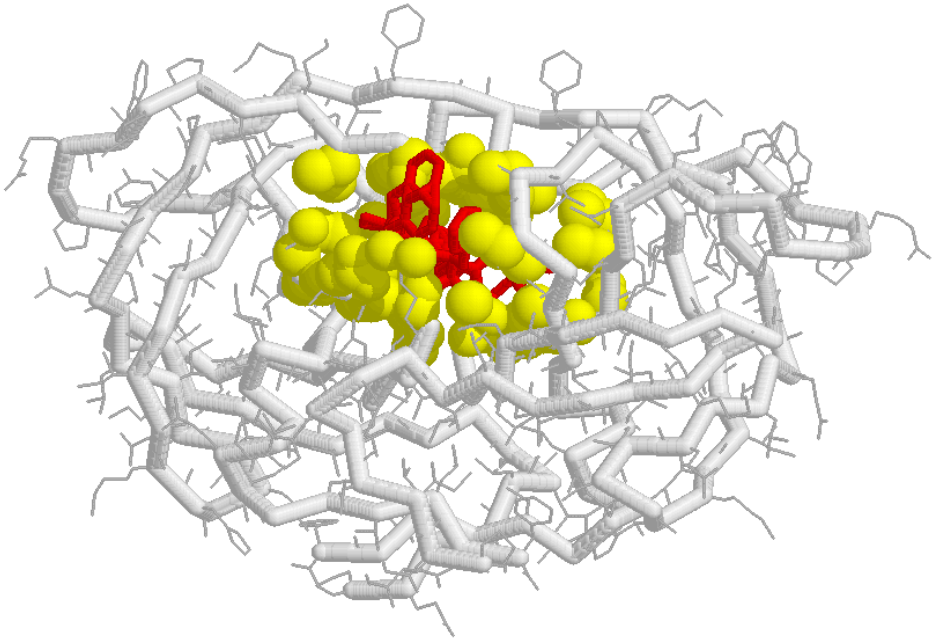
(Dundas, Adamian, and Liang, *J Mol Biol*, 2011, 406(5):713-29)

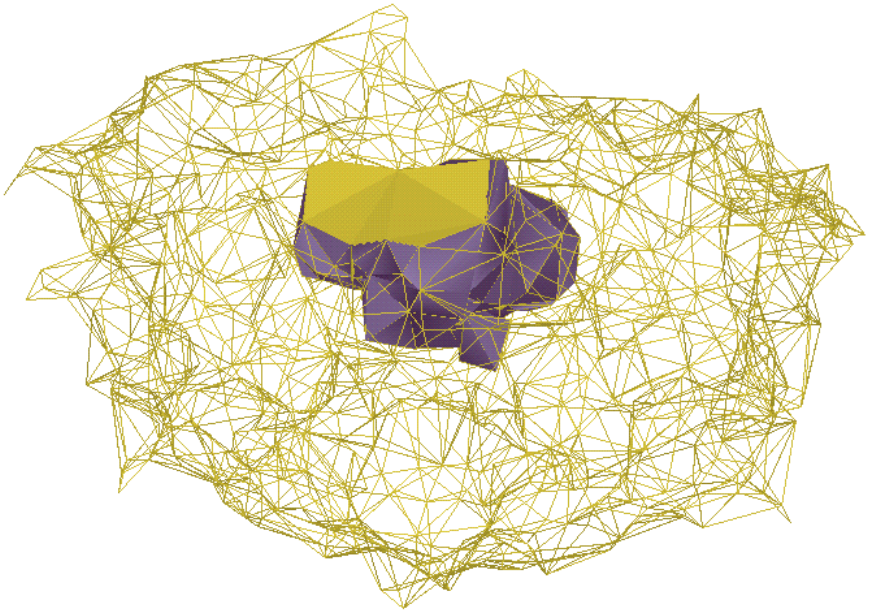
Signatures from Real and Modeled Structures



(Zhao, Dundas, Kachalo, Ouyang, and Liang, *J Struct Funct Genomics*, 2011, 12(2):97-107)

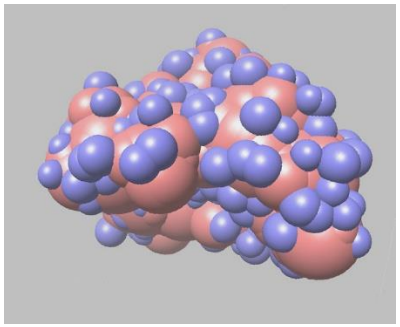
Geometry of Binding Sites

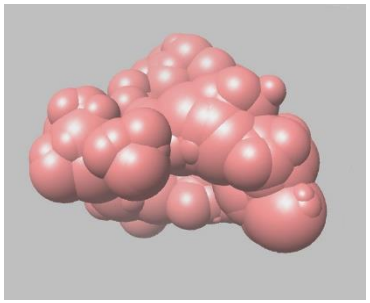




Structure Based Drug Discovery

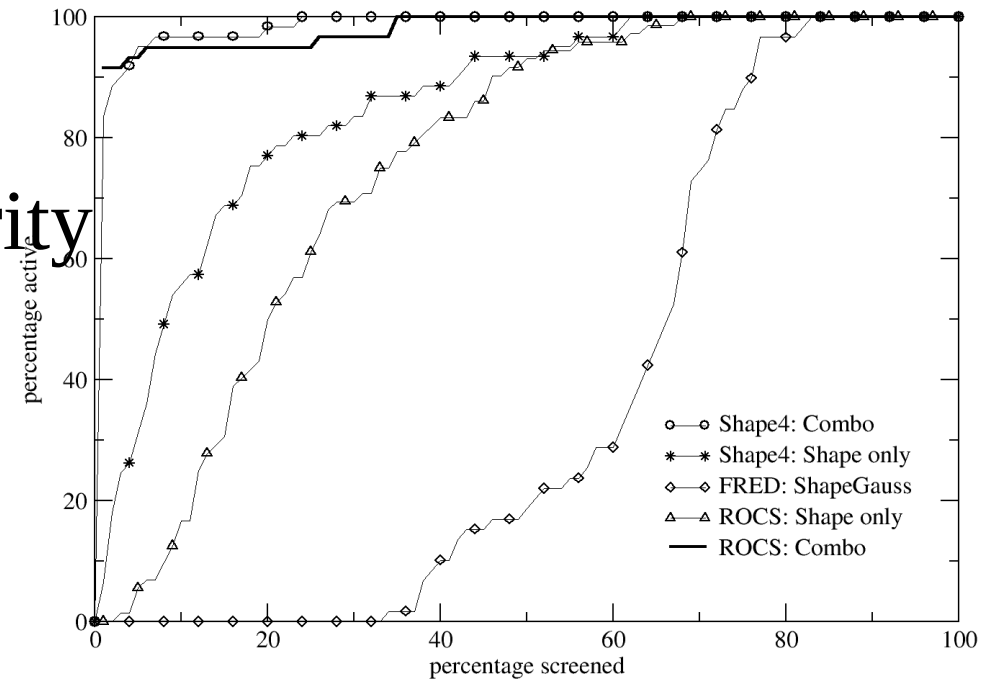
- Binding pocket based cast of negative imprint:
  - As template for compound search





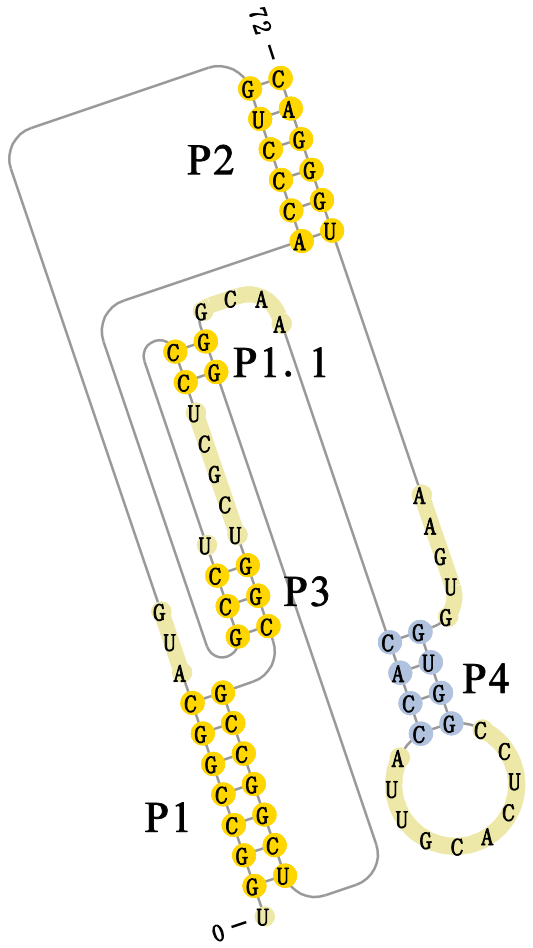
(Zheng Ouyang)

- Shown to be effective in compound similarity search.
  - Several biological activity classes.

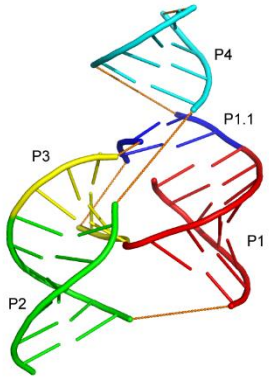


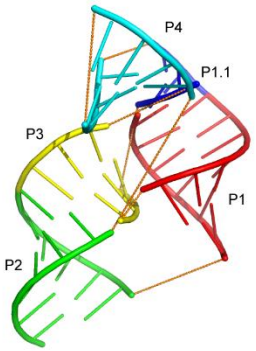
(Ebalunode, Ouyang, Liang, and Weifan Zheng, *J Chemical Information and Modeling*, 48(4):889-901,2008)

Other work: Prediction of pseudo-knotted RNA



- Assembly of candidate stable stems
- Sampling for entropy of models of complicated 3D loops





(Zhang, Lin, Chen, Wang, and Liang, RNA, 2009, *15: 2248-2263*)

Summary

- Space filling structures of proteins:
  - volume and surface models,
- Geometric constructs and algorithms:
  - Voronoi diagram, Delaunay triangulation, and alpha shape
- Application in proteins packing and function prediction

Collaborators

- **Andrew Binkowski** (Argonne National Lab)
- **Jeffrey Yan-Yuan Tseng** (Wayne State U)
- **Youfang Cao (**UIC)
- **Dennis Gessmann (UIC)**
- **Gamze Gursoy** (UIC)
- Sema Kachalo (now Intel graphics unit)
- Ronald Jackups, Jr (Wash U, Assist Prof)
- Yingzi Li (SJTU)
- **Meishan Lin** (UIC)
- **Hammad Naveed** (UIC/now Toyota Institute)
- **Ke Tang** (UIC)
- **Anna Terebus** (UIC)
- Wei Tian (UIC)
- **Yun Xu** (UIC)
- **Jieling Zhao** (UIC)
- Jinfeng Zhang (FSU, Associate Prof)

- David Stone (UIC)
- **Luisa Di Pietro (UIC)**
- **Linda Kenney (UIC)**
- **Amy Kenter (UIC)**
- John Marko (NWU)
- Lisa Xu (SJTU)

Papers: http://gila.bioe.uic.edu/lab/ (left column)

Acknowledgement

- NSF DBI
- NIH NIGMS