Inferring protein functions by matching binding surfaces through evolutionary models

Jie Liang
(Joint work with Jeffrey Tseng)

Dept. of Bioengineering
University of Illinois at Chicago

Outline

Methodology:

- Computational geometry of surface pattern:
  - Candidate motifs.
- Assessing surface similarity.
  - Sequence, shape, orientation, and *p*-values.
- Incorporation of evolutionary information by Bayesian Markov chain Monte Carlo.

Discovery:

- Protein functional prediction.

The Universe of Protein Structures

- Human genome: 3 billion nucleotides
- Number of genes: 30,000
- Protein families: 10,000-30,000
- Number of folds: 1,000 - 4,000

All β

α/β

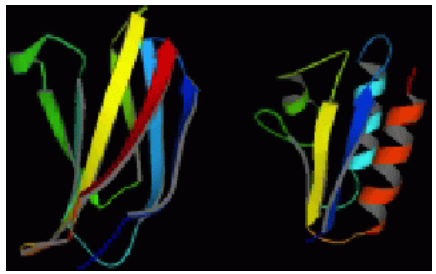
(from SCOP)

- Currently in PDB: < 700 folds
  - Comparative modeling: needs a structural template with sequence identities > 30-35%
    - eg. ~50% of ORFs and ~18% of residues of *S. cerevisiae* genome
- 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.

- Main chain folds:
  - Important for understanding evolution.
  - May not directly lead to understanding of function

Tenasin
1ten

Phosphotransferase
1poh



(SCOP)

All beta proteins
Ig like beta sandwich

a+b proteins
HPr fold

Tenasin
1ten

Phosphotransferase
1poh

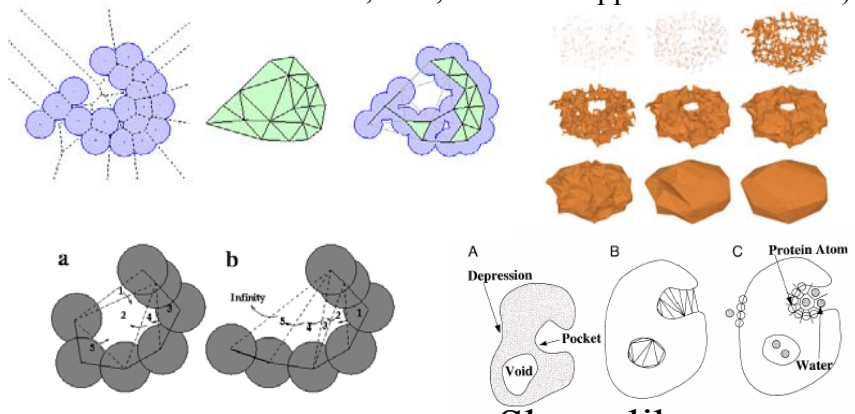
(from Jaroszewski & Godzik, ISMB 00)

Predicting protein function by matching surfaces

- Proteins from structural genomics often are of unknown functions.
  - Sequence homologs are often hypothetical proteins.
- Strategy: Matching automatically computed surfaces that may be binding sites.
- Three tasks:
  - Geometric computation: A library of >2 million surface patterns on > 20,000l PDBs. (cast.engr.uic.edu)
  - Similarity measure: Sequence patterns, coordinate RMSD, and orientational RMSD.
  - Scoring matrix.

(Mucke and Edelsbrunner, ACM Trans. Graphics. 1994.

Edelsbrunner, et al, Discrete Applied Math. 1998.)

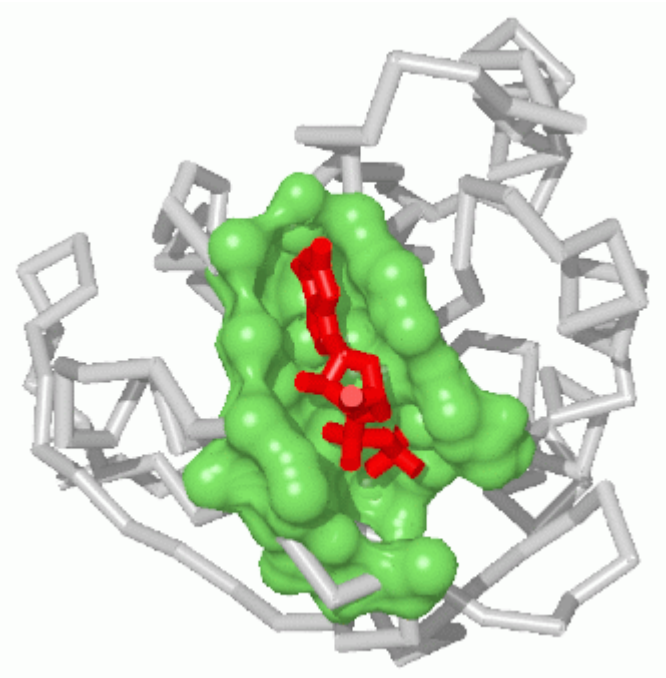


Shape library

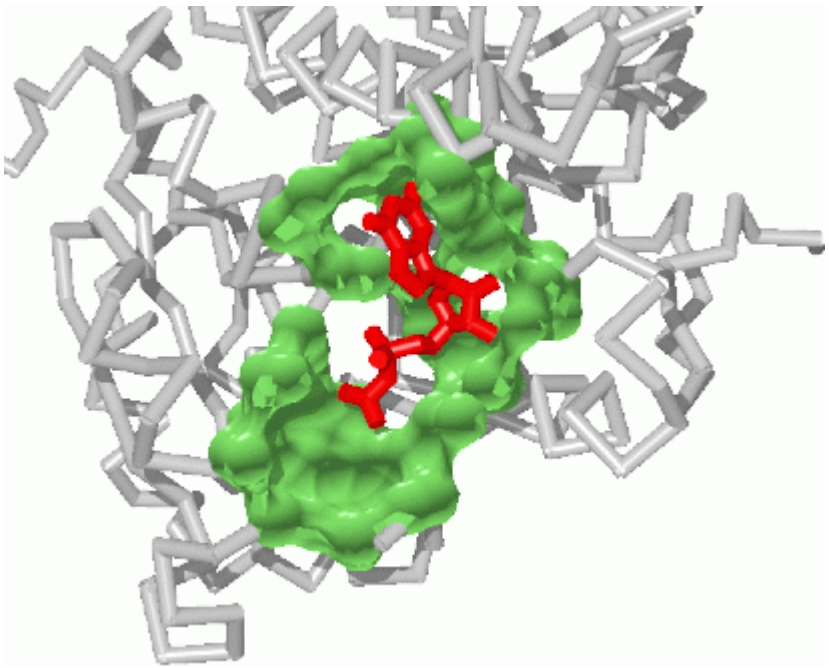
(Binkowski, Adamian, and Liang,
J. Mol. Biol. 332:505-526, 2003)

Protein Functional Surfaces

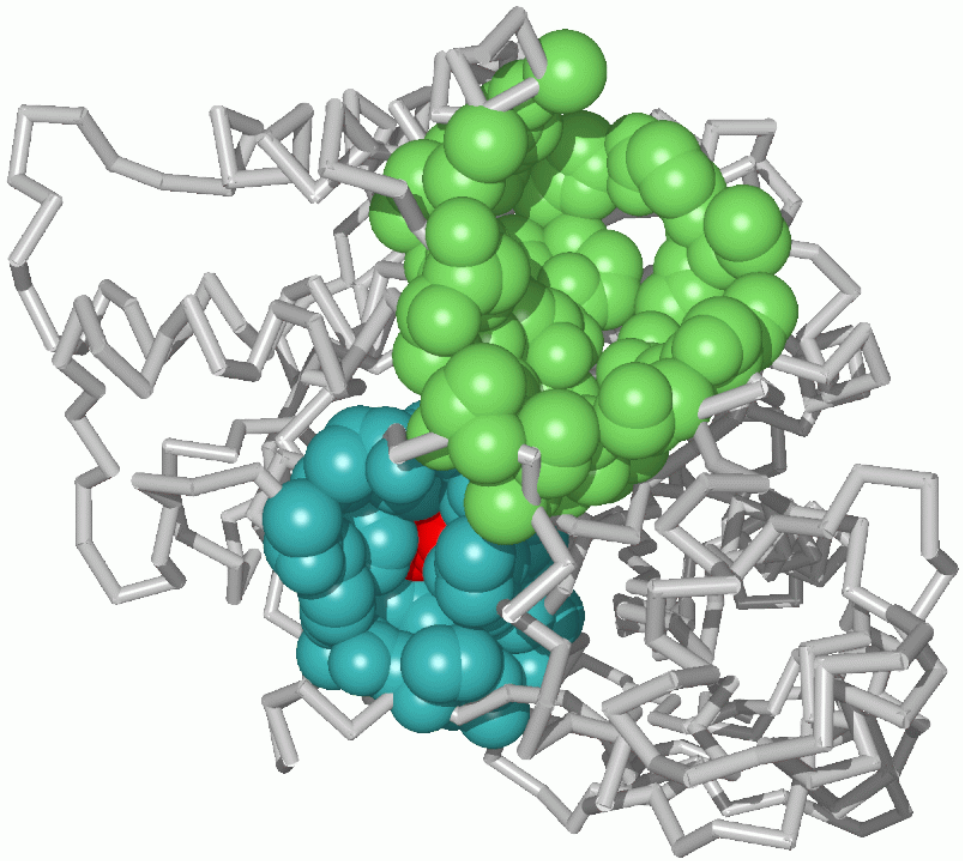
Ras 21



Fts Z

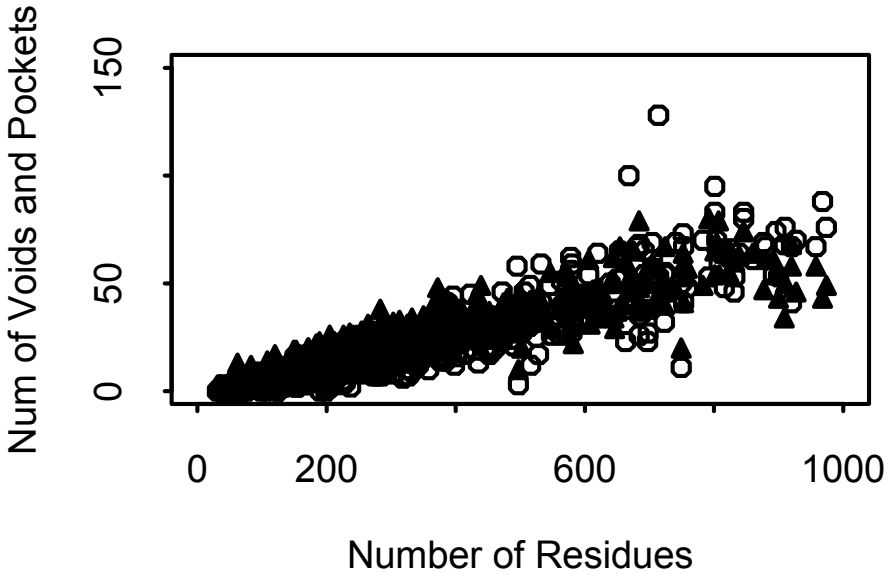


GDP Binding Pockets



http://cast.engr.uic.edu

Voids and Pockets in Soluble Proteins

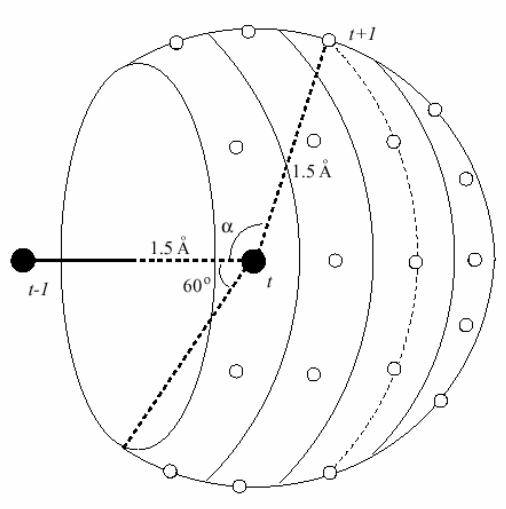


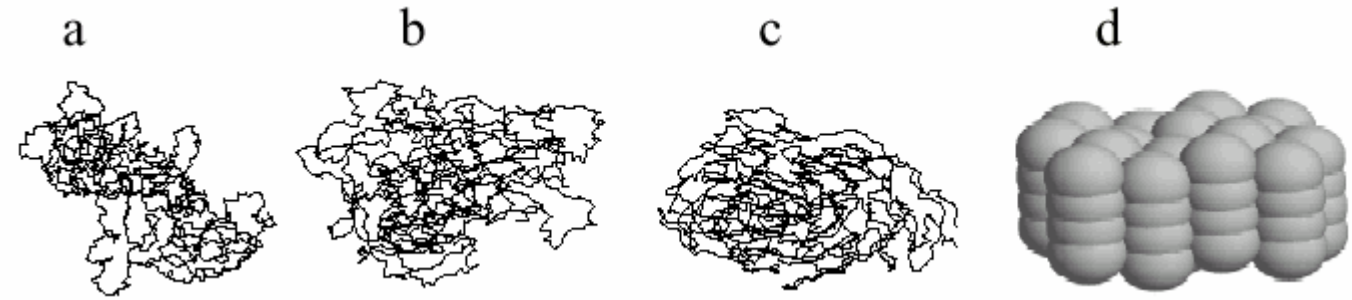
- Many voids and pockets.
  - At least 1 water molecule.
  - 15/100 residues.

(Liang & Dill, 2001, Bioph J)

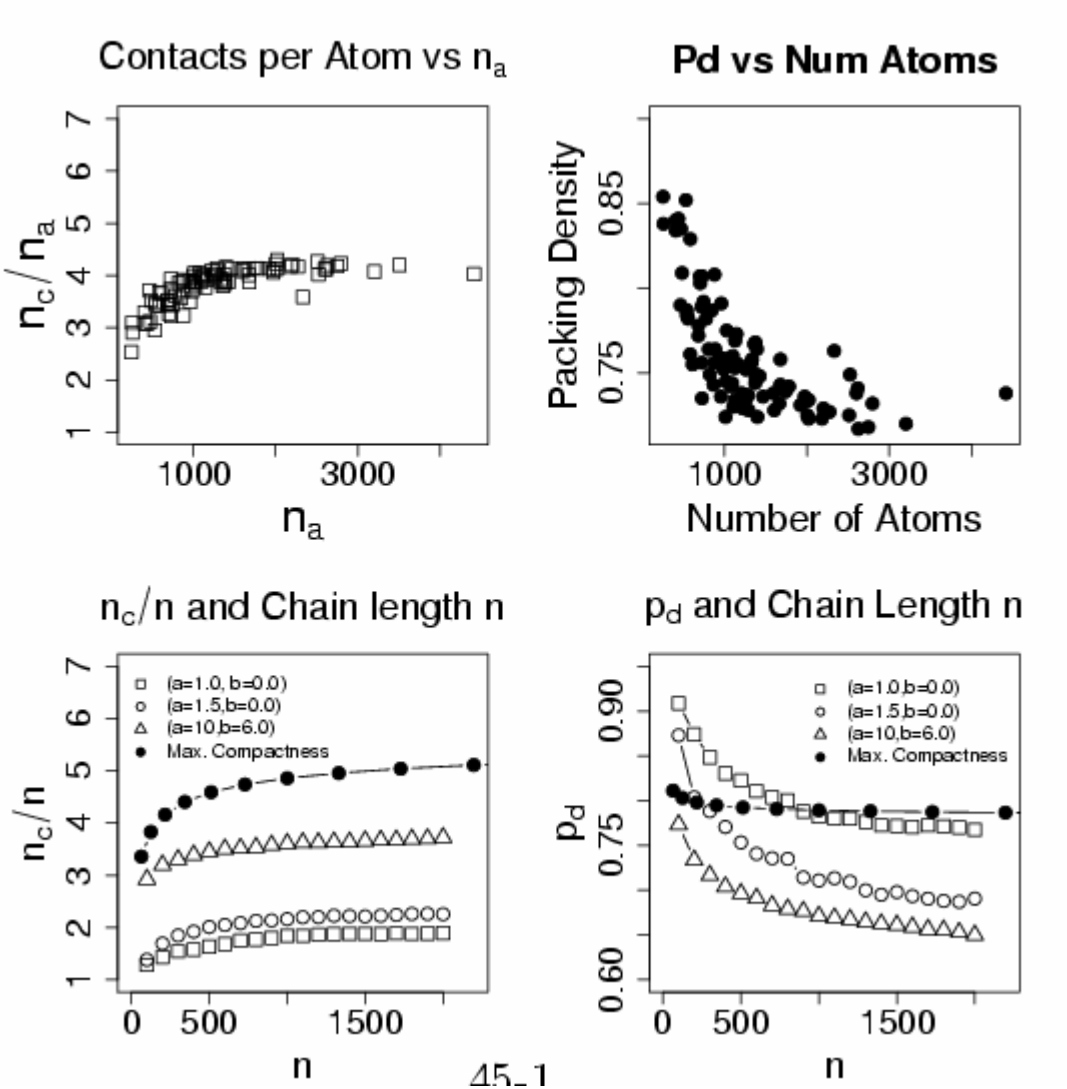
Simulating Protein Packing with Off-Lattice Chain Polymers

- 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*.)



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

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

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)
- Tyr Protein Kinase c-src (2src)

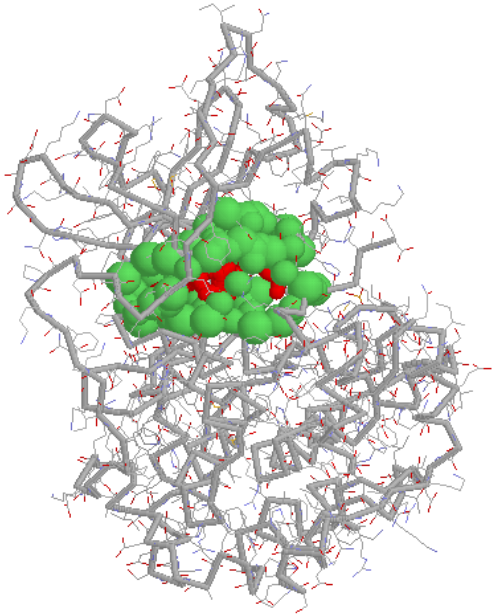
```
1cdk.A
   49LGTGSFGRVMLVKHKETGNHFAMKILDKQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEYSFKDNSNL
   YMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFG
   FAKRVKGRTWTLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFADQPIQIYEKIVSGKVR
   FPSHFSSDLKDLLRNLLQVDLTKRFGNLKDGVNDIKNHKWFATTDWIAIYQRKVEAPFIPKFKGPGDTSN
   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
```

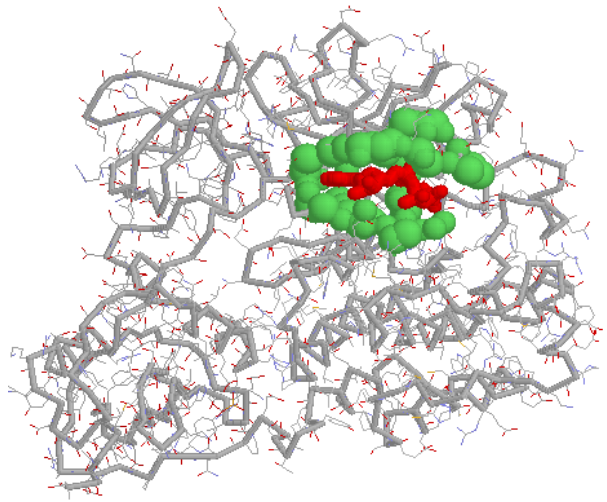
Low overall sequence identity: 13 %

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)
    - BLOSUM50 scoring matrix (Henikoff, 1994).
    - Not identity.
  - Order Dependent Sequence Pattern.

→ ***Statistical Significance !***

- Statistics of Null Model:
  - 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.

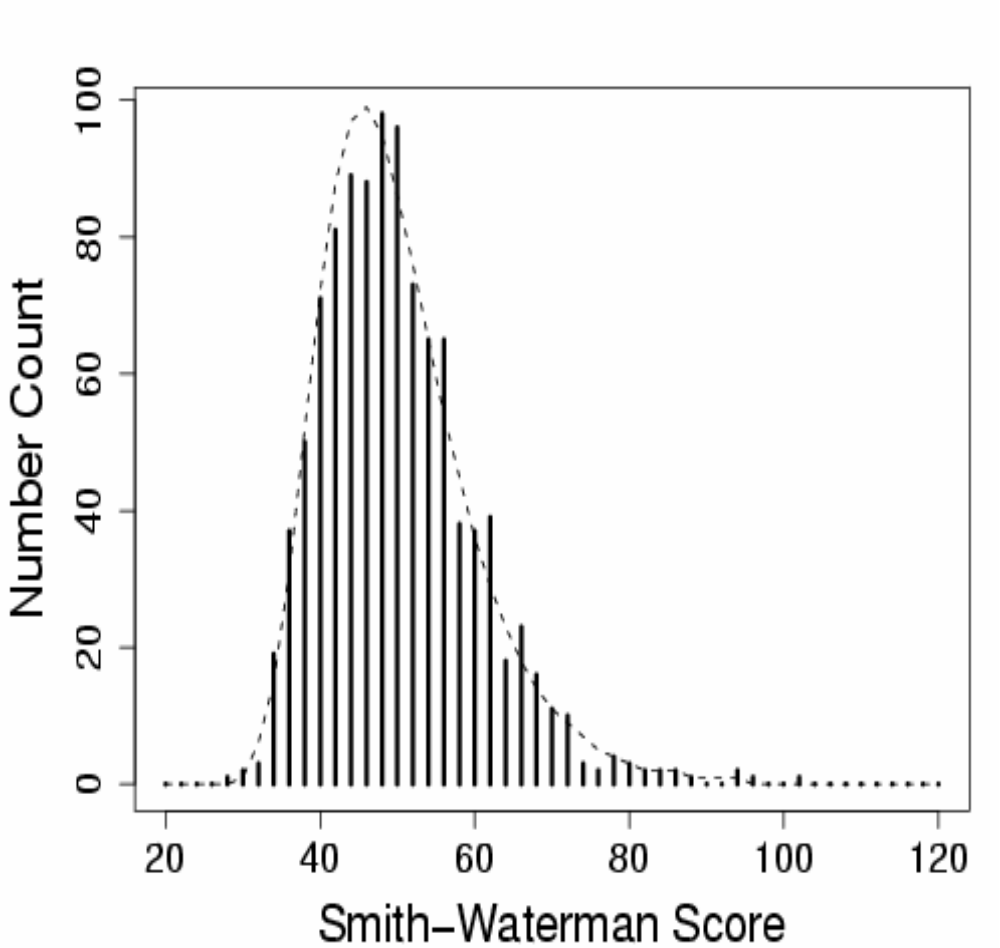
- Estimation of E-value:
  - Estimate p value of observed Smith-Waterman score by EVD.

$$S' = \lambda S - \ln Kmn,$$

$$p(S' \geq x) = 1 - \exp(-e^{-x})$$

  - Estimate E-value:

$$E = p \cdot (N_{\text{all}} - N_d) \leq p \cdot N_{\text{all}}$$



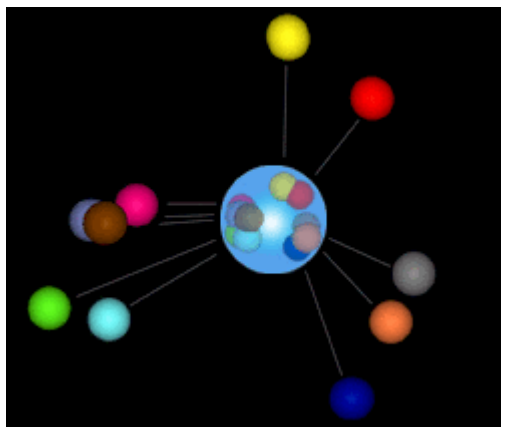
(Binkowski, Adamian, Liang, 2003, JMB, 332:505-526)

Shape Similarity Measure

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

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

  - Measuring RMSD between two sets of unit vectors.



(cf. uRMSD by Kedem and Chew, 2002)

Statistical Significance of Shape Similarity

- Estimate the probability p of obtaining a specific cRMSD or oRMSD value for random pockets with N_{res}
  - EVD and other parametric distributions not accurate.

  - Randomly select 2 pockets.
  - Calculate cRMSD for N_{res} randomly selected residues
  - Also calculate oRSMD

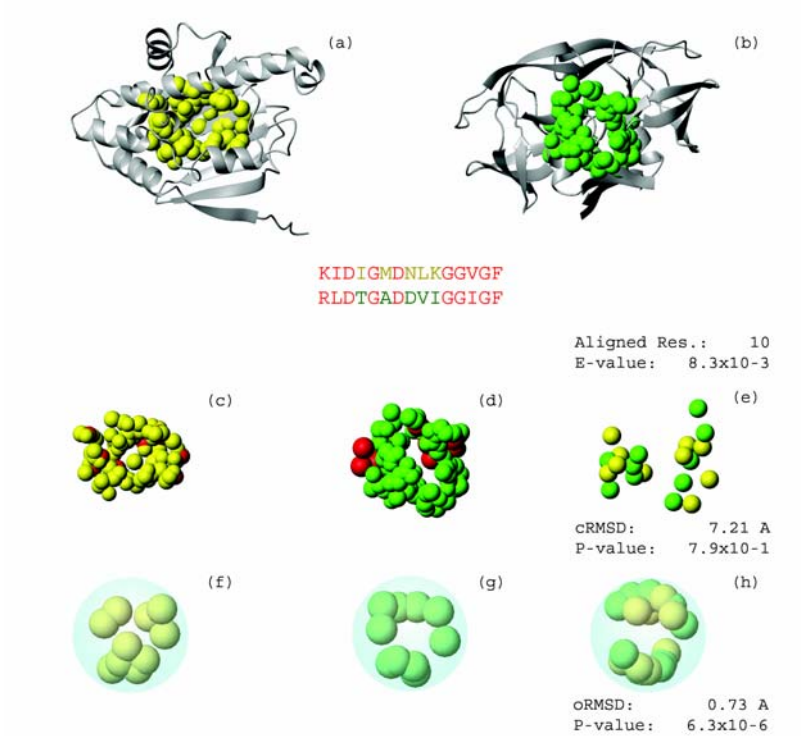
N_{res}	Random surfaces
3	10^{-8}
30	10^{-7}
100	10^{-6}

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

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	



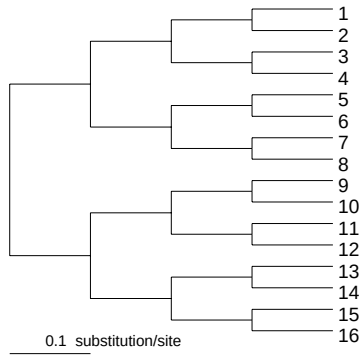
- Conserved residues both important in polypeptide binding
- Both pockets undergo conformational changes upon binding

How to incorporate evolutionary information?

What to do if related sequences all have unknown functions?

Likelihood function of a given phylogeny

- Given a set of multiple-aligned sequences $S = (\boldsymbol{x}_1, x_2, \cdots, \boldsymbol{x}_s)$ and a phylogenetic tree $T = (V, E)$,



A column x_h at poisition h is represented as:

$$x_h = (x_{1,h}, x_{2,h} \cdots, x_{s,h})^T$$

- The Likelihood function of observing these sequences is:

One column :

$$p(x_h \mid T, Q) = \pi_{x_k} \sum_{\substack{i \in I \\ x_{i \in A}}} \prod_{(i,j) \in \varepsilon} p_{x_i x_j}(t_{ij})$$

Whole sequence :

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

Estimation of instantaneous rates Q

- Posterior probability of rate matrix given the sequences and tree:

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

where

$\pi(Q)$: prior distribution,

$P(S \mid T,Q)$: likelihood distribution,

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

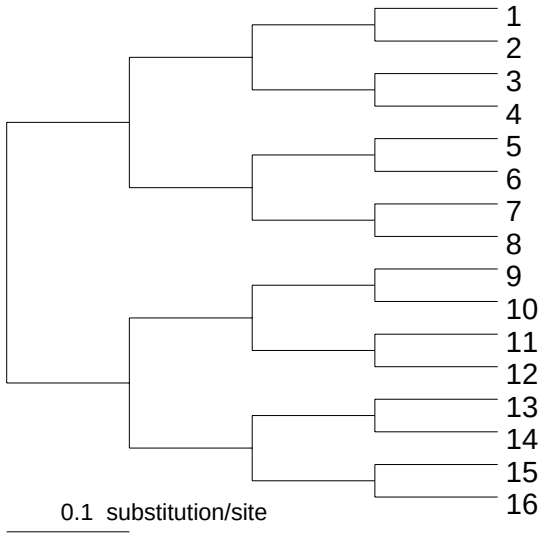
- Bayesian estimation of posterior mean of rates in Q :

$$\mathbb{E}_{\pi}(Q) = \int Q \cdot \pi\,(Q \mid S,\ T)\ \mathrm{d}\ Q,$$

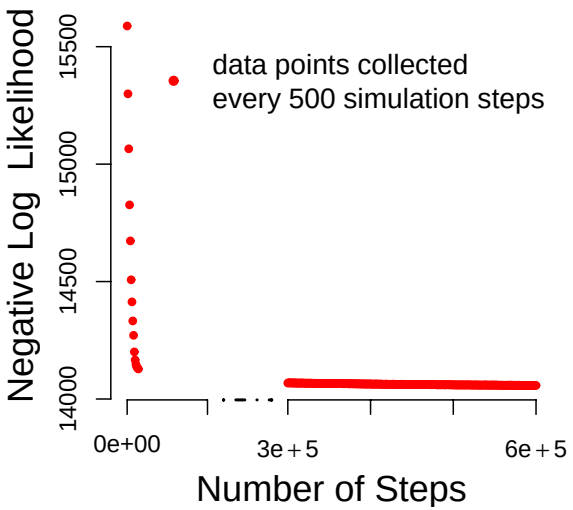
- Estimated by Markov chain Monte Carlo.

Validation by simulation

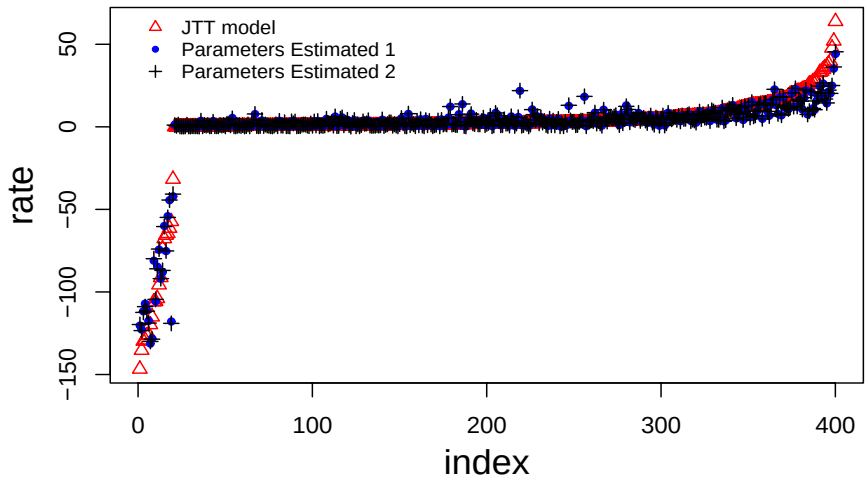
- Generate 16 artificial sequences from a known tree and known rates (JTT model)
  - Carboxypeptidase A2 precursor as ancestor, length = 147
- Goal: recovering the substitution rates



Phylogenetic tree used to generate 16 sequences

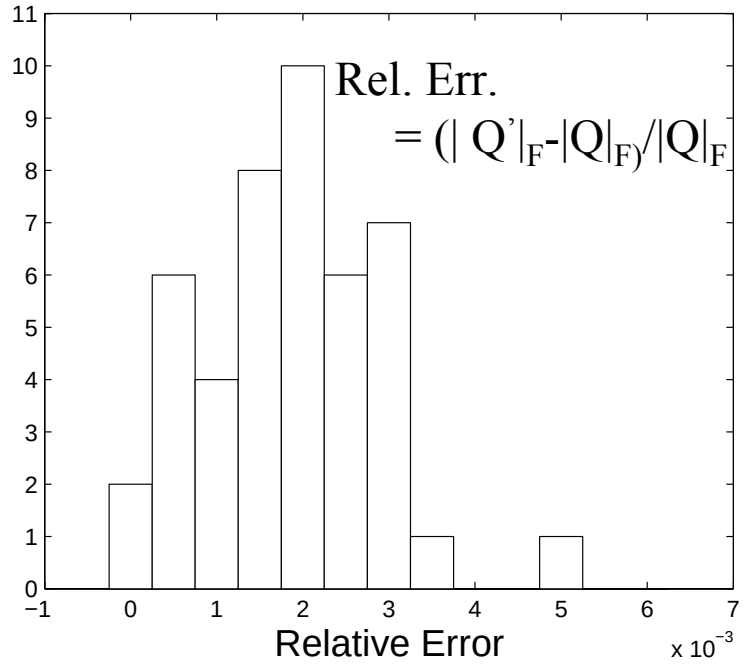


Convergence of the Markov chain



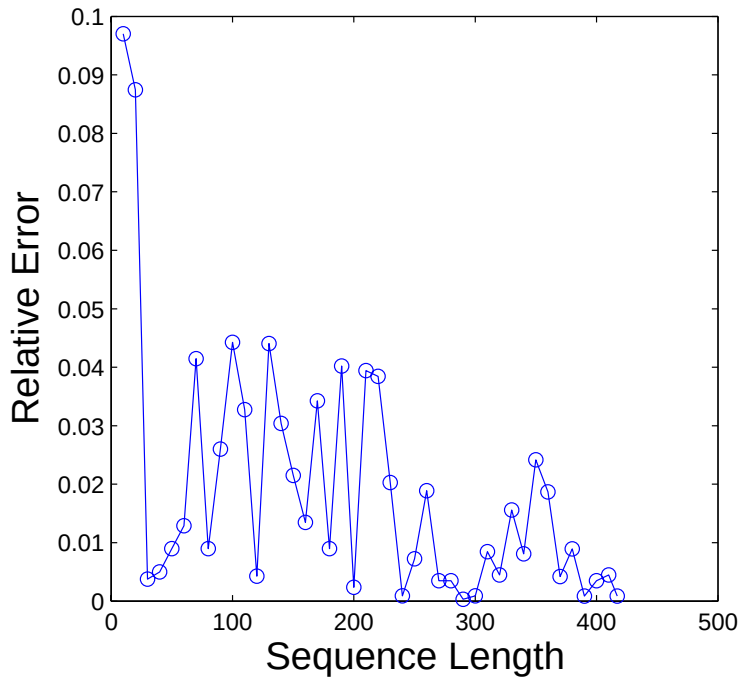
Estimations from two initial conditions are very similar to the true values of residue substitution rates.

Accurate Estimation with > 20 residues and random initial values



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

All Relative Error < 5%.



Accurate when > 20 residues in length.

Q' matrix estimated by Bayesian MCMC has small relative error by Frobenius norm (<5%) to Q.

Surface motifs known to be biologically important

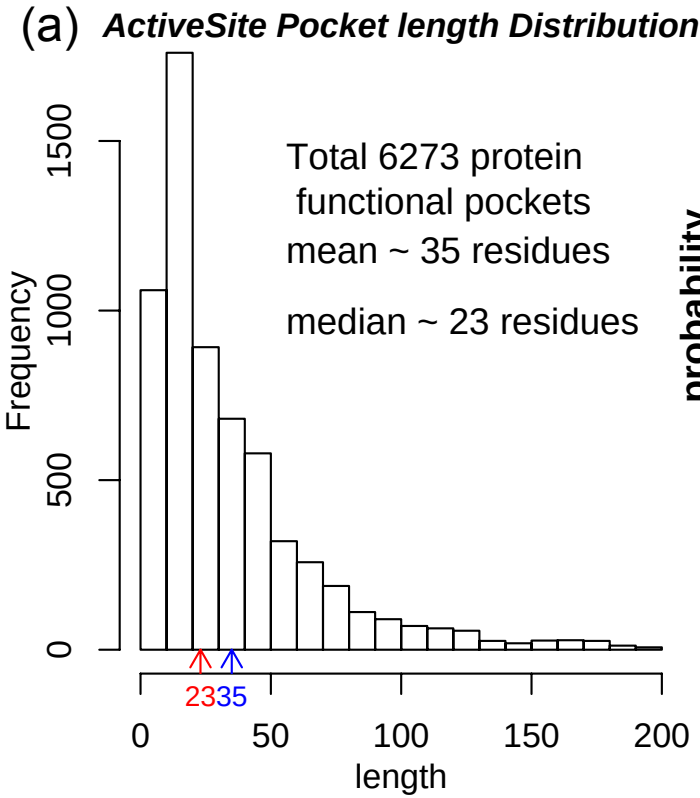


Fig (a). From 6,273 protein active site pockets, 80% have between 8 and 200 a.a.

- The average length: 35 residues.

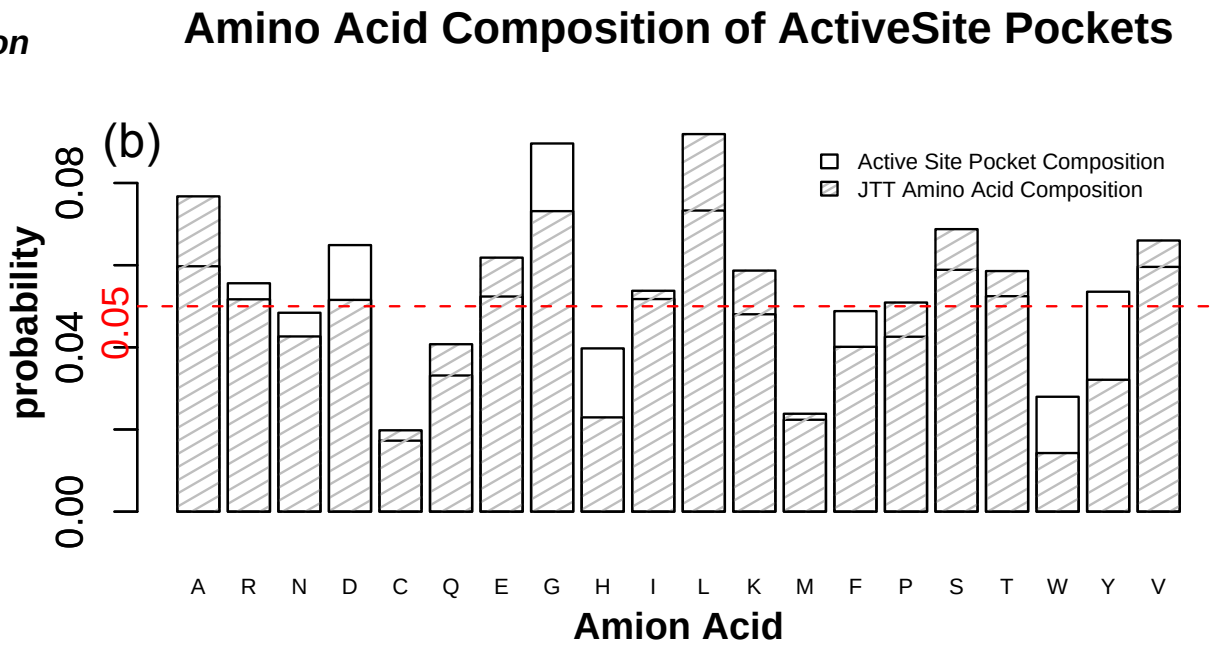
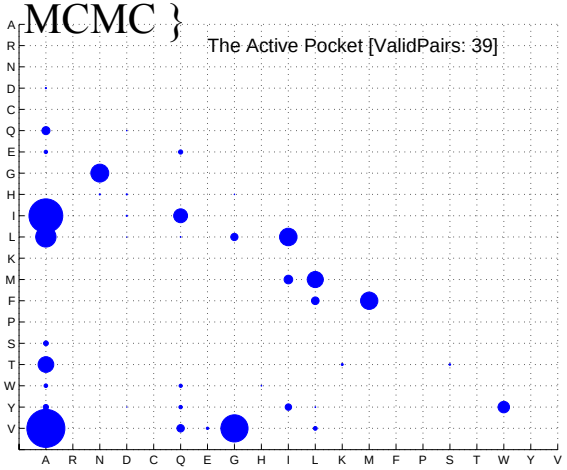


Fig (b). Compare amino acid composition of functional site pockets (7,173 protein pockets) with protein sequence database (16,300 proteins) by JTT.

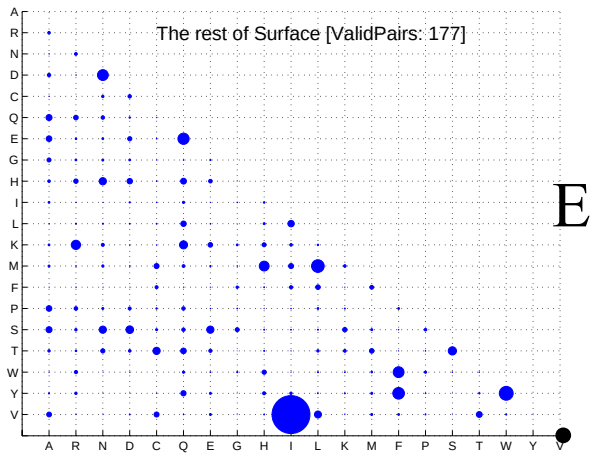
Functionally important residues:
His (H), Asp (D), Tyr (Y), Trp (W) and Gly (G)
Phe (F), Asn (N), and Arg (R).

Evolutionary rates of binding sites and other regions are different

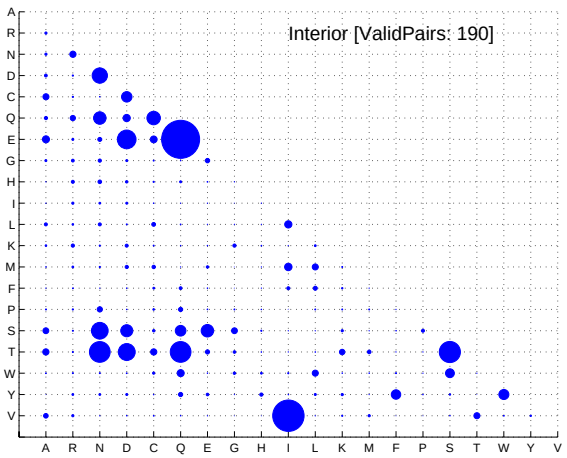
S_{ij} (i, j) are residues shown in the same column of MSA defined as Sampled Pairs and S_{ij} are estimated by Baysian MCMC }



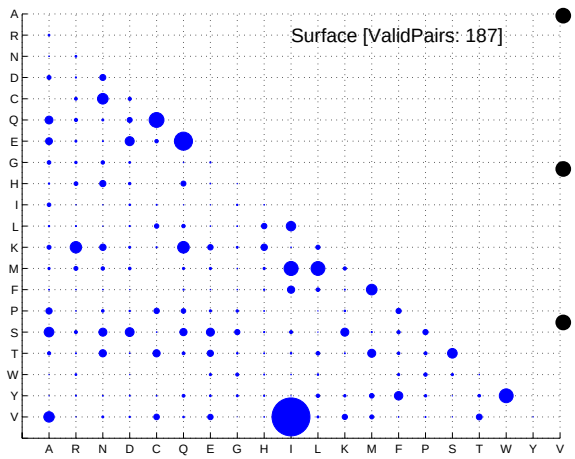
(a)



(b)



(c)



(d)

Residues on protein functional surface experience different selection pressure.

Estimated substitution rate matrices of amylase:

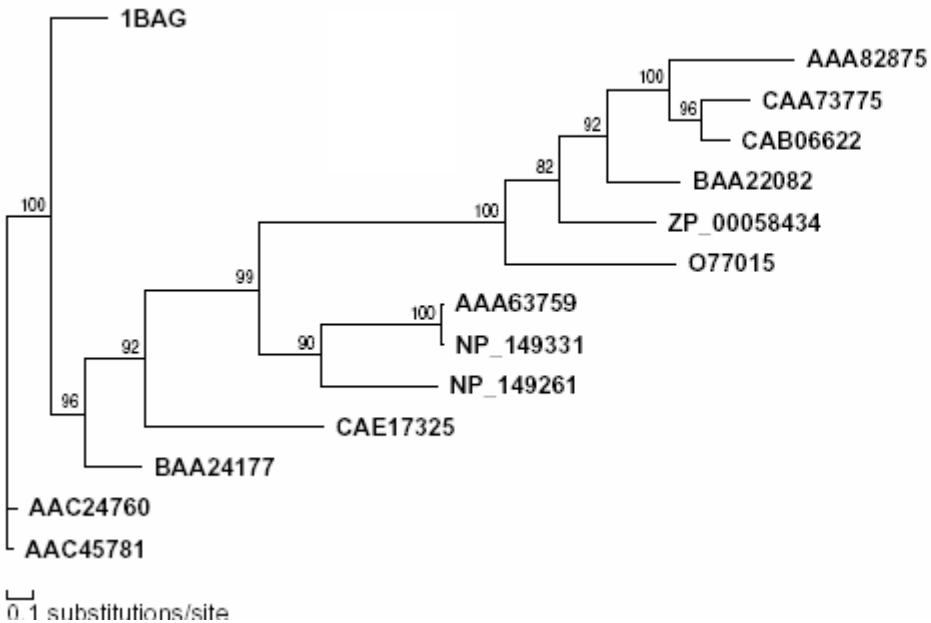
- functional surface residues.
- The remaining surface,
- The interior residues
- All surface residues.

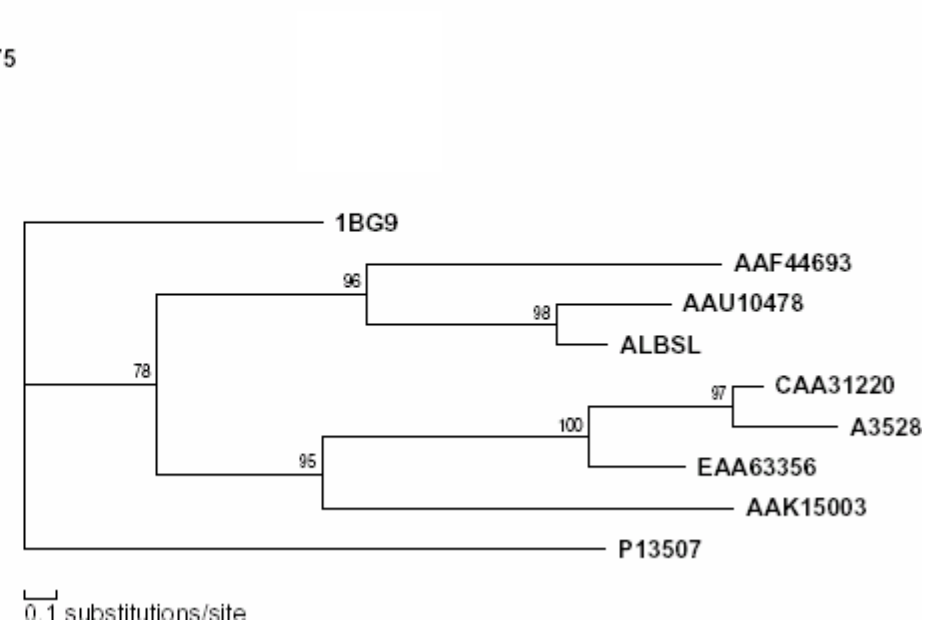
Improved functional prediction

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.

Results for Amylase

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

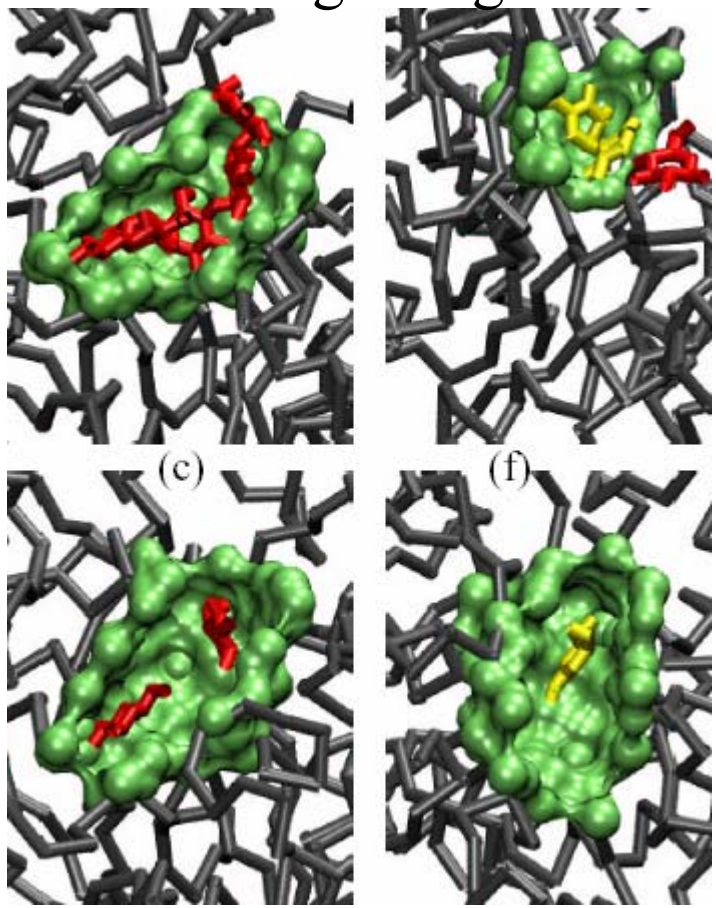
- Altogether: 69
  - All belong to amylase (EC 3.2.1.1)

Comparison:

- Annotated enzyme structure database (Thorton): 75.

Query: *B. subtilis* Barley

1bag 1bg9



Hits: human

1b2y 1u2y

22% 23%

Comparison with others

Benchmark data:

- Enzyme Structure Database (ESD):

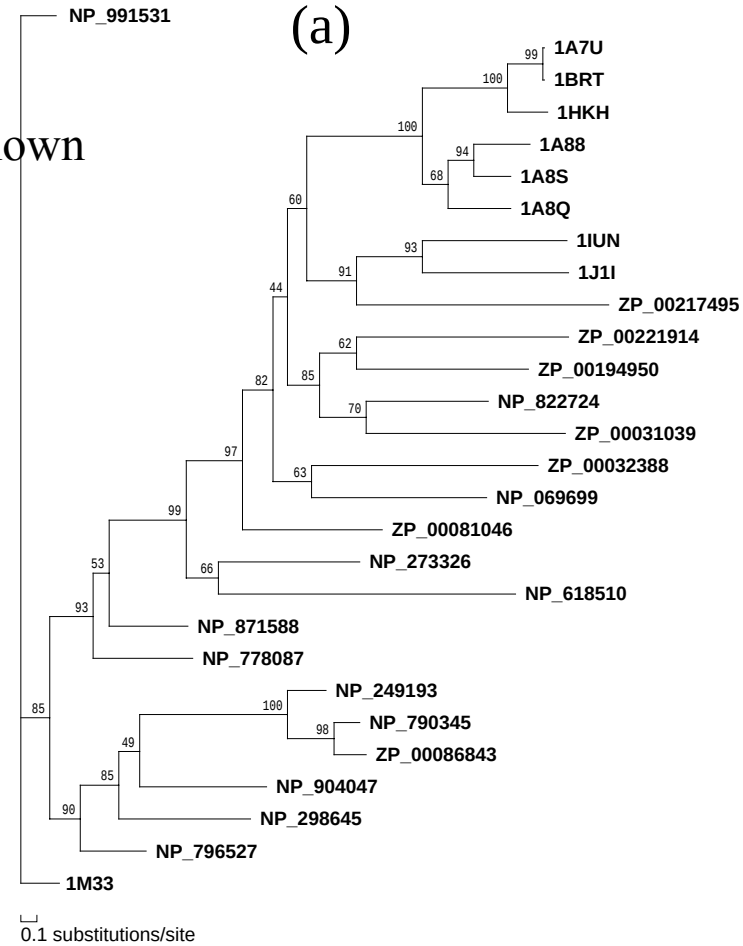
template	our results	ESD	psi-blast
1bag	58	75	31
1bg9	48	75	11
union	69	75	41

- Psi-blast: does not contain information about which surface region, active residues, and geometry; contains many uninterpretable false positives.
- Ssearch: 32 structures found.

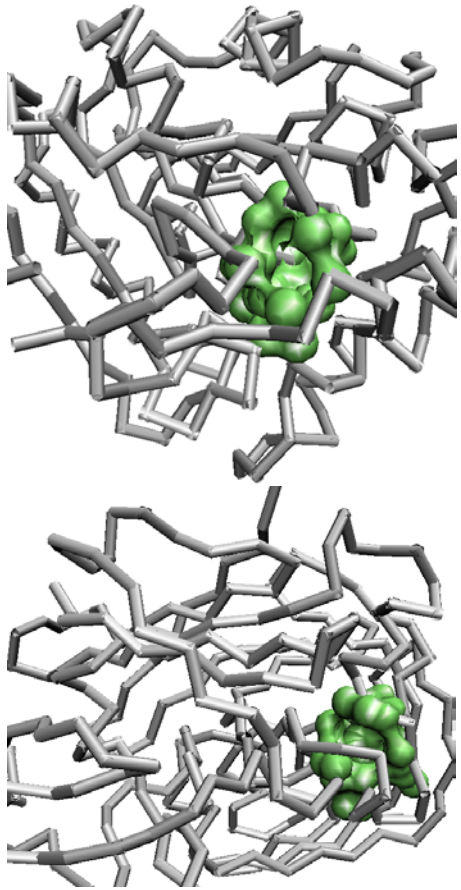
Inferring biological functions of protein BioH

Protein of unknown functions from structural genomics

(a)



(b)



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 proteinase A (2jxr) CASTp id=104.

Summary

- Model for evolution of binding surfaces:
  - Continuous Markov process for residue substitution.

- Bayesian Markov chain Monte Carlo works for residue rates:
  - Fast convergency, insensitive to perturbation of tree topology and representative sequences.
  - Small relative errors (<5%) for > 20 residues.

- Can be used for function prediction.
  - Database search of functionally related binding surfaces.

Collaborators

- Andrew Binkowski (UIC)
- Jinfeng Zhang (UIC)

Acknowledgement

- NSF CAREER DBI 0133856 and DBI 0078270
- NIH GM68958
- ONR MURI
- Whitaker Foundation