Mutating Proteins in the Computer

Bala Krishnamoorthy
Washington State University

www.wsu.edu/~kbala

What is a Protein?

* large biomolecules made of amino acids (AAs)

(image: www)

What is a Protein?

* large biomolecules made of amino acids (AAs)
* amino acid consists of backbone and an 'R' group (residue)

(image: www)

What is a Protein?

* large biomolecules made of amino acids (AAs)
* amino acid consists of backbone and an 'R' group (residue)
* 20 different AAs
* proteins are ~25 to more than 2000 AAs long

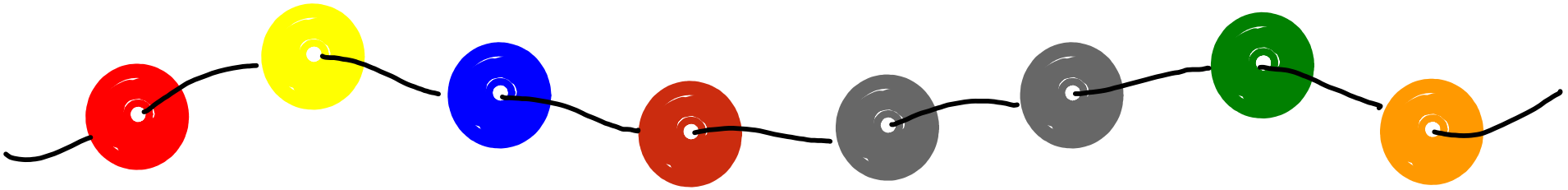
(image: www)

Levels of Protein Structure

* The AA-sequence (primary structure)

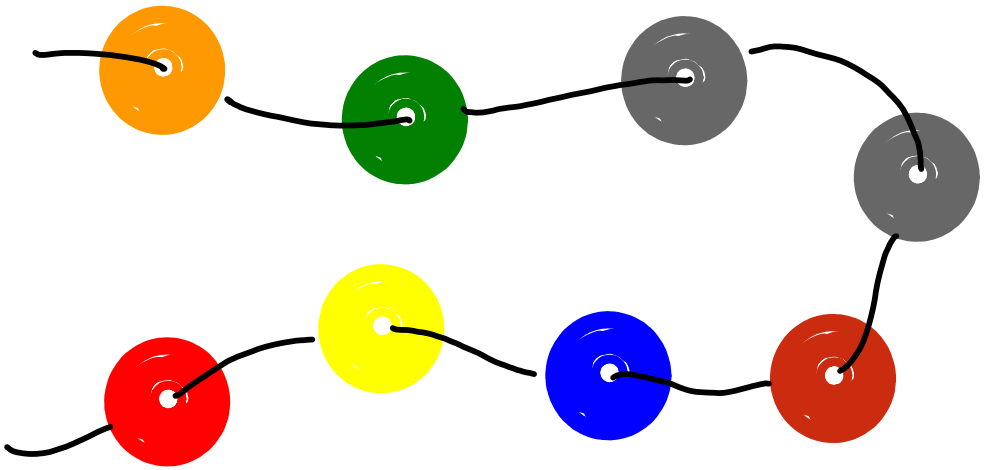
Levels of Protein Structure

* The AA-sequence (primary structure)
* Model - beads in a open necklace
  color of bead $\Leftrightarrow$ AA identity
  string $\Leftrightarrow$ backbone



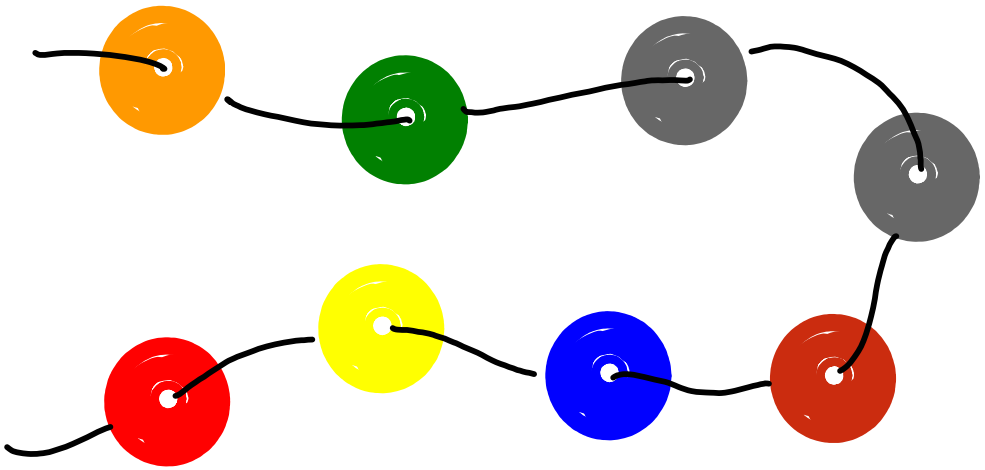
Secondary Structure

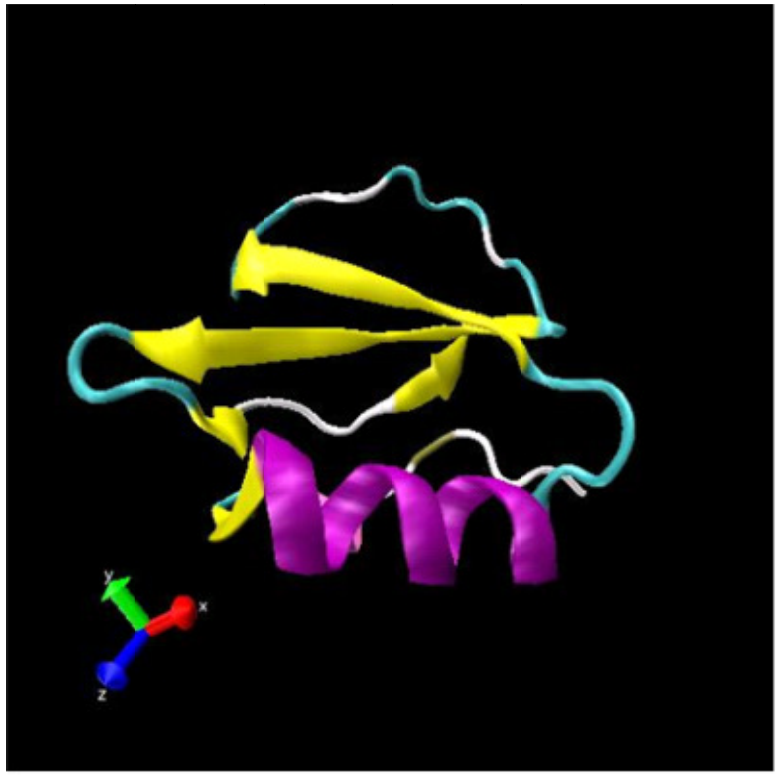
* folding of open necklace into 3D structural units



Secondary Structure

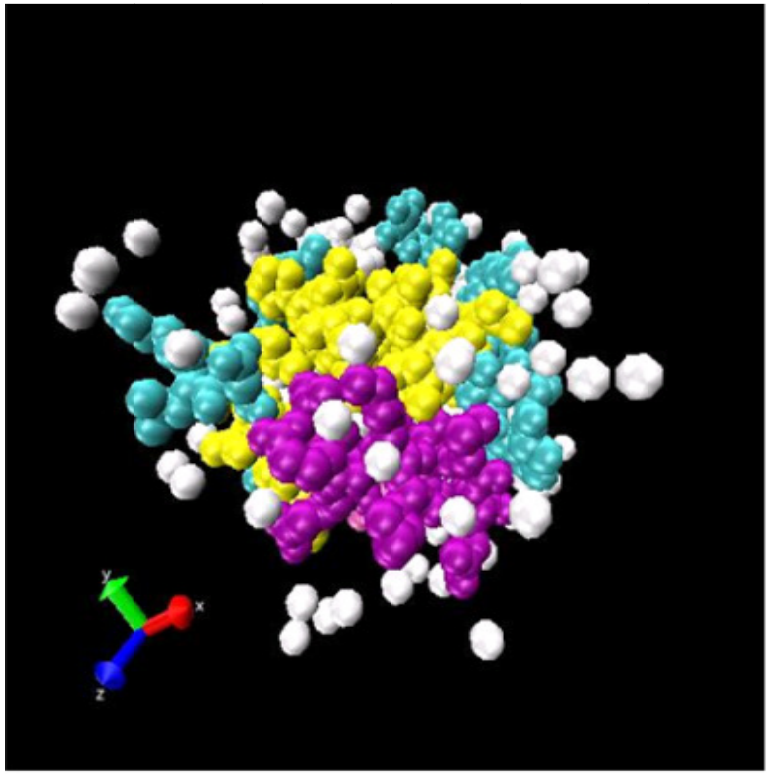
* folding of open necklace into 3D structural units
* alpha helices, beta sheets, and random coils





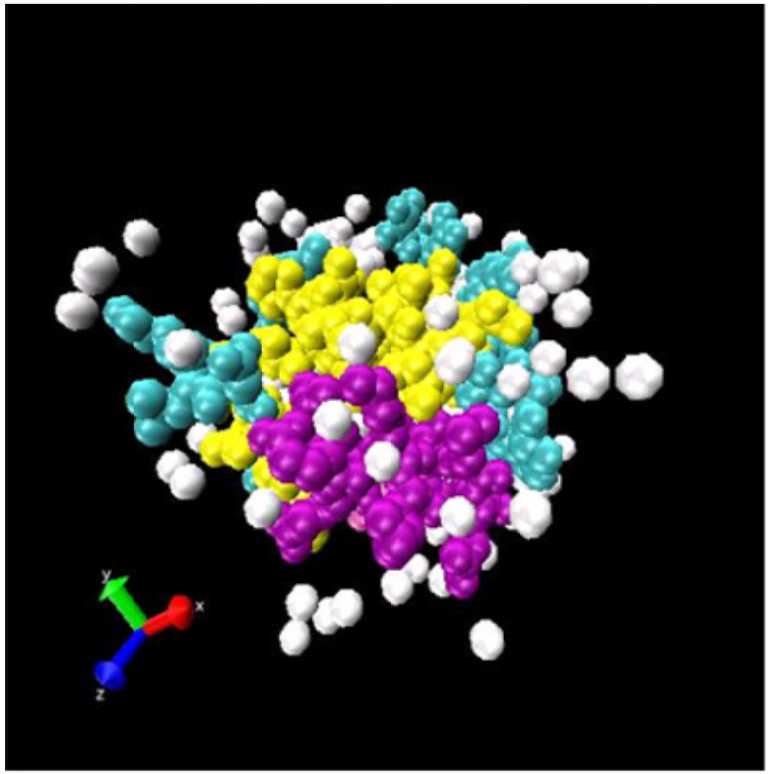
Tertiary Structure

* α-helices, β-sheets, and coils coming together



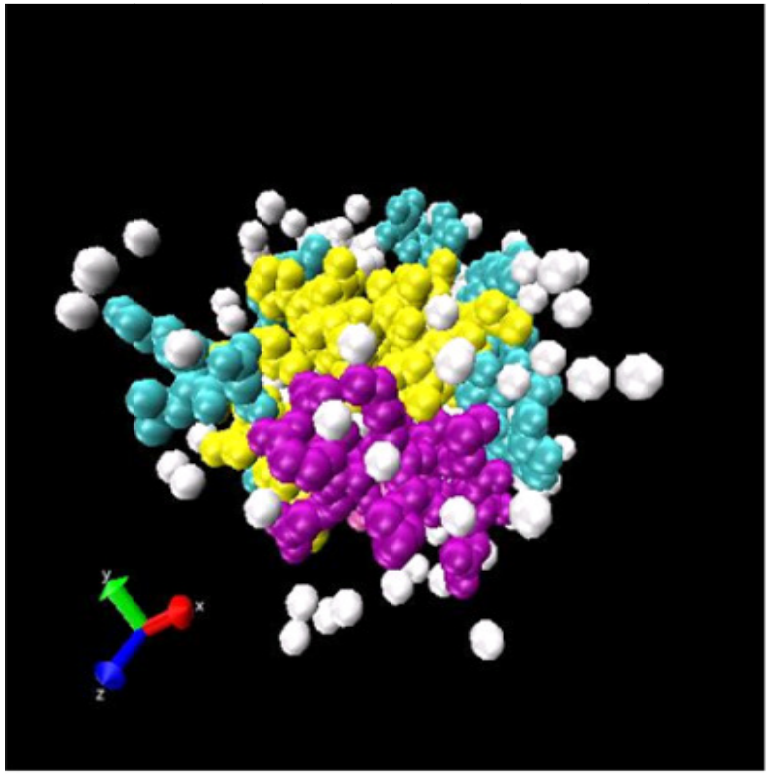
Tertiary Structure

* α-helices, β-sheets, and coils coming together
* minimum energy configuration



Tertiary Structure

* α-helices, β-sheets, and coils coming together
* minimum energy configuration
* sequence ⇒ structure ⇒ function



Tertiary Structure

* α-helices, β-sheets, and coils coming together
* minimum energy conformation
* sequence $\Rightarrow$ structure $\Rightarrow$ function

Protein folding problem

Mutagenesis

* process of changing one or more amino acids in a protein

Mutagenesis

* process of changing one or more amino acids in a protein

* want to make the protein more or less stable, reactive, soluble, etc.

Mutagenesis

* process of changing one or more amino acids in a protein

* want to make the protein more or less stable, reactive, soluble, etc.

* standard technique used in drug design, protein engineering, ...

Mutagenesis in the Lab

Lab Rule #1: If an experiment works, something's wrong ;)

(image: www)

Mutagenesis in the Lab

* long, hands-on process
* DNA is "spliced out" and replaced

(image: www)

Mutagenesis in the Lab

* long, hands-on process
* DNA is "spliced out" and replaced
* often requires large numbers of mutants to be created

(image: www)

Mutagenesis in the Lab

* long, hands-on process
* DNA is "spliced out" and replaced
* often requires large numbers of mutants to be created

? Can we **predict** the effects of mutations computationally ?

(image: www)

Results

✓ stability mutagenesis
- four-body scoring function

✓ solubility mutagenesis
- three-body scoring function optimized using linear programming

✓ temperature-sensitive (Ts) mutants
- 133 features tested:
sequence, structure, site, neighborhood

Results

✓ stability mutagenesis
- four-body scoring function → Delaunay tessellation

✓ solubility mutagenesis
- three-body scoring function
- optimized using linear programming

✓ temperature-sensitive (Ts) mutants
- 133 features tested:
  sequence, structure, site, neighborhood

Results

✓ stability mutagenesis
- four-body scoring function → Delaunay tessellation
✓ solubility mutagenesis
- three-body scoring function
- optimized using linear programming
✓ temperature-sensitive (Ts) mutants
- 133 features tested:
  - sequence, structure, neighborhood
  - mutual information of Ts mutants

Stability Mutagenesis

* Bioinformatics, 2007; with C. Deutsch
* predict whether stability increases or decreases after mutation(s)

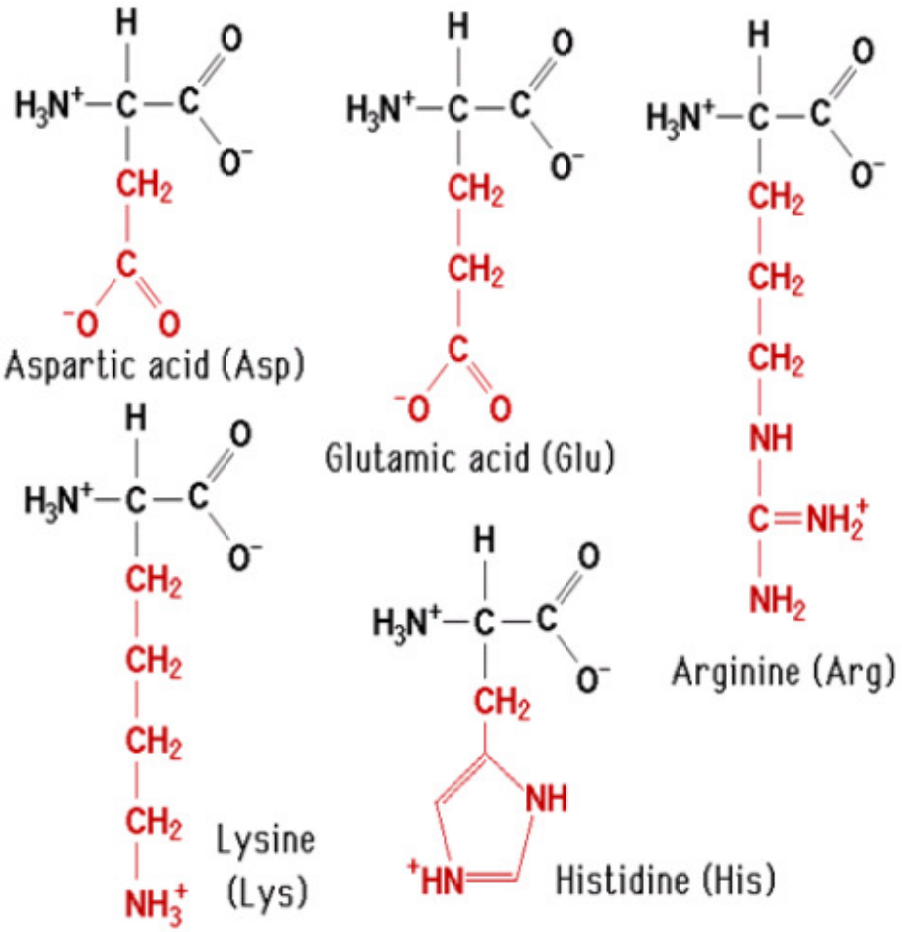
Stability Mutagenesis

* Bioinformatics, 2007; with C. Deutsch
* predict whether stability increases or decreases after mutation(s)
* scoring function – knowledge-based method

Stability Mutagenesis

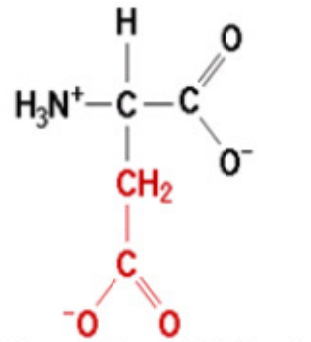
* Bioinformatics, 2007; with C. Deutsch
* predict whether stability increases or decreases after mutation(s)
* scoring function – knowledge-based method
* probabilities of how often do groups of AAs appear close-by in proteins
* need (simpler) representation of protein structure

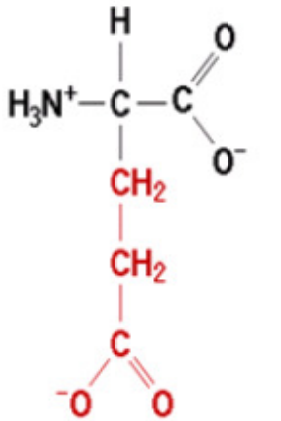
Representing the Amino Acids

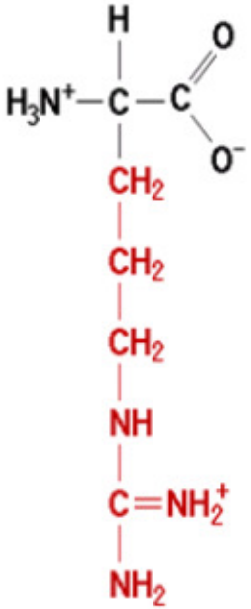


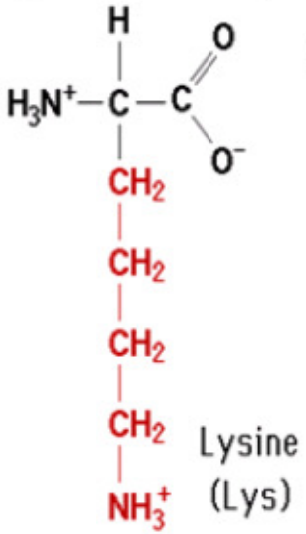
* 'R' group varies
* alpha C (C_α) is the backbone carbon atom
* beta C (C_β) is the first carbon atom in "R"

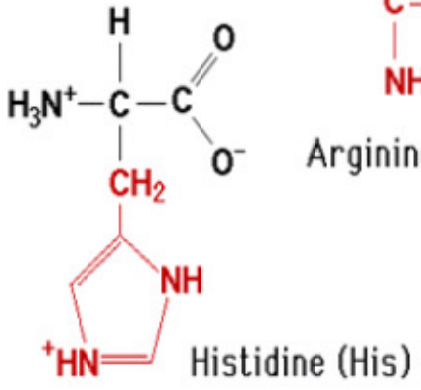
Representing the Amino Acids











* 'R' group varies
* alpha C (C_α) is the backbone carbon atom
* beta C (C_β) is the first carbon atom in "R"
* side chain center — centroid of all atoms in 'R' (including C_α)

Two-Body Potentials

* previous scoring functions counted how many times pairs of AAs are "close to each other"
  - Miyazawa-Jernighan, 1985, 1996
  - Sippl, 1990

Two-Body Potentials

* previous scoring functions counted how many times pairs of AAs are "close to each other"
  - Miyazawa-Jernighan, 1985, 1996
  - Sippl, 1990
* "close to each other" defined using distance cut-offs, e.g.,

  C_α-C_α distance $\leq 8\text{A}^\circ$

Two-Body Potentials

* previous scoring functions counted how many times pairs of AAs are "close to each other"
  - Miyazawa-Jernighan, 1985, 1996
  - Sippl, 1990
* "close to each other" defined using distance cut-offs, e.g.,

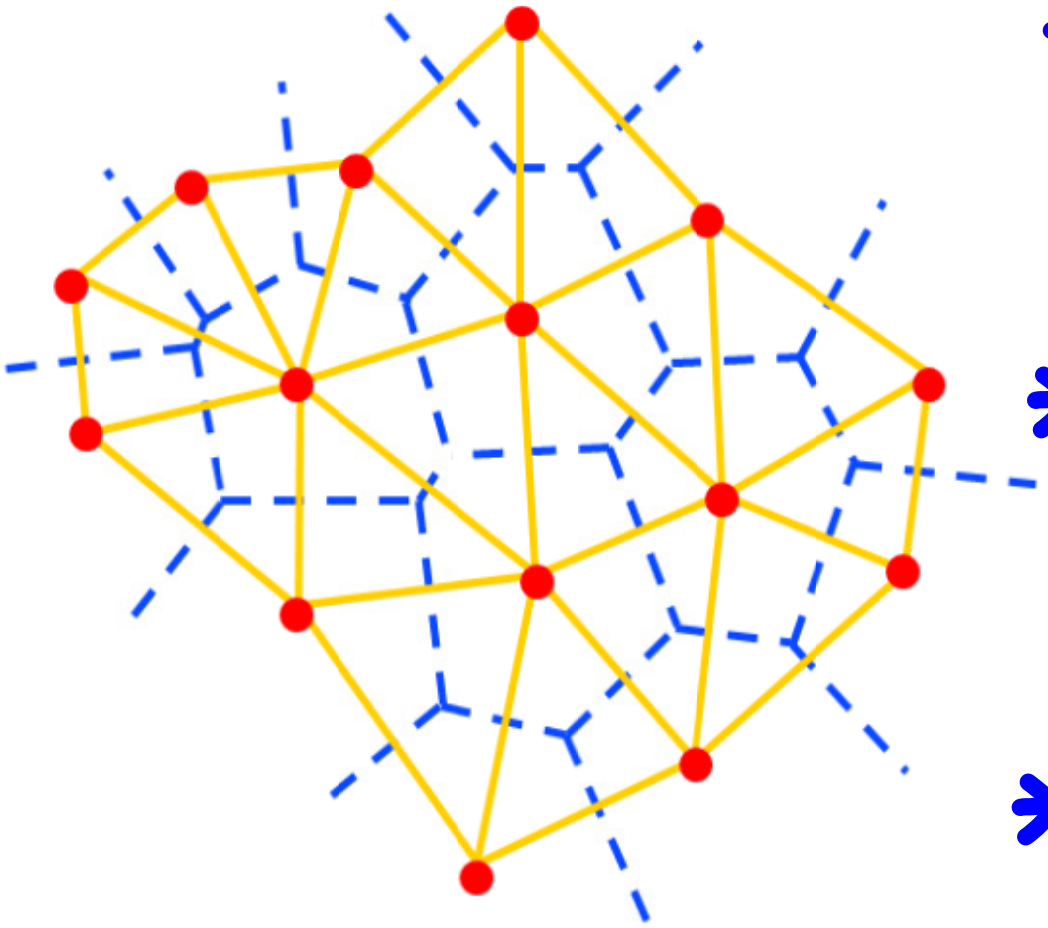
  C_α-C_α distance $\leq 8A^\circ$

$\Rightarrow$ 7.99 A°: neighbors; 8.01 A°: not neighbors!

Voronoi/Delaunay Tessellation

* Voronoi cells - neighborhoods of points (convex)
* Delaunay triangles - clusters of three nearest neighbors

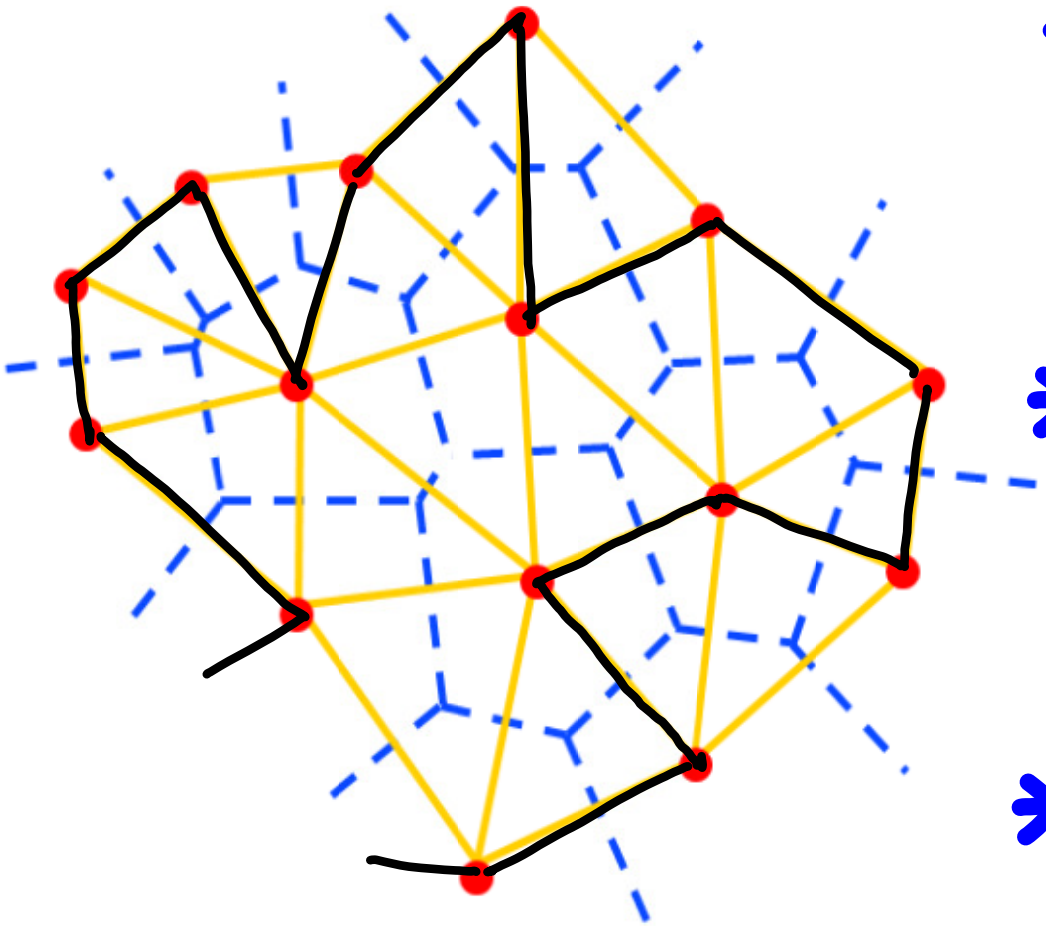
Voronoi/Delaunay Tessellation



* Voronoi cells - neighborhoods of points (convex)
* Delaunay triangles - clusters of three nearest neighbors
* Get tetrahedra in 3D

* fast algorithms available ($O(n^2)$ in 3D)

Voronoi/Delaunay Tessellation

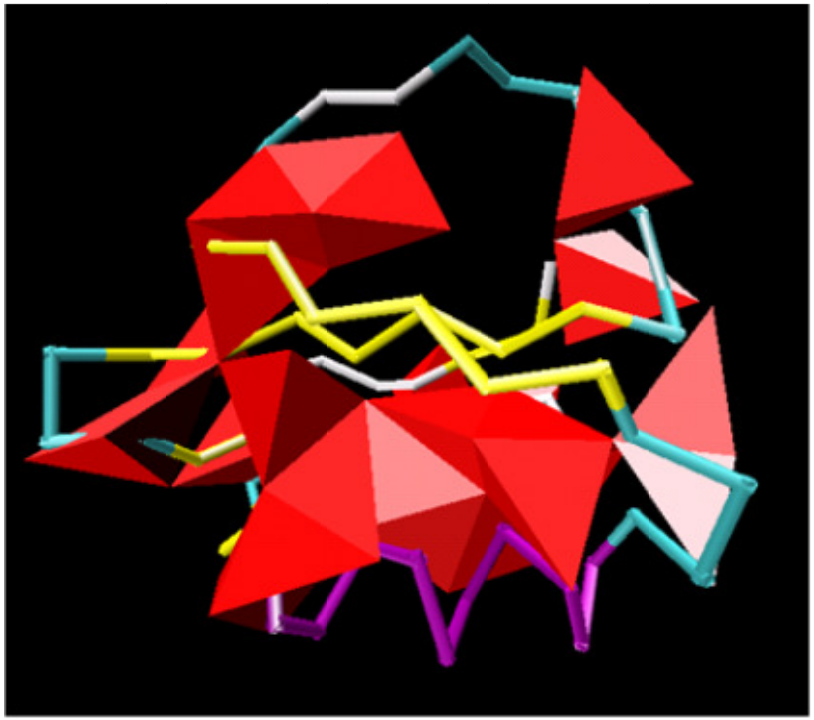


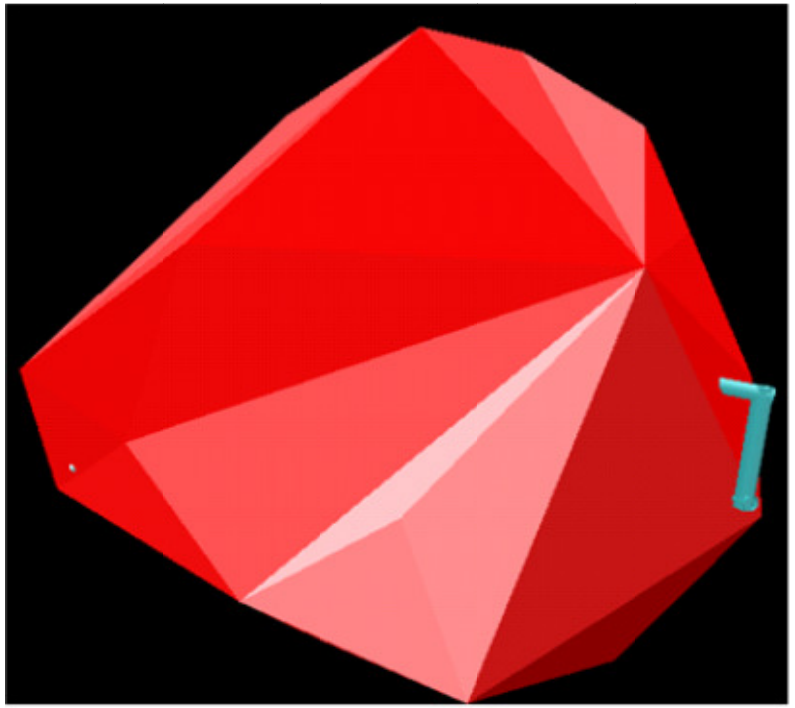
* Voronoi cells - neighborhoods of points (convex)
* Delaunay triangles - clusters of three nearest neighbors
* Get tetrahedra in 3D

* fast algorithms available ($O(n^2)$ in 3D)

Delaunay Tessellation of Proteins

* represent each AA by a point at the side chain center
* discard biologically irrelevant tetrahedra (>12 Å°)



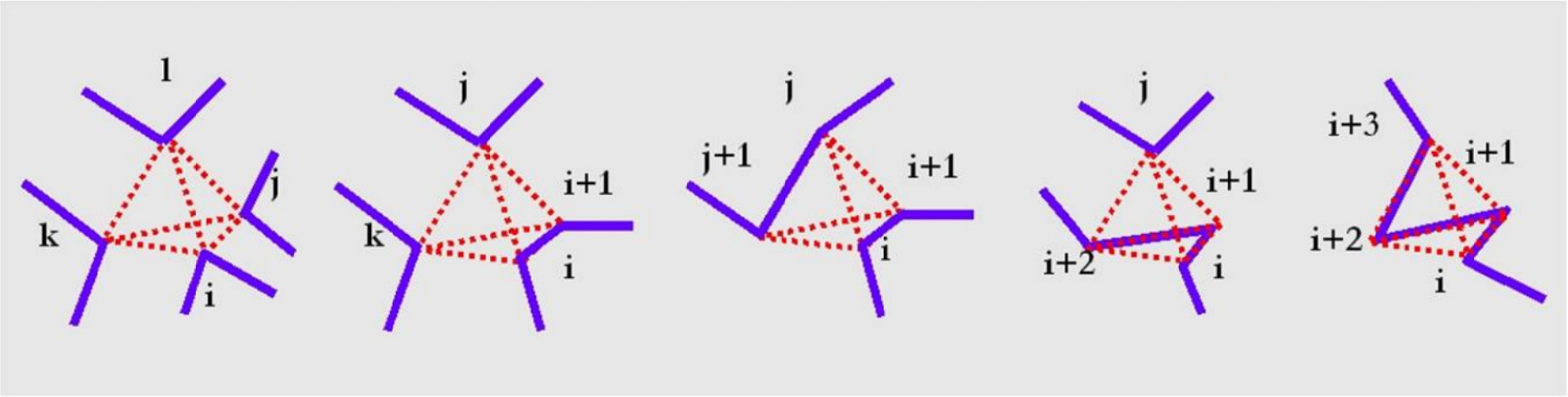


Types of Tetrahedra

* because proteins have backbone
* tetrahedra formed due to backbone connectivity less indicative of 3D structure

Types of Tetrahedra

* because proteins have backbone
* tetrahedra formed due to backbone connectivity less indicative of 3D structure
* 5 types of tets – based on how many AAs in the tet are connected



Score of a Tetrahedron

$$q^t_{ijkl} = \log\left(\frac{f^t_{ijkl}}{p^t_{ijkl}}\right)$$

f^t_{ijkl} → observed

p^t_{ijkl} → expected

Score of a Tetrahedron

$$q^t_{ijkl} = \log\left(\frac{f^t_{ijkl}}{p^t_{ijkl}}\right)$$

f^t_{ijkl} → observed

p^t_{ijkl} → expected

f^t_{ijkl} = frequency of observing quadruplet of AA $\{i,j,k,l\}$ in tet type t

Score of a Tetrahedron

$$q^t_{ijkl} = \log\left(\frac{f^t_{ijkl}}{p^t_{ijkl}}\right)$$

f^t_{ijkl} → observed

p^t_{ijkl} → expected

f^t_{ijkl} = frequency of observing quadruplet of AA $\{i,j,k,l\}$ in tet type t

p^t_{ijkl} = expected frequency of AA-quadruplet $\{i,j,k,l\}$ in tet type t

$$p^t_{ijkl} = C\, a_i a_j a_k a_l P_t$$

a_i = frequency of AA i

P_t = frequency of tet type t

C = combinatorial factor

Mutagenesis Score

* total score of protein $S = \sum_{\text{all tets}} q^{t}_{ijkl}$
  - used for fold recognition
    Krishnamoorthy & Tropsha, Bioinformatics, 2003.

Mutagenesis Score

* total score of protein $S = \sum_{\text{all tets}} q^{t}_{ijkl}$
  - used for fold recognition
    Krishnamoorthy & Tropsha, Bioinformatics, 2003.
* use wild-type (WT) structure for mutant, but change sequence due to mutation

Mutagenesis Score

* total score of protein $S = \sum_{\text{all tets}} q^{t}_{ijkl}$
  - used for fold recognition
    Krishnamoorthy & Tropsha, Bioinformatics, 2003.
* use wild-type (WT) structure for mutant, but change sequence due to mutation
* mutagenesis score $\Delta = (S_{MUT} - S_{WT})/S_{WT}$

Mutagenesis Score

* total score of protein $S = \sum_{\text{all tets}} q^{t}_{ijkl}$
  - used for fold recognition
    Krishnamoorthy & Tropsha, Bioinformatics, 2003.
* use wild-type (WT) structure for mutant, but change sequence due to mutation
* mutagenesis score $\Delta = (S_{MUT} - S_{WT})/S_{WT}$
* $\Delta > 0 \iff$ increase in stability

A Dataset of Mutants

* databases (e.g., ProTherm), and previous studies (e.g., Chang et al., 2006) have collections of single-point mutations

A Dataset of Mutants

* databases (e.g., ProTherm), and previous studies (e.g., Chang et al., 2006) have collections of single-point mutations
* single- and multi-point mutations handled in the same way by our scoring function

A Dataset of Mutants

* databases (e.g., ProTherm) and previous studies (e.g., Chang et al., 2006) have collections of single-point mutations handled in the same way by varying function
* comprehensive literature search to assemble database of single- and multi-point mutations from 24 papers

Stability Mutagenesis - Results

* accuracy on our single-/multi-point mutant dataset = 80.5%

Stability Mutagenesis - Results

* accuracy on our single-/multi-point mutant dataset = 80.5%

* accuracy on single-point mutant set (Chang et al., 2006; Guerois et al., 2002; Topham et al., 1997) = 66%

Stability Mutagenesis - Results

* accuracy on our single-/multi-point mutant dataset = 80.5%

* accuracy on single-point mutant set (Chang et al., 2006; Guerois et al., 2002; Topham et al., 1997) = 66%

* FOLD-X (Guerois et al., 2002) has an accuracy of 68% on our dataset

Stability Mutagenesis – Results

* stability changes quantified for 130 mutants:
  Spearman rank correlation with Δ: 0.67

Stability Mutagenesis - Results

* stability changes quantified for 130 mutants: Spearman rank correlation with Δ: 0.67
* Combinatorial mutagenesis: change each of a set of AAs to all other 19 AAs
  - 6 AAs mutated → 64 million mutants!
  - (re)score only tets formed by at least one of the 6 mutated AAs

Stability Mutagenesis - Results

* stability changes quantified for 130 mutants: Spearman rank correlation with Δ: 0.67
* Combinatorial mutagenesis: change each of a set of AAs to all other 19 AAs
  - 6 AAs mutated → 64 million mutants!
  - (re)score only tets formed by at least one of the 6 mutated AAs

Noteworthy:

✓ scoring function not "trained"

Stability Mutagenesis - Results

* stability changes quantified for 130 mutants: Spearman rank correlation with Δ: 0.67
* Combinatorial mutagenesis: change each of a set of AAs to all other 19 AAs
  - 6 AAs mutated → 64 million mutants!
  - (re)score only tets formed by at least one of the 6 mutated AAs

Noteworthy:

✓ scoring function not "trained"

✓ more accurate for multi-point mutants

Solubility Mutagenesis

* AlMoB, 2010; with Y. Tian & C. Deutsch
* predominantly surface property

Solubility Mutagenesis

* AlMoB, 2010; with Y. Tian & C. Deutsch
* predominantly surface property
* solvent accessible surface area
  — measures exposure to solvent

Solubility Mutagenesis

* AlMoB, 2010; with Y. Tian & C. Deutsch
* predominantly surface property
* solvent accessible surface area
  — measures exposure to solvent
* correlate propensities of groups of AAs to be on/near surface with solubility

Solubility Mutagenesis

* AlMoB, 2010; with Y. Tian & C. Deutsch
* predominantly surface property
* solvent accessible surface area
  — measures exposure to solvent
* correlate propensities of groups of AAs to be on/near surface with solubility
* use the framework of Delaunay tessellation (DT)

Three Body Contacts

* protein surface represented by collection of "exposed" triangles in DT

Three Body Contacts

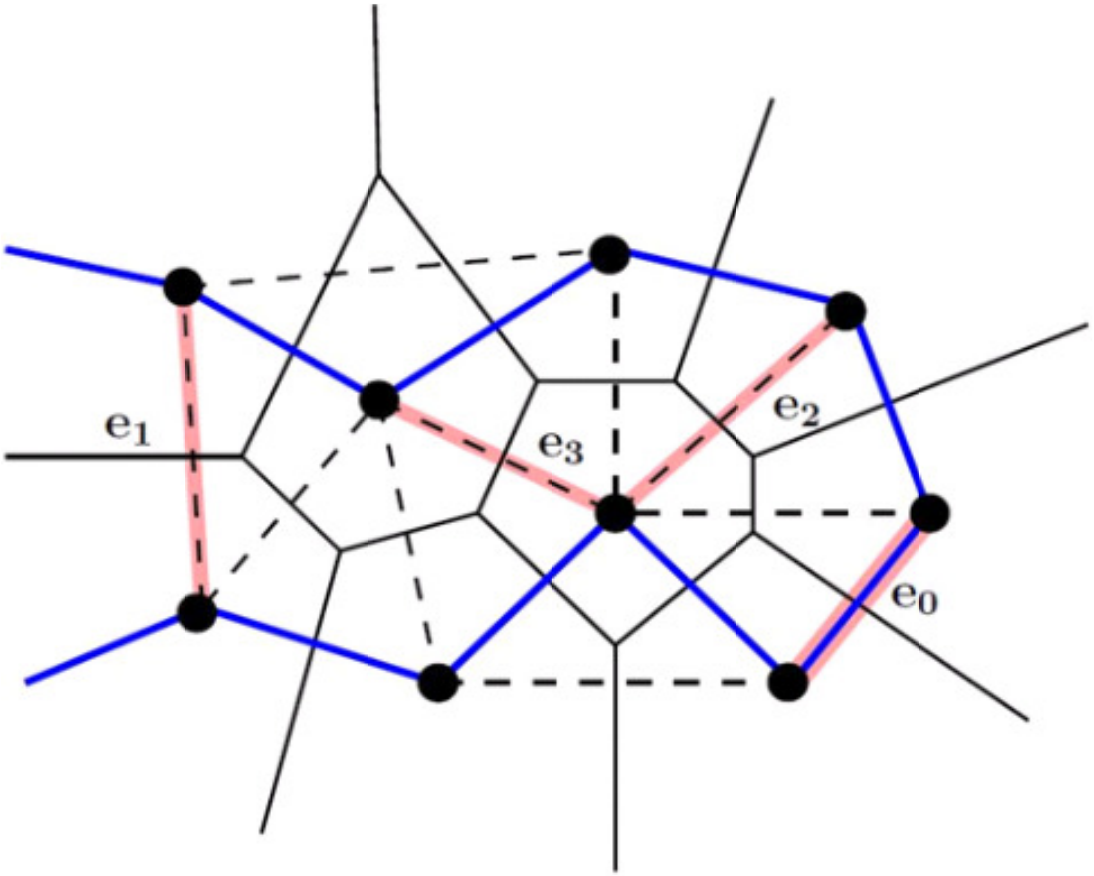
* protein surface represented by collection of "exposed" triangles in DT
* 3-body log-likelihoods

Three Body Contacts

* protein surface represented by collection of "exposed" triangles in DT
* 3-body log-likelihoods
* buriedness defined in the DT
  - combinatorial definition
  - combines sequence info (AA types)

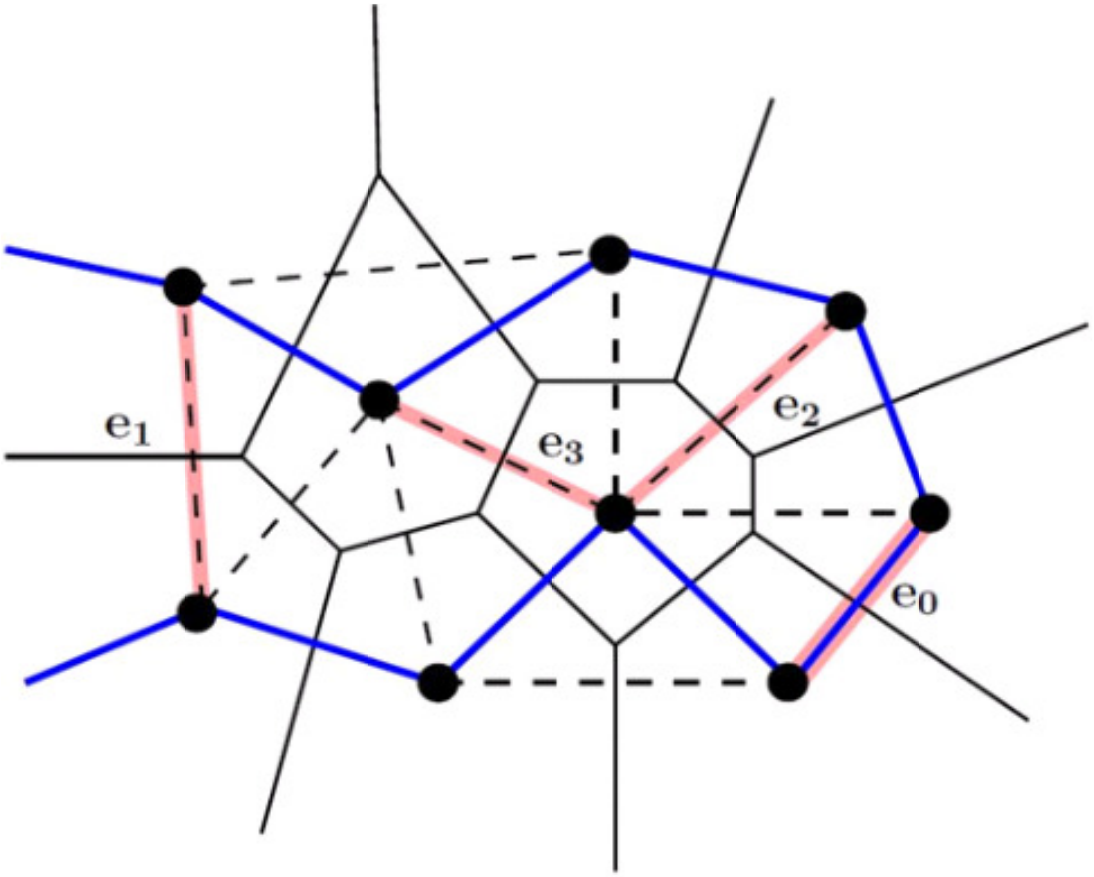
Delaunay Buriedness

* edge is buried if it is part of two triangles

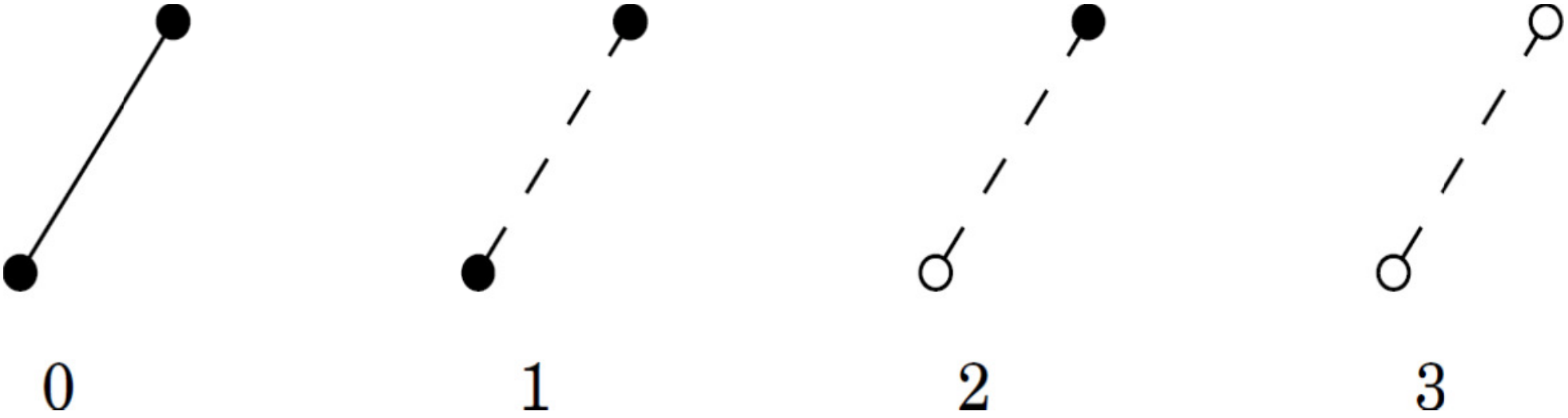


Delaunay Buriedness

* edge (triangle) is buried if it is part of two triangles (tetrahedra)

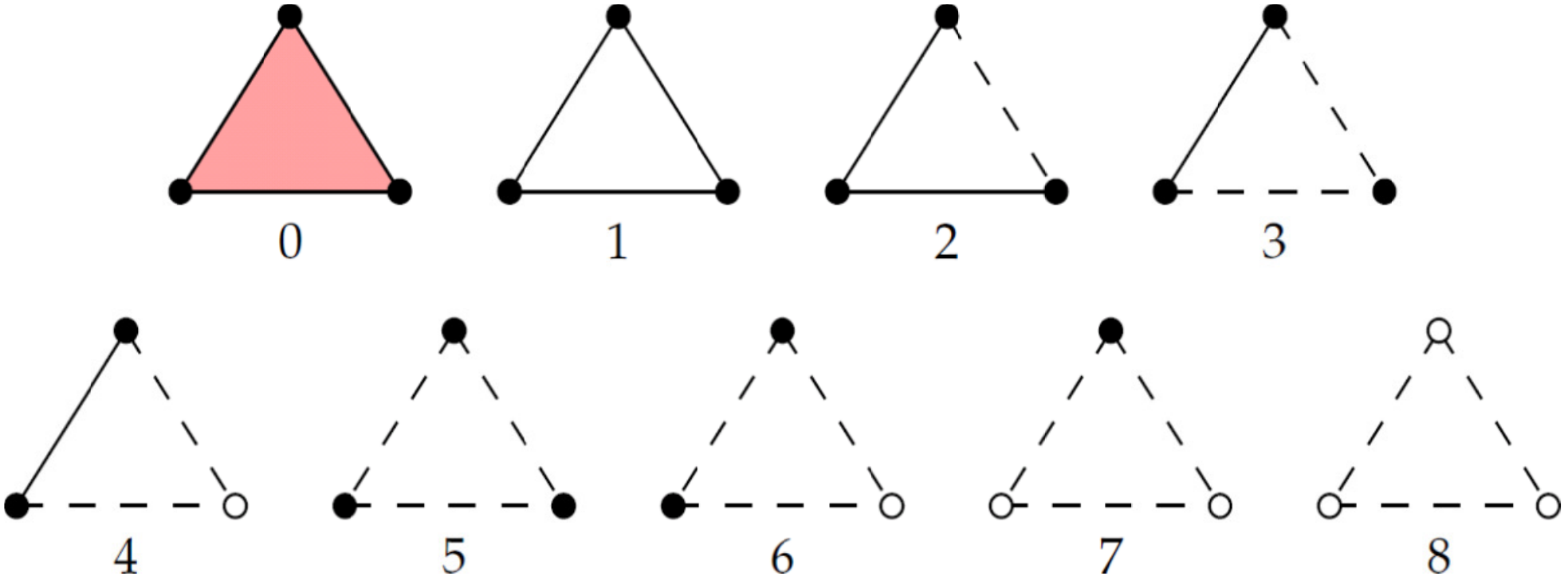


Edge Buriedness – 4 Levels



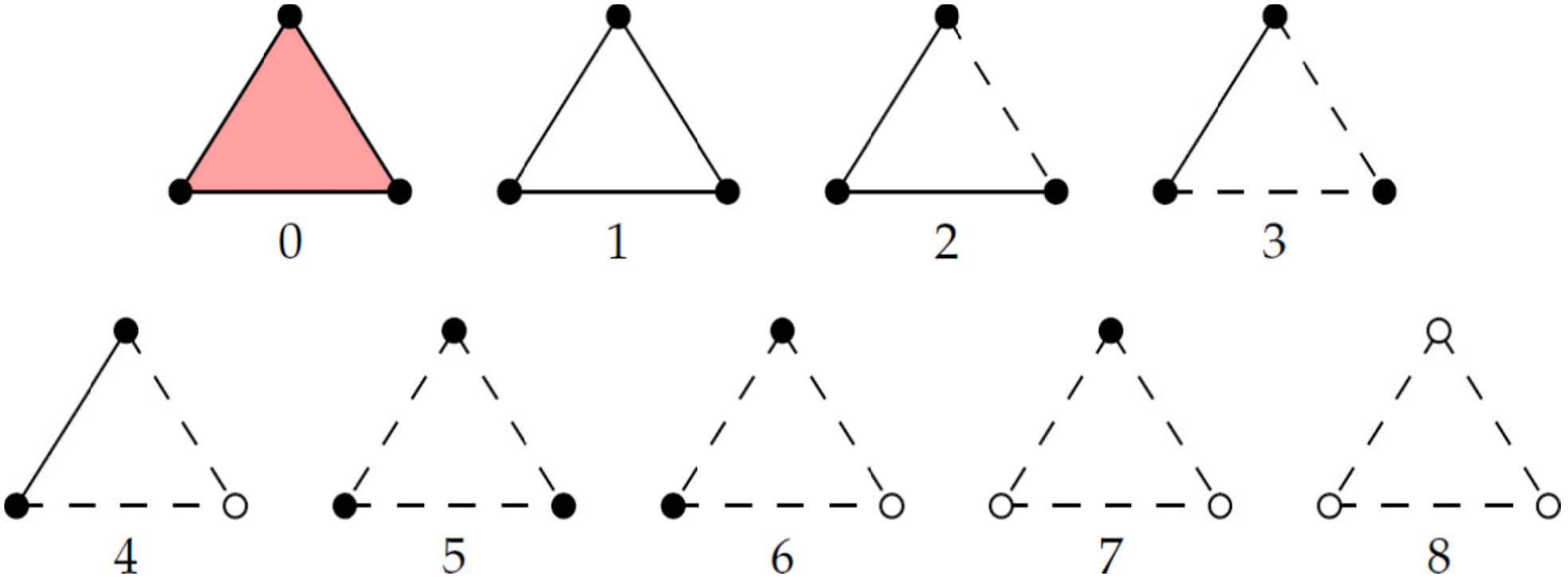
* varies from most non-buried (0), i.e., on the surface, to most buried (3)

Triplet Buriedness – 9 Levels



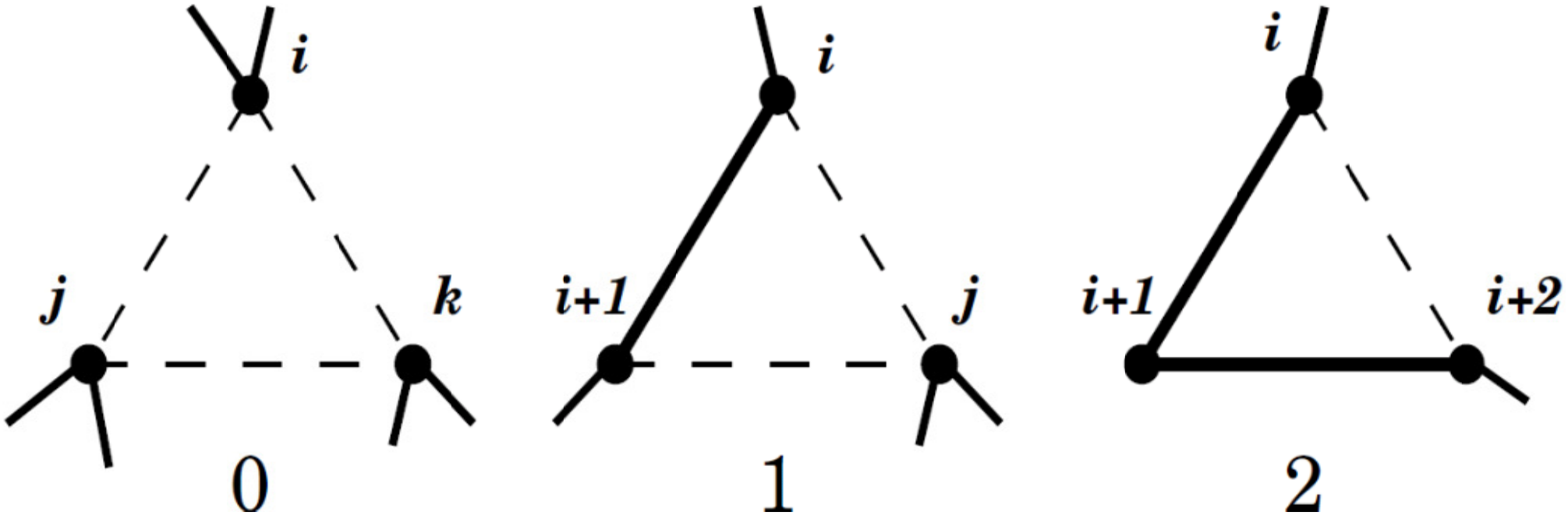
* from most non-buried (0) to most buried (8)

Triplet Buriedness – 9 Levels



* from most non-buried (0) to most buried (8)
* use levels 0-4 (non-buried) for solubility mutagenesis

Triplet Connectivity - 3 Classes



3-Body Scoring Function

* score of a triplet with AAs $\{i,j,k\}$, connectivity class c(0-2), buriedness class b(0-4)

$$q_{ijk}^{cb} = \log\left(\frac{f_{ijk}^{cb}}{p_{ijk}^{cb}}\right)$$

3-Body Scoring Function

* score of a triplet with AAs $\{i,j,k\}$, connectivity class c(0-2), buriedness class b(0-4)

$$q_{ijk}^{cb} = \log\left(\frac{f_{ijk}^{cb}}{p_{ijk}^{cb}}\right)$$

* weighted total score $S = \sum_t w_t q_t$, $t = (i,j,k,c,b)$

3-Body Scoring Function

* score of a triplet with AAs $\{i,j,k\}$, connectivity class c(0-2), buriedness class b(0-4)

$$q_{ijk}^{cb} = \log\left(\frac{f_{ijk}^{cb}}{p_{ijk}^{cb}}\right)$$

* weighted total score $S = \sum_t w_t q_t$, $t = (i,j,k,c,b)$

* mutagenesis score $\Delta = S_{MUT} - S_{WT}$

3-Body Scoring Function

* score of a triplet with AAs $\{i,j,k\}$, connectivity class c (0-2), buriedness class b (0-4)

$$q_{ijk}^{cb} = \log\left(\frac{f_{ijk}^{cb}}{p_{ijk}^{cb}}\right)$$

* weighted total score $S = \sum_t w_t q_t$, $t = (i,j,k,c,b)$
* mutagenesis score $\Delta = S_{MUT} - S_{WT}$
* find w_t by training

A Dataset of Solubility Mutants

* 137 single- and multi-point mutants from 15 different studies

A Dataset of Solubility Mutants

* 137 single- and multi-point mutants from 15 different studies
* mutant also soluble except in 5 cases
  - Idicula-Thomas and Balaji, 2005, 2006
  - Smialowski et al., 2007

    predict whether protein is soluble using only sequence info

A Dataset of Solubility Mutants

* 137 single- and multi-point mutants from 15 different studies
* mutant also soluble except in 5 cases
  - Idicula-Thomas and Balaji, 2005, 2006
  - Smialowski et al., 2007

  predict whether protein is soluble using only sequence info
* our dataset is for increase/decrease of solubility

Training by Linear Programming

* similar to support vector machines (SVM) optimization model, but allows us to set meaningful bounds on w_t

Training by Linear Programming

* similar to support vector machines (SVM) optimization model, but allows us to set meaningful bounds on w_t
* LASSO – Tibshirani, 1996

Training by Linear Programming

* similar to support vector machines (SVM) optimization model, but allows us to set meaningful bounds on w_t
* LASSO – Tibshirani, 1996

$$
\begin{array}{lrcl}
\max & \mu & & \\
\text{s.t.} & \sum_{t\in M_i} w_t Q_t - \sum_{t\in W_i} w_t Q_t & \geq & 1+\varepsilon_i, \forall i \in I; \\
& \sum_{t\in M_i} w_t Q_t - \sum_{t\in W_i} w_t Q_t & \leq & -1-\varepsilon_i, \forall i \in D; \\
& \mu & \leq & \varepsilon_i, \forall i \in I, D; \\
0 \leq & w_t & \leq & 2, \forall t.
\end{array}
$$

← mutants seeing increase & decrease in solubility

M_i, W_i – triplets seeing changes in mutant and WT

Solubility Mutagenesis – Results

* leave one out cross validation (LOOCV)

Measure	LP	SVM	Lasso
Accuracy	0.810	0.708	0.701
MCC	0.617	0.405	0.423
Precision (class I)	0.762	0.661	0.909
Precision (class D)	0.851	0.735	0.661

* 10-fold CV

Measure	LP	SVM	Lasso
Accuracy	0.766	0.752	0.708
MCC	0.545	0.496	0.448
Precision (class I)	0.719	0.705	0.952
Precision (class D)	0.822	0.790	0.664

Solubility Mutagenesis – Results

Noteworthy:

* LP strategy may not work well on other types of data

Solubility Mutagenesis – Results

Noteworthy:

* LP strategy may not work well on other types of data
* scoring function most accurate for surface mutations

Solubility Mutagenesis – Results

Noteworthy:

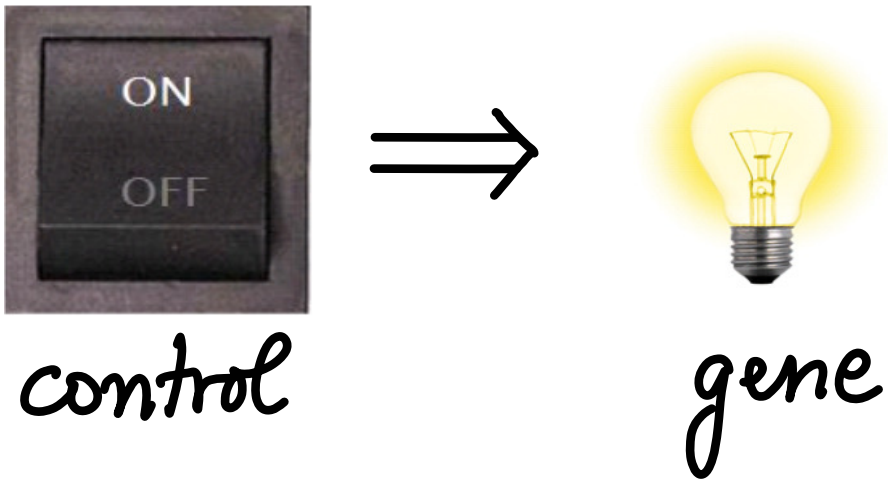
* LP strategy may not work well on other types of data
* scoring function most accurate for surface mutations
* DT-based scoring functions can predict stability and solubility mutagenesis simultaneously

Temperature-Sensitive (Ts) Mutants

* with P. Ye (Molecular Biosciences) and S. Lockwood (EECS)

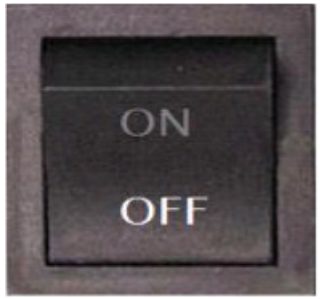
Temperature-Sensitive (Ts) Mutants

* with P. Ye (Molecular Biosciences) and S. Lockwood (EECS)
* control of gene expression



Temperature-Sensitive (Ts) Mutants

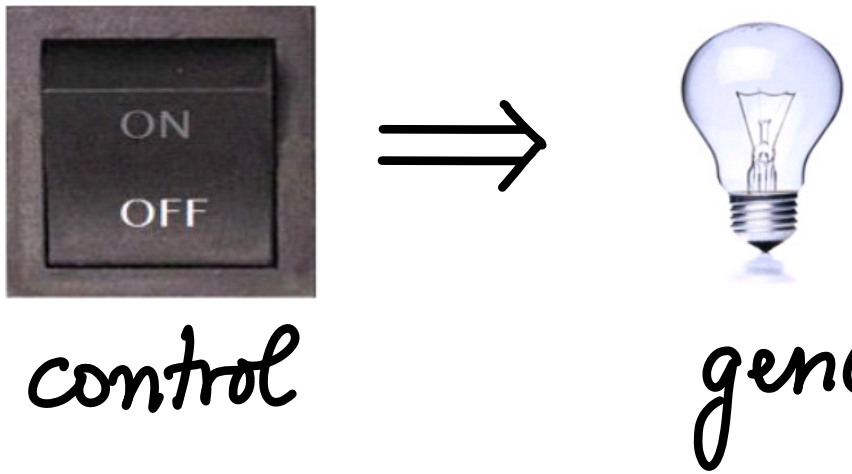
* with P. Ye (Molecular Biosciences) and S. Lockwood (EECS)
* control of gene expression



control ⟹ gene

Temperature-Sensitive (Ts) Mutants

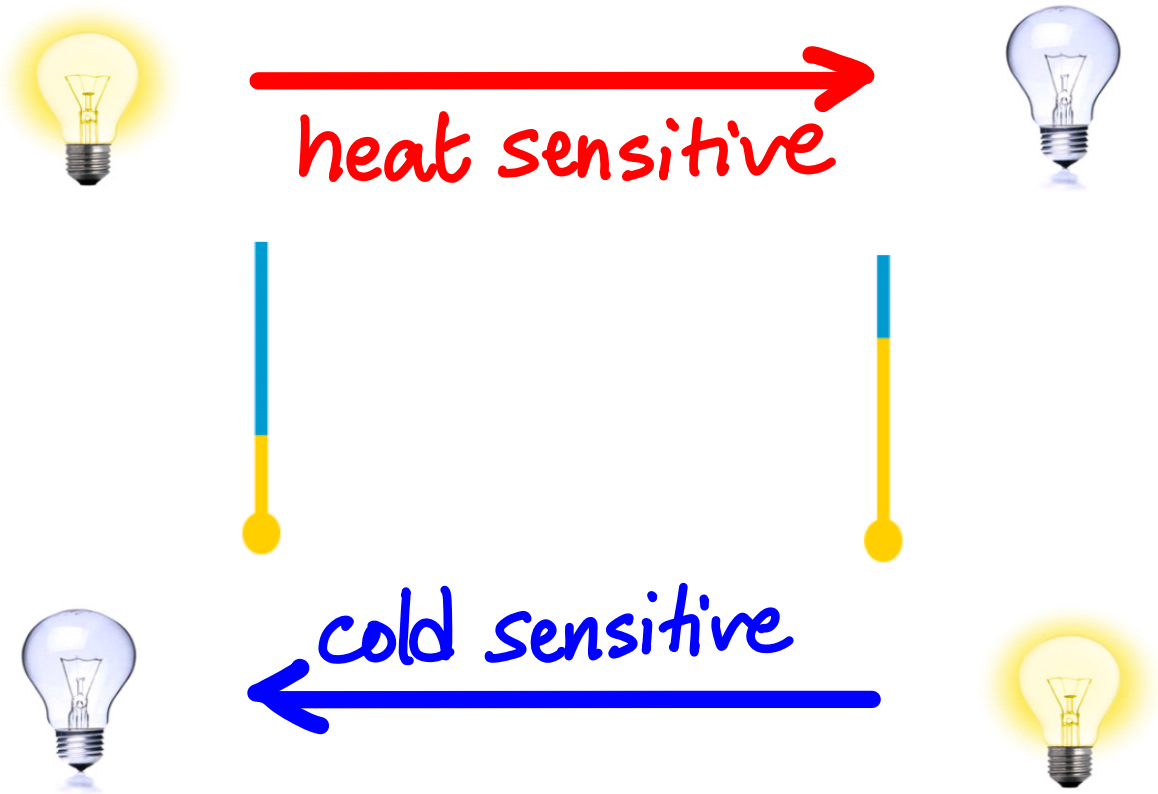
* with P. Ye (Molecular Biosciences) and S. Lockwood (EECS)
* control of gene expression



* useful for essential gene studies

Ts Mutants

Ts Mutants



Ts Mutants

Ts Mutants

* Explain mechanism of Ts mutants computationally ?

Descriptive Framework

* logistic regression $y = \frac{1}{1+e^{-z}}$ where $z \to \sum_{j=1}^{n} \beta_j x_j$
  - predict 0/1 response
  - variables need not be independent
  - can include interaction terms

Descriptive Framework

* logistic regression $y = \frac{1}{1+e^{-z}}$ where $z = \sum_{j=1}^{n} \beta_j x_j$
  - predict 0/1 response
  - variables need not be independent
  - can include interaction terms

* 10-fold CV

Descriptive Framework

* logistic regression $y = \frac{1}{1+e^{-z}}$ where $z \rightarrow \sum_{j=1}^{n} \beta_j x_j$
  - predict 0/1 response
  - variables need not be independent
  - can include interaction terms

* 10-fold CV
* dataset available : 6231 mutants
  - 747 (12%) are Ts mutants

133 Features Studied

Structure

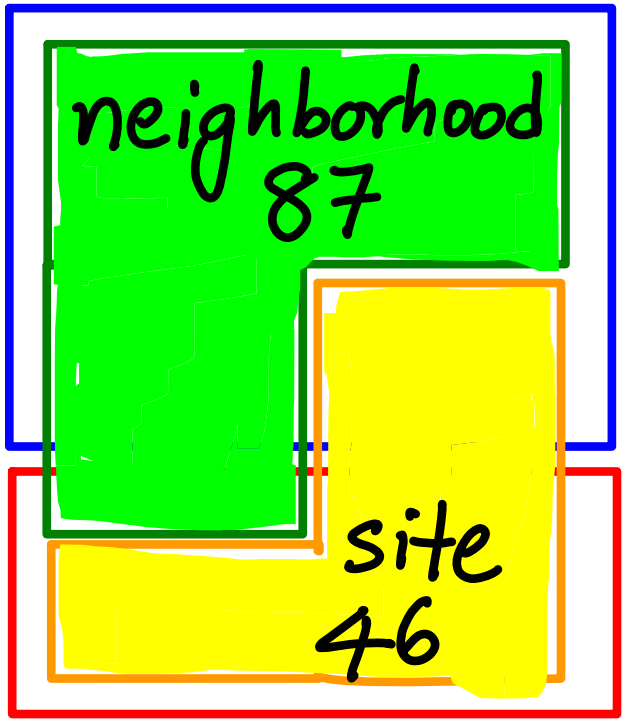
86

sequence

47

133 Features Studied

Structure

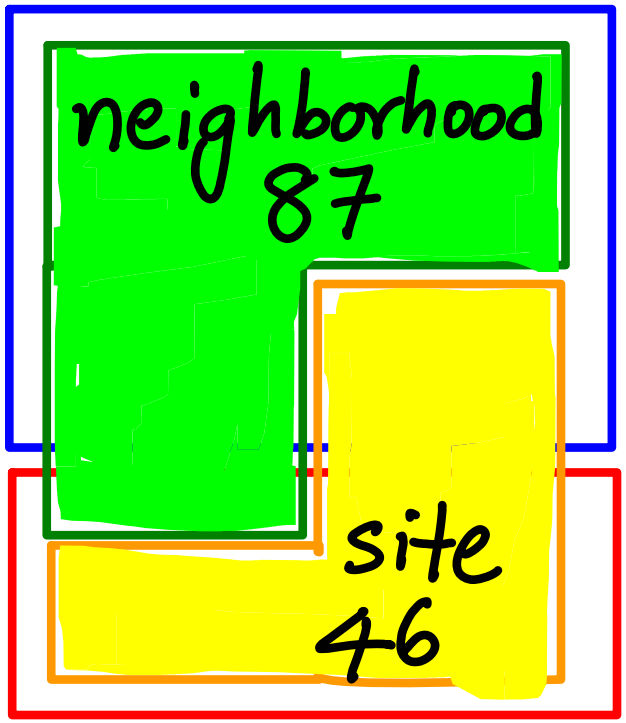


86

sequence

47

133 Features Studied



structure 86

sequence 47

* neighborhood further divided into sequence, Euclidean, and Delaunay

Mechanism of Ts Mutants

* evaluate predictive power of each feature individually

Mechanism of Ts Mutants

* evaluate predictive power of each feature individually

* most predictive features explain Ts mutation mechanism

Mechanism of Ts Mutants

* evaluate predictive power of each feature individually
* most predictive features explain Ts mutation mechanism
* Can build predictive models using subsets of features, e.g., sequence-only, site-only, neighborhood only, all, etc.

Mechanism of Ts Mutants

traditional view – amino acid & position
(i.e., site specific)

Mechanism of Ts Mutants

traditional view – amino acid & position (i.e., site specific)

✓ 11 of top 20 most predictive features are neighborhood-based

Mechanism of Ts Mutants

traditional view – amino acid & position (i.e., site specific)

✓ 11 of top 20 most predictive features are neighborhood-based

New view – neighborhood features are as (more) important as (than) site features in defining a mutation as Ts.

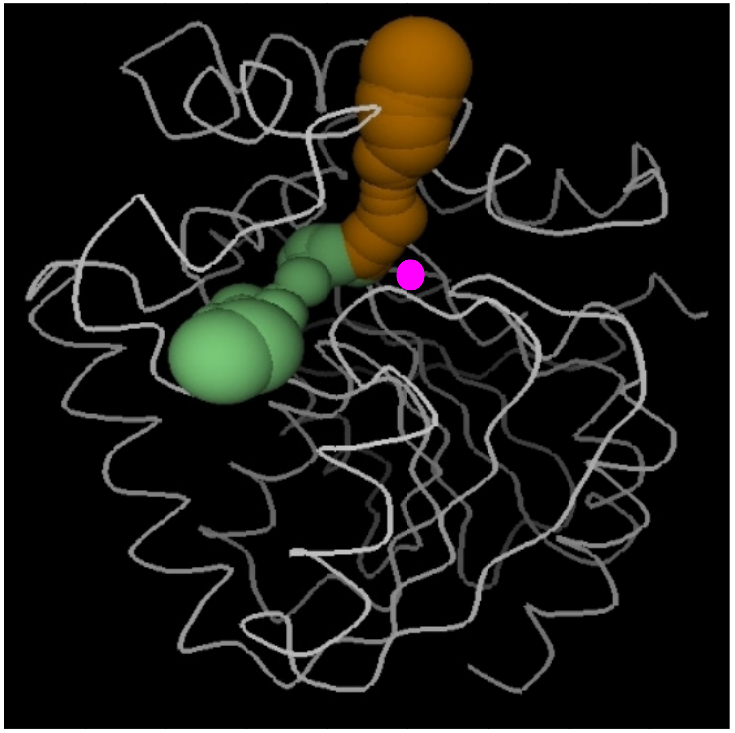
Prediction Models for Ts Mutants

Model name	ACC	MCC	AUC	KL	DD
Site features	0.78	0.39	0.87	0.17	0.48
Neighborhood features	0.82	0.46	0.91	0.10	0.68
Sequence neighborhood	0.79	0.37	0.84	0.11	0.48
Euclidean neighborhood	0.81	0.41	0.88	0.11	0.55
Delaunay neighborhood	0.78	0.39	0.86	0.15	0.52
All features	0.84	0.51	0.93	0.08	0.80
Sequence features	0.81	0.45	0.89	0.11	0.60
Structural features	0.83	0.49	0.92	0.09	0.75

Do Proteins Inspire New Mathematics?

Do Proteins Inspire New Mathematics?

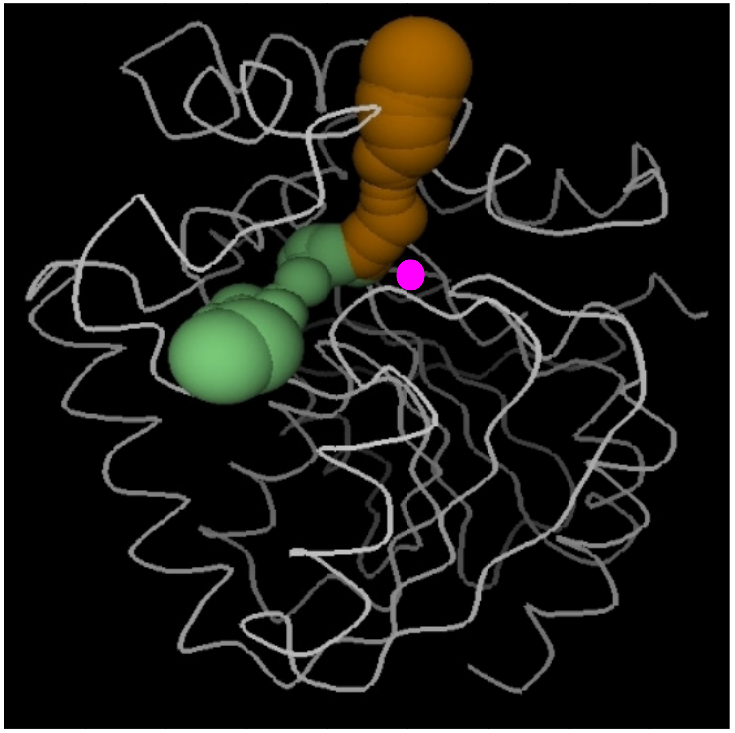
tunnels in proteins —
access to active site



(image: CAVER)

Do Proteins Inspire New Mathematics?

tunnels in proteins — access to active site



(image: CAVER)

Substrate can react with protein if the "narrowest neck" of tunnel is "big enough"

Proteins Do Inspire New Mathematics!

✓ finding bottlenecks in tunnels $\Longrightarrow$
optimal homologous chain problem (OHCP)

Proteins Do Inspire New Mathematics!

✓ finding bottlenecks in tunnels ⟹
optimal homologous chain problem (OHCP)

✓ with T. Dey (Ohio State U.) & A. Hirani (Illinois)
— results connecting concepts from
algebraic topology and matroid theory

Proteins Do Inspire New Mathematics!

- ✓ finding bottlenecks in tunnels $\Longrightarrow$ optimal homologous chain problem (OHCP)
- ✓ with T. Dey (Ohio State U.) & A. Hirani (Illinois)
  - results connecting concepts from algebraic topology and matroid theory
  - funded by NSF

Proteins Do Inspire New Mathematics!

- ✓ finding bottlenecks in tunnels ⟹ optimal homologous chain problem (OHCP)
- ✓ with T. Dey (Ohio State U.) & A. Hirani (Illinois)
  - results connecting concepts from algebraic topology and matroid theory
  - funded by NSF
- ✓ Edelsbrunner et al., 1995 – alpha shapes

Proteins Do Inspire New Mathematics!

✓ finding bottlenecks in tunnels ⟹
optimal homologous chain problem (OHCP)

✓ with T. Dey (Ohio State U.) & A. Hirani (Illinois)
— results connecting concepts from
algebraic topology and matroid theory
— funded by NSF

✓ Edelsbrunner et al., 1995 – alpha shapes

? How to handle moving proteins? ⟹
inspires more fundamental questions!