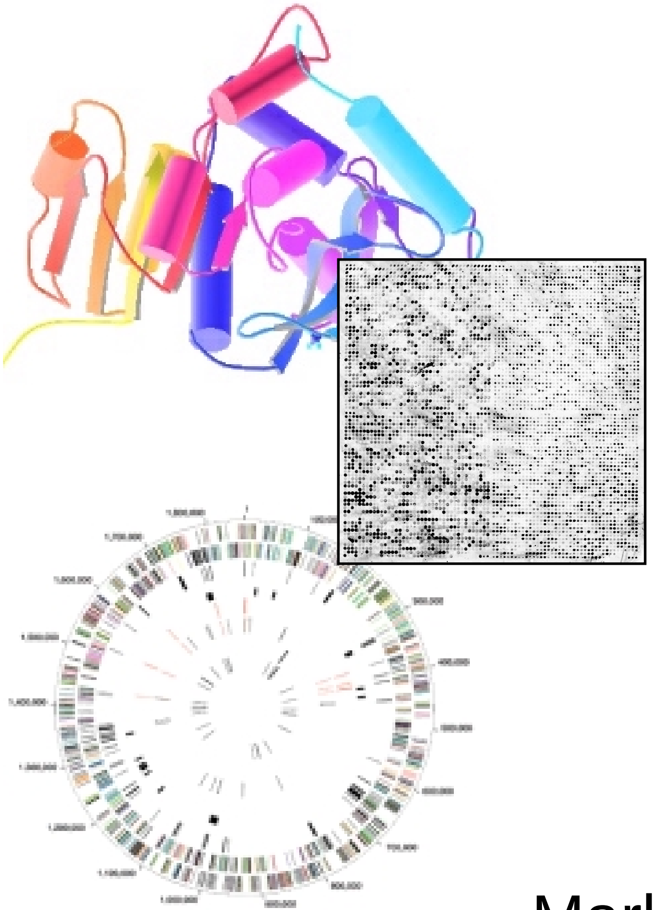
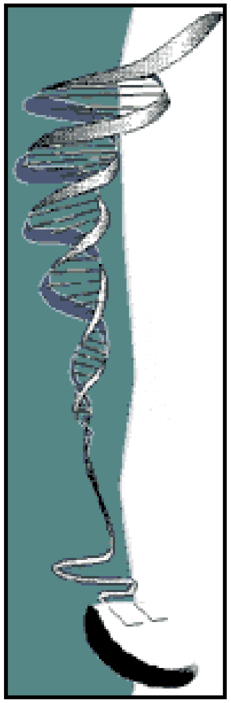
BIOINFORMATICS
Structures





Mark Gerstein, Yale University

bioinfo.mbb.yale.edu/mbb452a

Contents: Structures

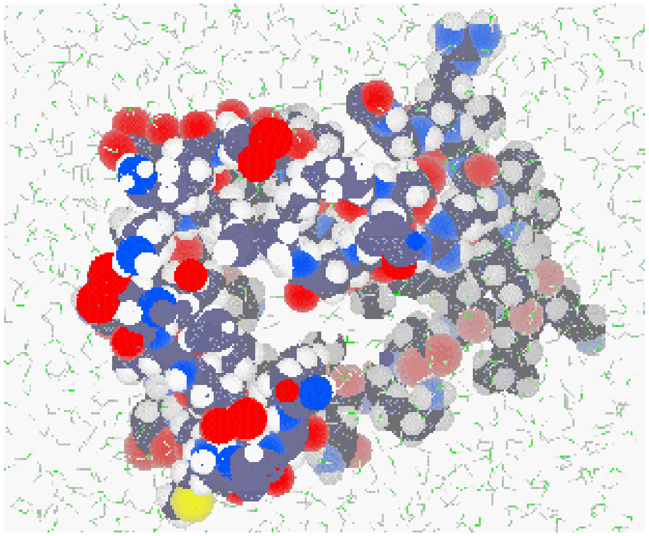
- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
    - ◊ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
    - ◊ Distance Matrix based methods
    - ◊ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity

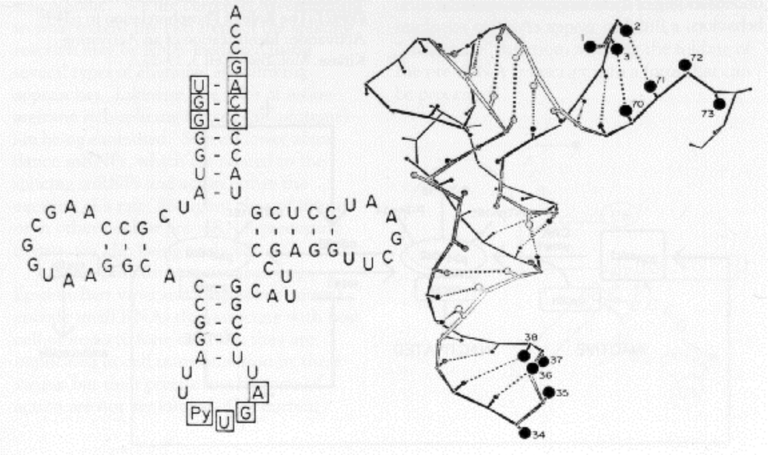
- Protein Geometry
- Surfaces I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking

Molecular Biology Information: Macromolecular Structure

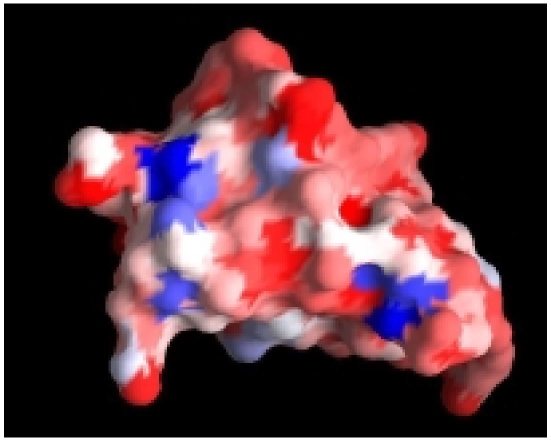
- DNA/RNA/Protein
  - ◊ Almost all protein

(RNA Adapted From D Soll Web Page,
Right Hand Top Protein from M Levitt web page)





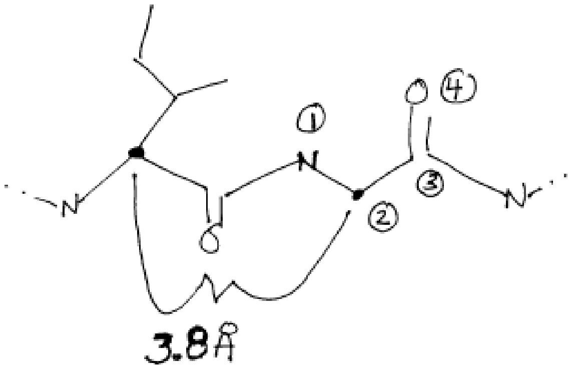
'Identity elements' in *Escherichia coli* glutamine tRNA.



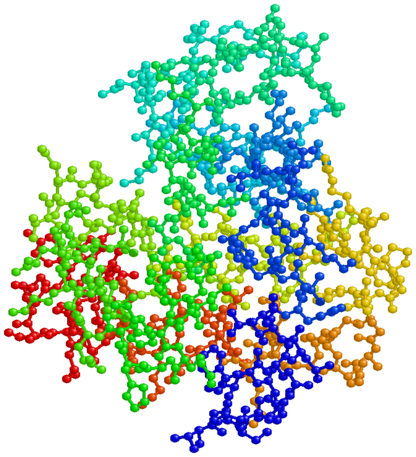
Molecular Biology Information: Protein Structure Details

- Statistics on Number of XYZ triplets
  - ◊ 200 residues/domain -> 200 CA atoms, separated by 3.8 A
  - ◊ Avg. Residue is Leu: 4 backbone atoms + 4 sidechain atoms, 150 cubic A
    - o => ~1500 xyz triplets (=8x200) per protein domain
  - ◊ 10 K known domain, ~300 folds

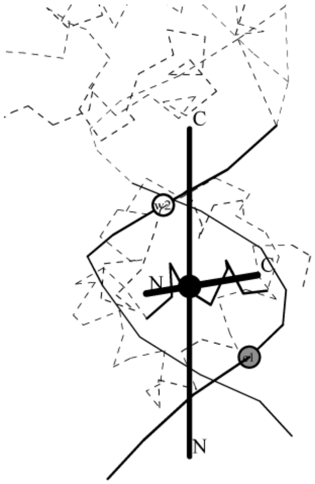
```
ATOM      1  C   ACE     0       9.401  30.166  60.595  1.00 49.88      1GKY  67
ATOM      2  O   ACE     0      10.432  30.832  60.722  1.00 50.35      1GKY  68
ATOM      3  CH3 ACE     0       8.876  29.767  59.226  1.00 50.04      1GKY  69
ATOM      4  N   SER     1       8.753  29.755  61.685  1.00 49.13      1GKY  70
ATOM      5  CA  SER     1       9.242  30.200  62.974  1.00 46.62      1GKY  71
ATOM      6  C   SER     1      10.453  29.500  63.579  1.00 41.99      1GKY  72
ATOM      7  O   SER     1      10.593  29.607  64.814  1.00 43.24      1GKY  73
ATOM      8  CB  SER     1       8.052  30.189  63.974  1.00 53.00      1GKY  74
ATOM      9  OG  SER     1       7.294  31.409  63.930  1.00 57.79      1GKY  75
ATOM     10  N   ARG     2      11.360  28.819  62.827  1.00 36.48      1GKY  76
ATOM     11  CA  ARG     2      12.548  28.316  63.532  1.00 30.20      1GKY  77
ATOM     12  C   ARG     2      13.502  29.501  63.500  1.00 25.54      1GKY  78
. . .
ATOM   1444  CB  LYS   186      13.836  22.263  57.567  1.00 55.06      1GKY1510
ATOM   1445  CG  LYS   186      12.422  22.452  58.180  1.00 53.45      1GKY1511
ATOM   1446  CD  LYS   186      11.531  21.198  58.185  1.00 49.88      1GKY1512
ATOM   1447  CE  LYS   186      11.452  20.402  56.860  1.00 48.15      1GKY1513
ATOM   1448  NZ  LYS   186      10.735  21.104  55.811  1.00 48.41      1GKY1514
ATOM   1449  OXT LYS   186      16.887  23.841  56.647  1.00 62.94      1GKY1515
TER    1450      LYS   186                                              1GKY1516
```



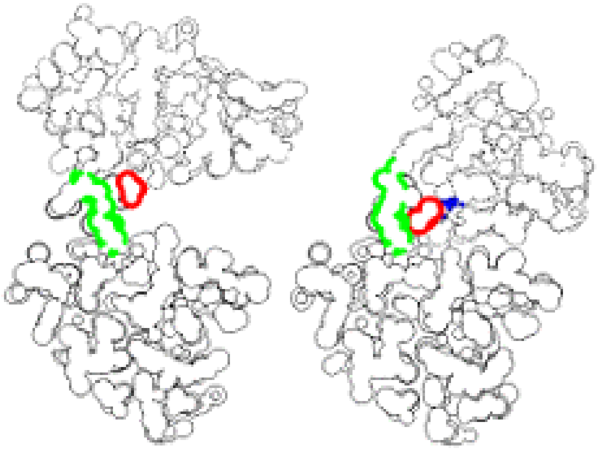
Other Aspects of Structure, Besides just Comparing Atom Positions



Atom Position, XYZ triplets



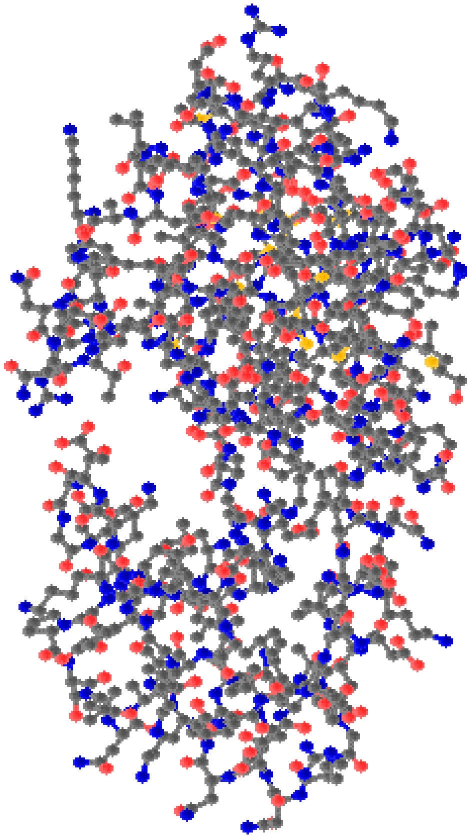
Lines, Axes, Angles



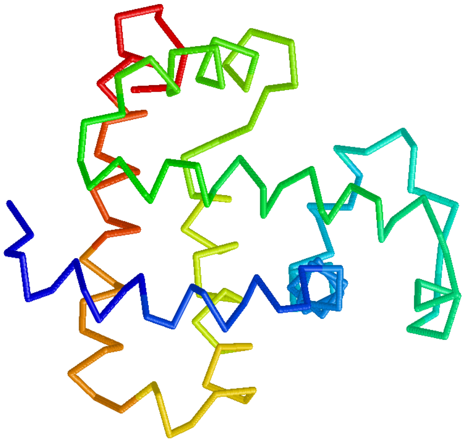
Surfaces, Volumes

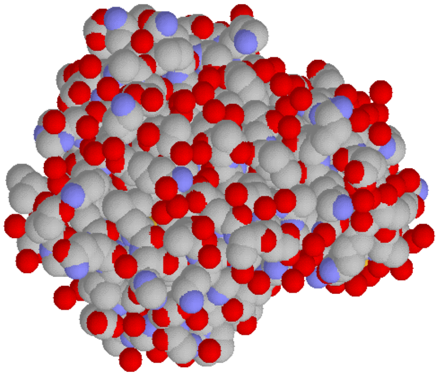
What is Protein Geometry?

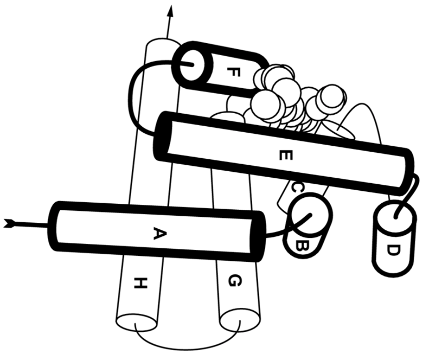
- Coordinates (X, Y, Z’s)
- Derivative Concepts
  - ◊ Distance, Surface Area, Volume, Cavity, Groove, Axes, Angle, &c
- Relation to
  - ◊ Function, Energies (E(x)), Dynamics (dx/dt)

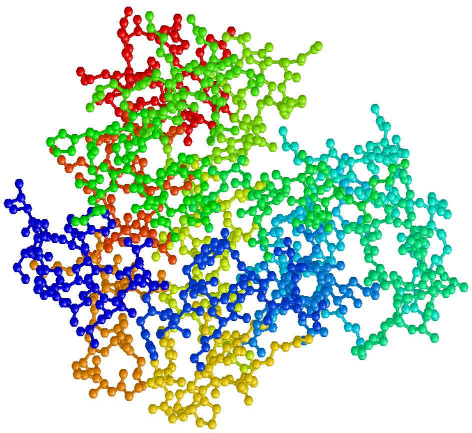


Depicting Protein Structure: Sperm Whale Myoglobin

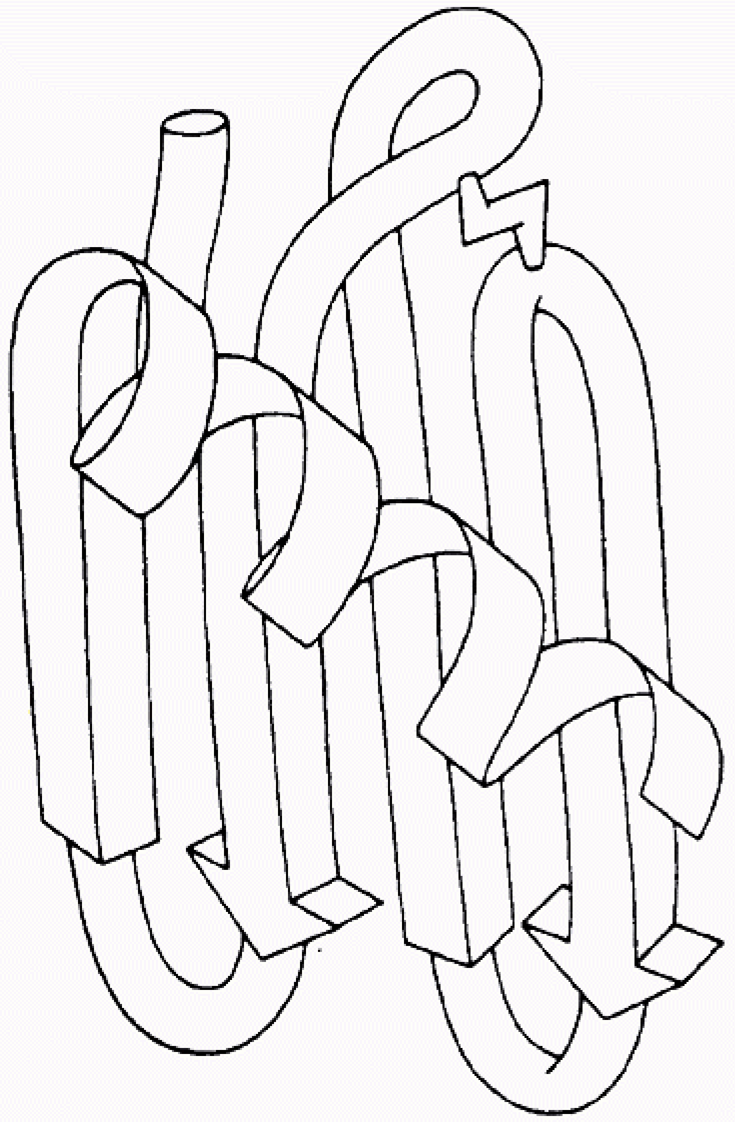








Incredulase



Incredulase

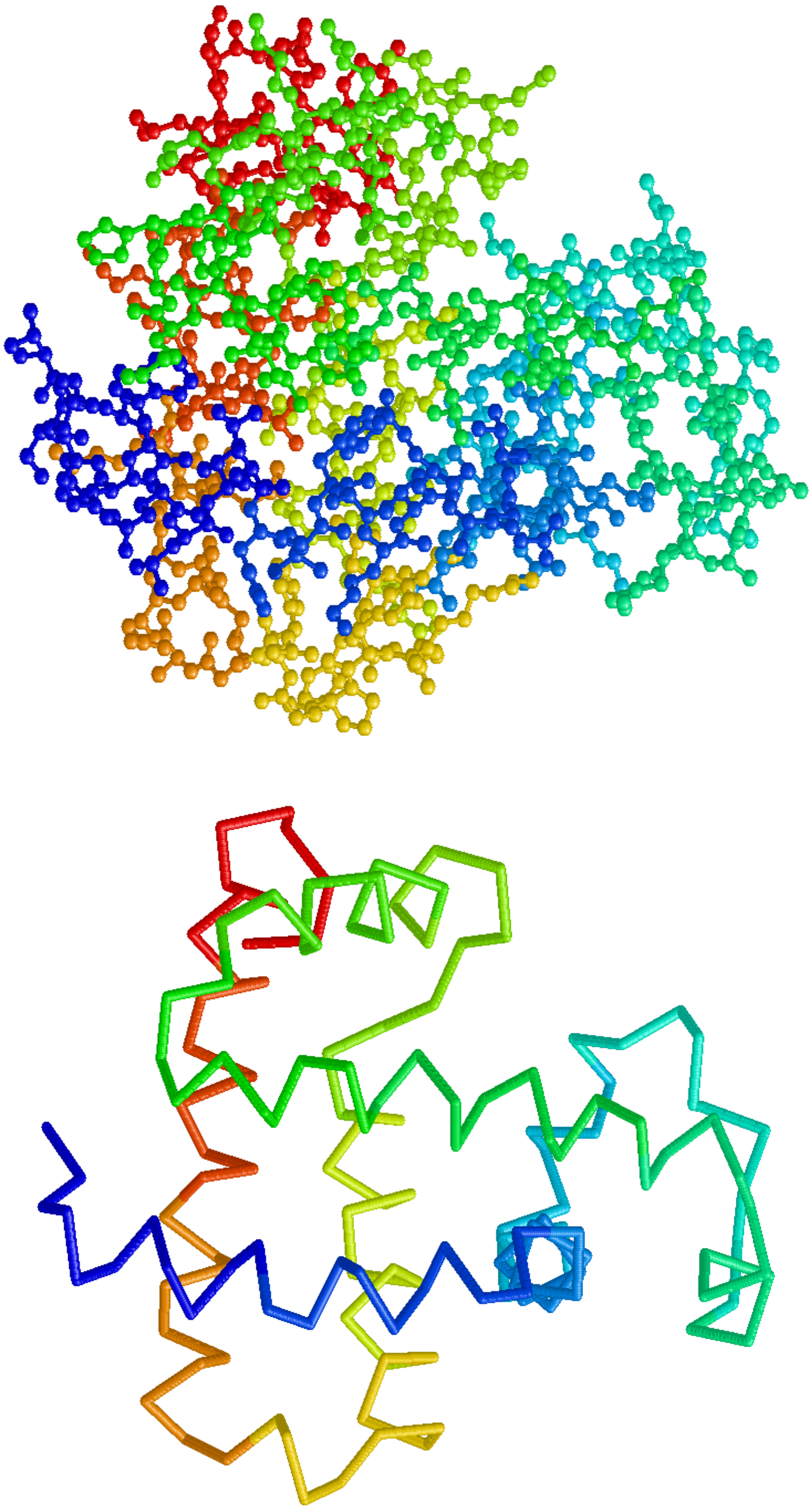
J.S. Richardson and D.C. Richardson, "*Some design principles: Betabellin*", in D.L. Oxender and C.F. Fox (Eds.), "*Protein Engineering*", Alan R. Liss, 1987, p. 149-163

Structure alignment - Method

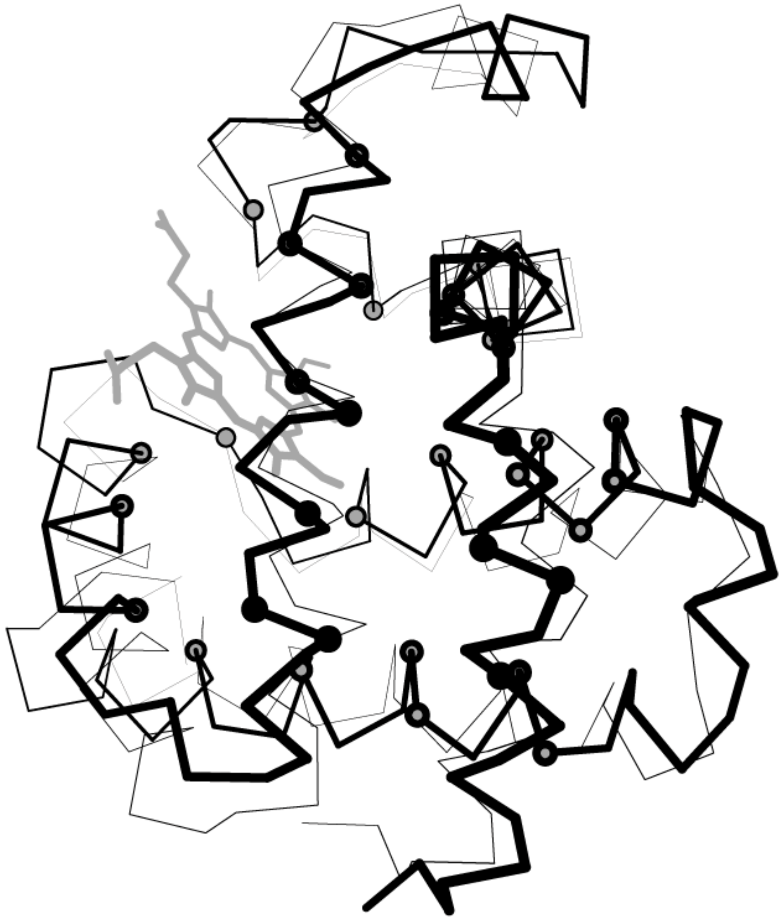
- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
    - ◊ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
    - ◊ Distance Matrix based methods
    - ◊ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity

- Protein Geometry
- Surface I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking

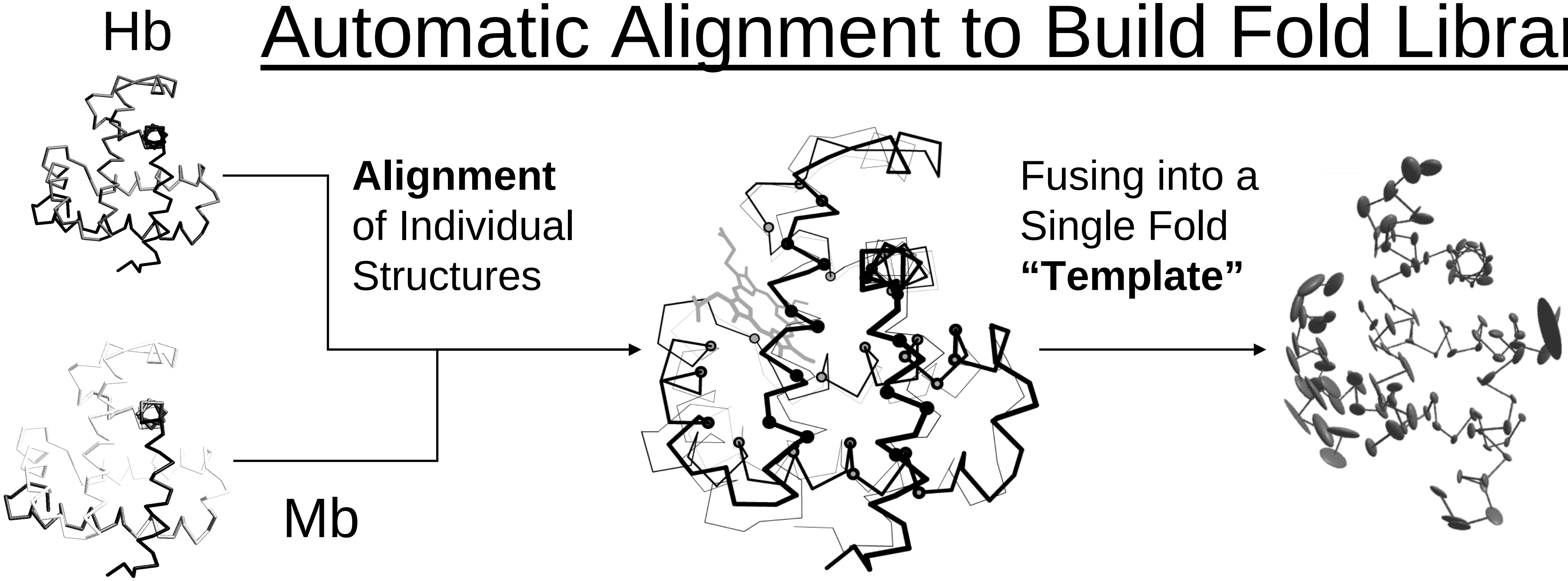
Sperm Whale Myoglobin



Structural Alignment of Two Globins



Automatic Alignment to Build Fold Library



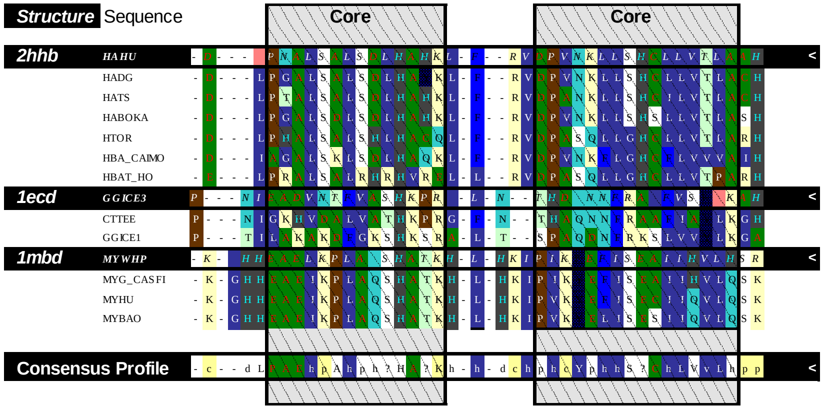
Alignment of Individual Structures

Fusing into a Single Fold **“Template”**

```
Hb VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSAQVKGHGKKVADALTNAV
      ||| ..   |   |.|| |  .  | .  | | | |   |  |  |       .|  .| ||  |  ||   .
Mb VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDLKKHGVTVLTALGAIL

Hb AHVD-DMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
             |  | . || | .. .     .|  ..  |    |..|     . . | |      .  ||.
Mb KK-KGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
```

Elements: <u>Domain</u> definitions; <u>Aligned</u> structures, collecting together <u>Non-homologous Sequences</u>; <u>Core</u> annotation

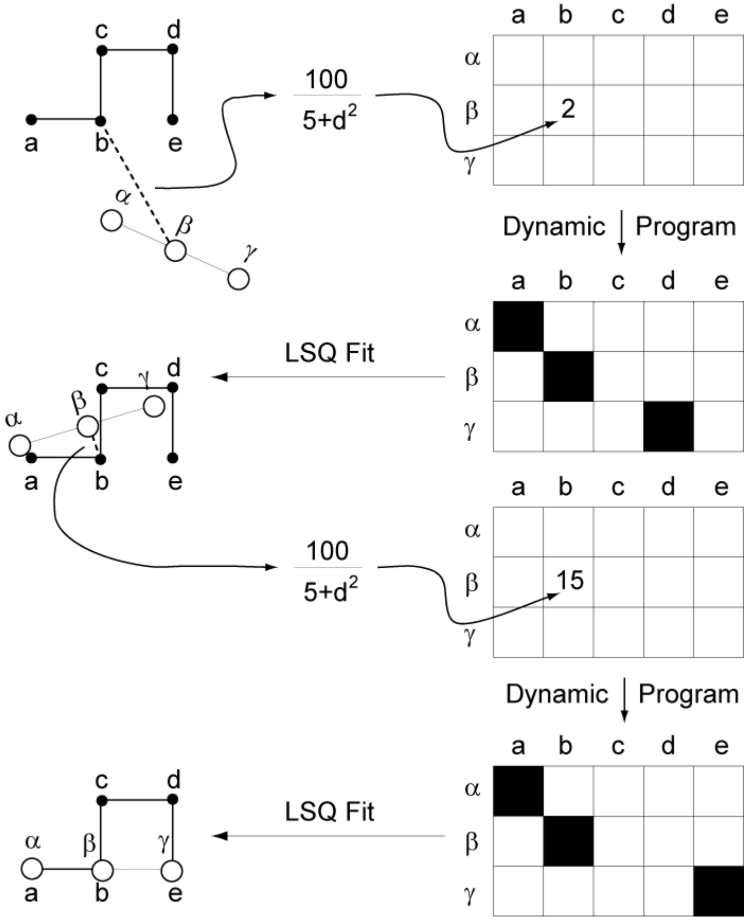


Previous work: Remington, Matthews ‘80; **Taylor, Orengo ‘89**, ‘94; Artymiuk, Rice, Willett ‘89; Sali, Blundell, ‘90; Vriend, Sander ‘91; Russell, Barton ‘92; **Holm, Sander ‘93**; Godzik, Skolnick ‘94; Gibrat, Madej, Bryant ‘96; Falicov, F Cohen, ‘96; Feng, Sippl ‘96; G Cohen ‘97; Singh & Brutlag, ‘98

Automatically Comparing Protein Structures

- Given
  2 Structures (A & B),
  2 Basic
  Comparison Operations
  1. Given an alignment optimally **SUPERIMPOSE** A onto B

     Find Best R & T to move A onto B
  2. **Find an Alignment** between A and B based on their 3D coordinates



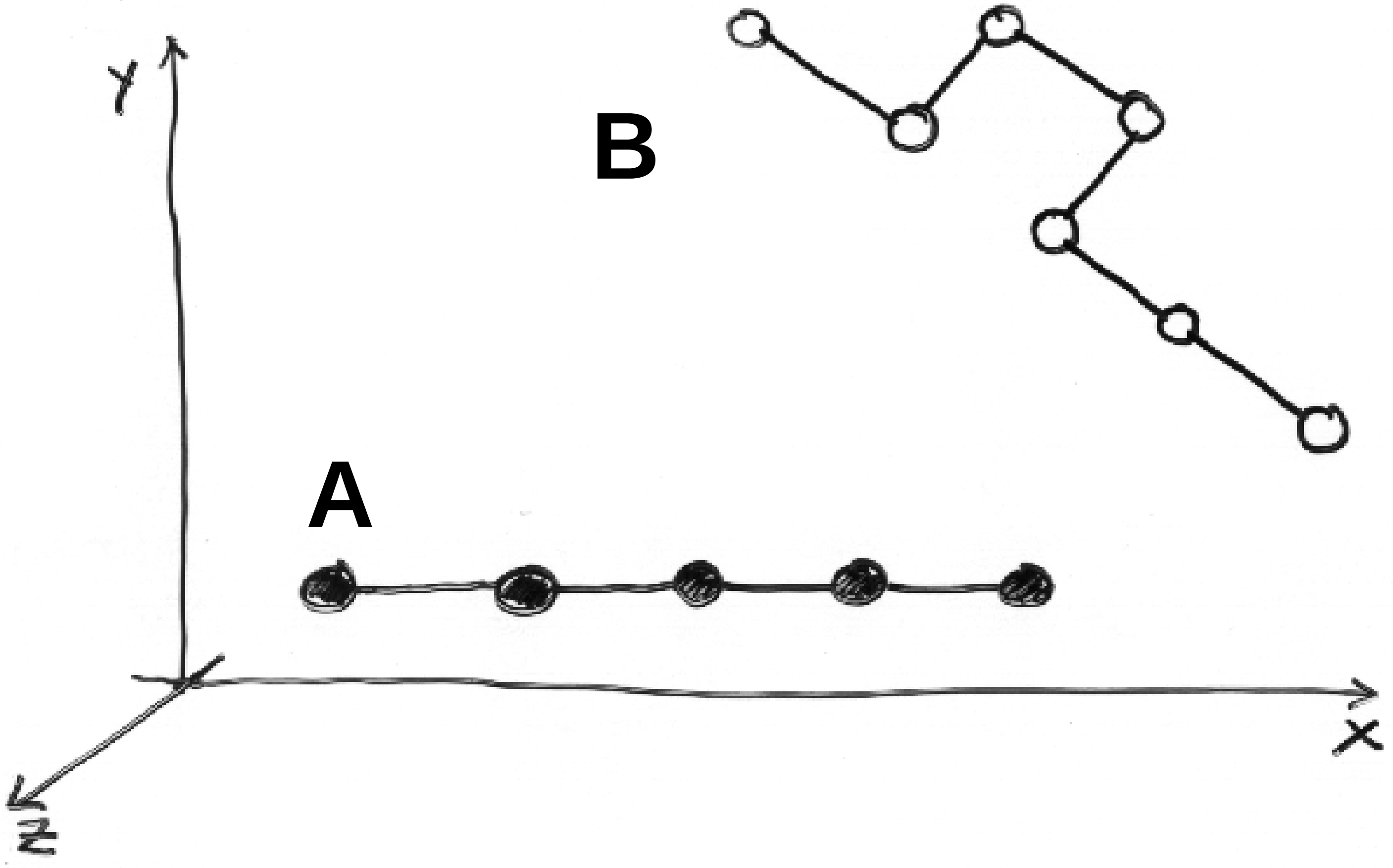
	a	b	c	d	e
α					
β		2			
γ					

	a	b	c	d	e
α	■				
β		■			
γ				■	

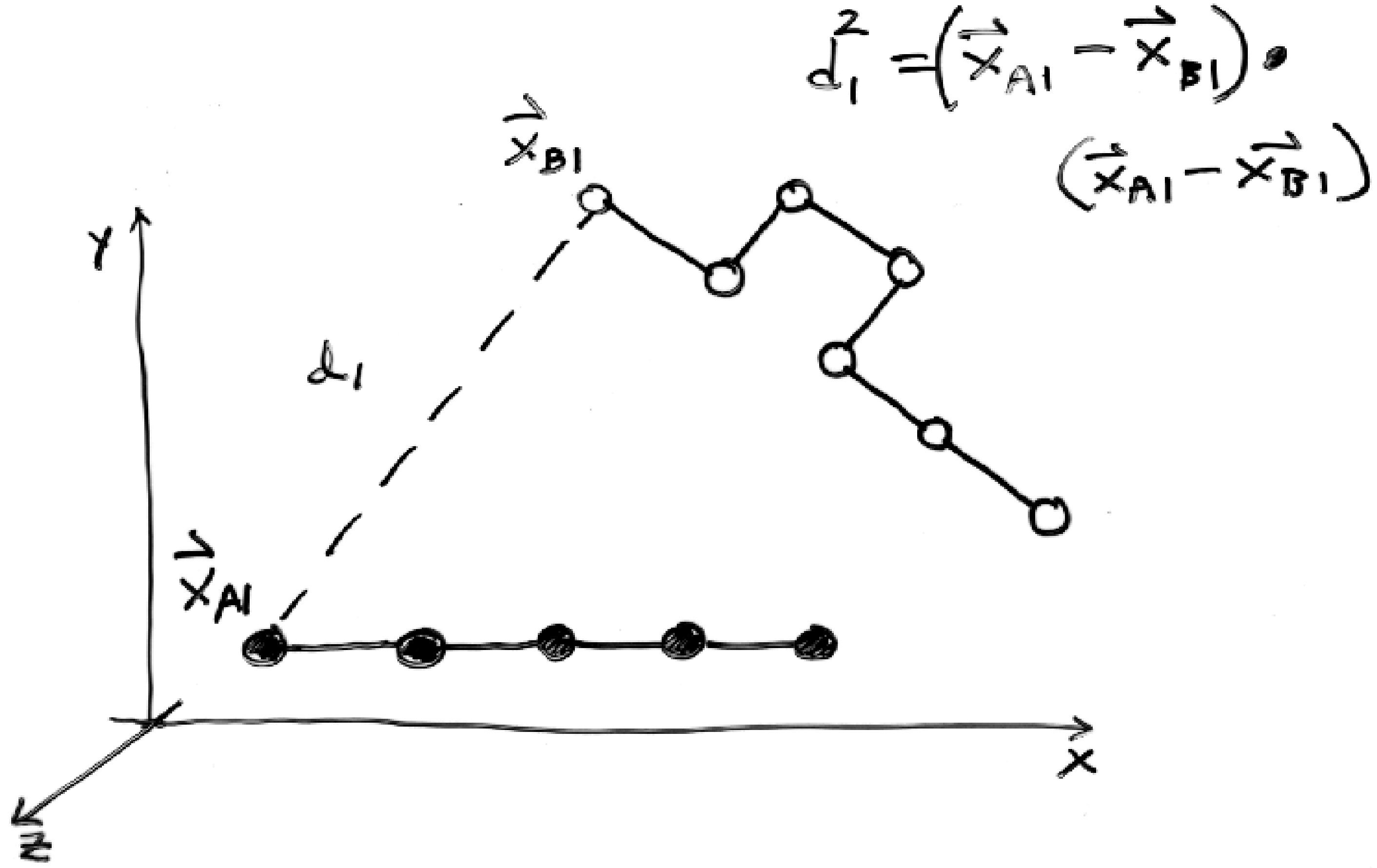
	a	b	c	d	e
α					
β		15			
γ					

	a	b	c	d	e
α	■				
β		■			
γ					■

RMS Superposition (1)



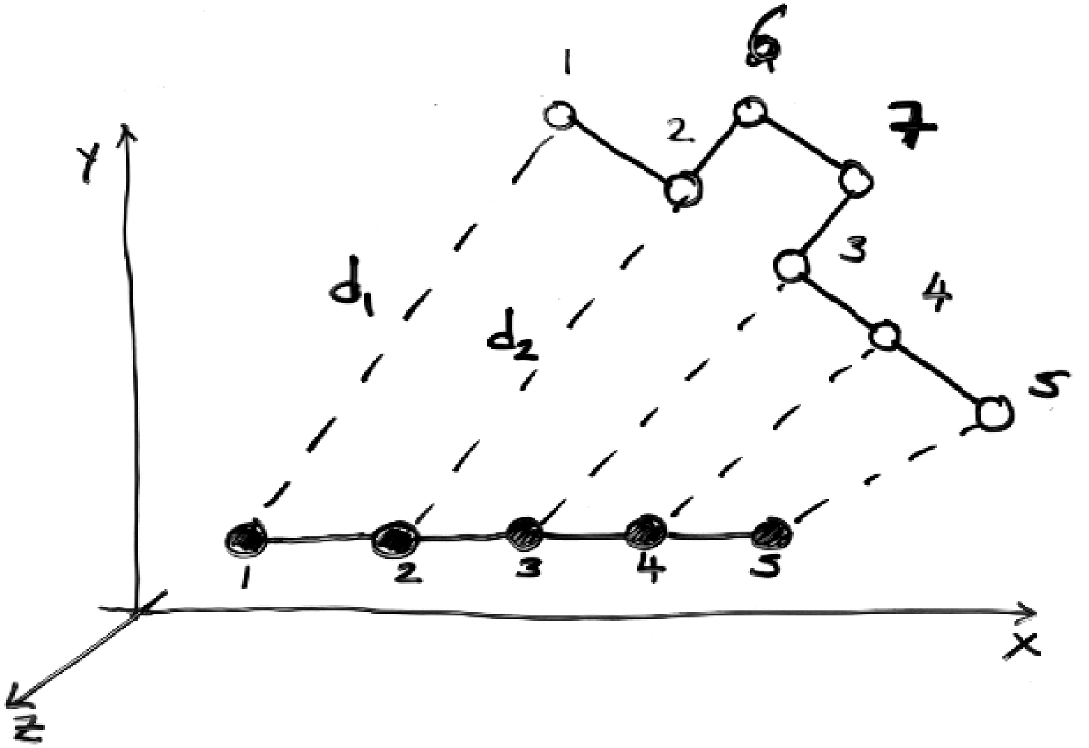
RMS Superposition (2): Distance Between an Atom in 2 Structures



$$d_1^2 = (\vec{x}_{A1} - \vec{x}_{B1}) \bullet (\vec{x}_{A1} - \vec{x}_{B1})$$

RMS Superposition (3): RMS Distance Between Aligned Atoms in 2 Structures

$$RMS = \sqrt{\frac{\sum_{i=1}^{5} (\vec{x}_{Ai} - \vec{x}_{Bi})^2}{5}} \sim \frac{d_1 + d_2 + d_3 + d_4 + d_5}{5}$$



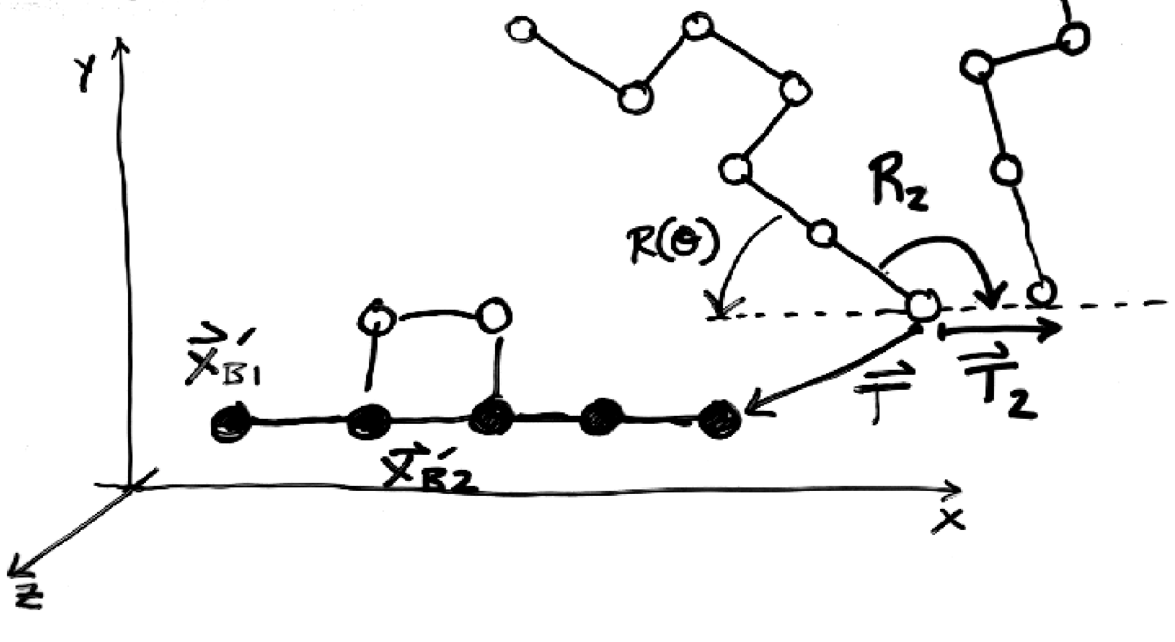
RMS Superposition (4): Rigid-Body Rotation and Translation of One Structure (B)

$$\vec{x}'_{Bi} = R(\theta)\vec{x}_{Bi} + \vec{T}$$

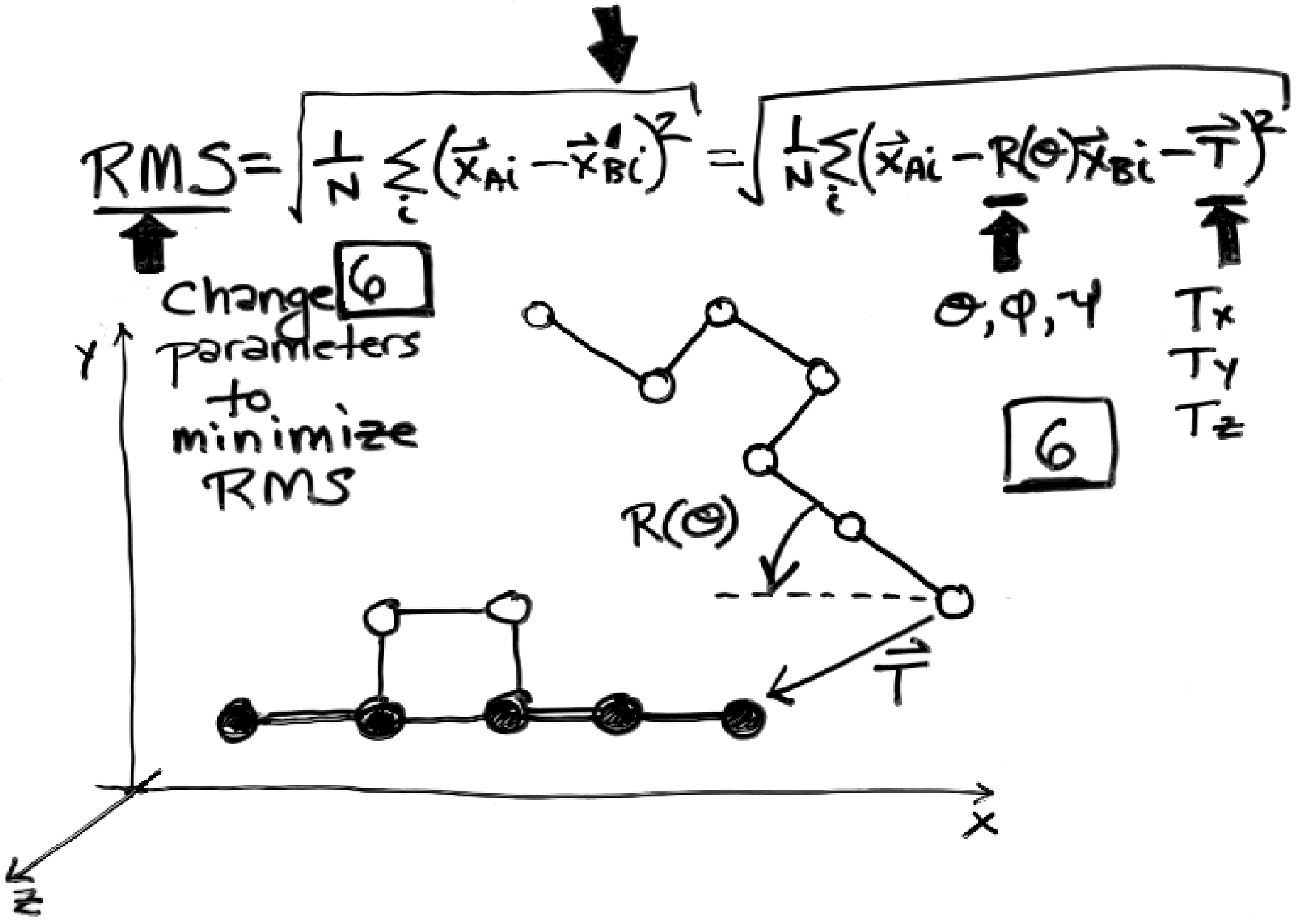
ROTATE & TRANSLATE

$\boxed{6}$ parameters

$\vec{T} = (T_x\ T_y\ T_z)$ $R(\theta, \varphi, \psi)$



RMS Superposition (5): Optimal Movement of One Structure to Minimize the RMS



$$RMS = \sqrt{\frac{1}{N}\sum_i \left(\vec{x}_{Ai} - \vec{x}_{Bi}\right)^2} = \sqrt{\frac{1}{N}\sum_i \left(\vec{x}_{Ai} - R(\theta)\vec{x}_{Bi} - \vec{T}\right)^2}$$

Methods of Solution:

springs (F ~ kx)

SVD

Kabsch

Alignment (1) Make a Similarity Matrix (Like Dot Plot)

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
B		1											
P												1	

Structural Alignment (1b)
Make a Similarity Matrix
(Generalized Similarity Matrix)

- PAM(A,V) = 0.5
  - ◊ Applies at every position
- S(aa @ i, aa @ J)
  - ◊ Specific Matrix for each pair of residues
    i in protein 1 and
    J in protein 2
  - ◊ Example is Y near N-term. matches any C-term. residue (Y at J=2)
- S(i,J)
  - ◊ Doesn't need to depend on a.a. identities at all!
  - ◊ Just need to make up a score for matching residue i in protein 1 with residue J in protein 2

i →

		1	2	3	4	5	6	7	8	9	10	11	12	13
		A	B	C	N	Y	R	Q	C	L	C	R	P	M
1	A	**1**												
2	**Y**					**1**			**5**	**5**	**5**	**5**	**5**	**5**
3	C			**1**					**1**		**1**			
4	Y					**1**								
5	N				**1**									
6	R						**1**					**1**		
7	C			**1**					**1**		**1**			
8	K													
9	C			**1**					**1**		**1**			
10	R						**1**					**1**		
11	B		**1**											
12	P												**1**	

J ↓

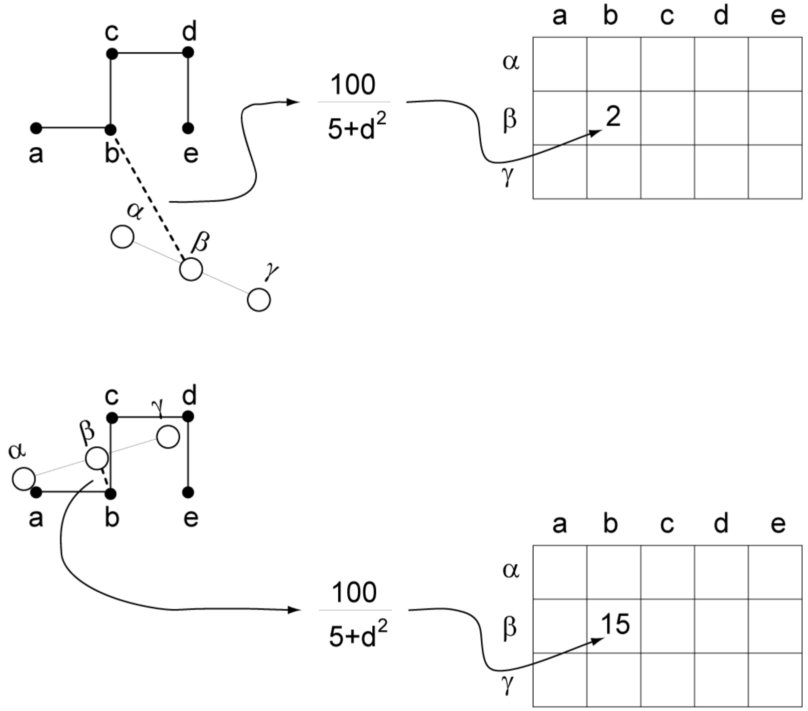
Structural Alignment (1c*) Similarity Matrix for Structural Alignment

- Structural Alignment
  - ◊ Similarity Matrix S(i,J) depends on the 3D coordinates of residues i and J
  - ◊ Distance between CA of i and J

$$d = \sqrt{(x_i - x_J)^2 + (y_i - y_J)^2 + (z_i - z_J)^2}$$

  - ◊ M(i,j) = 100 / (5 + d^2)

- Threading
  - ◊ S(i,J) depends on the how well the amino acid at position i in protein 1 fits into the 3D structural environment at position J of protein 2



Alignment (2): Dynamic Programming, Start Computing the Sum Matrix

```
new_value_cell(R,C) <=
  cell(R,C)                      { Old value, either 1 or 0    }
  + Max[
        cell (R+1, C+1),         { Diagonally Down, no gaps    }
        cells(R+1, C+2 to C_max),{ Down a row, making col. gap }
        cells(R+2 to R_max, C+2) { Down a col., making row gap }
       ]
```

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
B		1											
P												1	

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					*2*	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Alignment (3):Dynamic Programming, Keep Going

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					**1**		
C			1					1		1			
K													
C			1					1		1			
R						1					***2***	**0**	**0**
B	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
P	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						***5***	**4**	**3**	**3**	**2**	**2**	**0**	**0**
C	**3**	**3**	**4**	**3**	**3**	**3**	**3**	**4**	**3**	**3**	**1**	**0**	**0**
K	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**2**	**1**	**0**	**0**
C	**2**	**2**	**3**	**2**	**2**	**2**	**2**	**3**	**2**	**3**	**1**	**0**	**0**
R	**2**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**2**	**0**	**0**
B	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
P	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**

Alignment (4): Dynamic Programming, Sum Matrix All Done

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						***5***	**4**	**3**	**3**	**2**	**2**	**0**	**0**
C	**3**	**3**	**4**	**3**	**3**	**3**	**3**	**4**	**3**	**3**	**1**	**0**	**0**
K	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**2**	**1**	**0**	**0**
C	**2**	**2**	**3**	**2**	**2**	**2**	**2**	**3**	**2**	**3**	**1**	**0**	**0**
R	**2**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**2**	**0**	**0**
B	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
P	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	***8***	**7**	**6**	**6**	**5**	**4**	**4**	**3**	**3**	**2**	**1**	**0**	**0**
Y	**7**	**7**	**6**	**6**	**6**	**4**	**4**	**3**	**3**	**2**	**1**	**0**	**0**
C	**6**	**6**	**7**	**6**	**5**	**4**	**4**	**4**	**3**	**3**	**1**	**0**	**0**
Y	**6**	**6**	**6**	**5**	**6**	**4**	**4**	**3**	**3**	**2**	**1**	**0**	**0**
N	**5**	**5**	**5**	**6**	**5**	**4**	**4**	**3**	**3**	**2**	**1**	**0**	**0**
R	**4**	**4**	**4**	**4**	**4**	**5**	**4**	**3**	**3**	**2**	**2**	**0**	**0**
C	**3**	**3**	**4**	**3**	**3**	**3**	**3**	**4**	**3**	**3**	**1**	**0**	**0**
K	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**2**	**1**	**0**	**0**
C	**2**	**2**	**3**	**2**	**2**	**2**	**2**	**3**	**2**	**3**	**1**	**0**	**0**
R	**2**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**2**	**0**	**0**
B	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**
P	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**

(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Alignment (5): Traceback

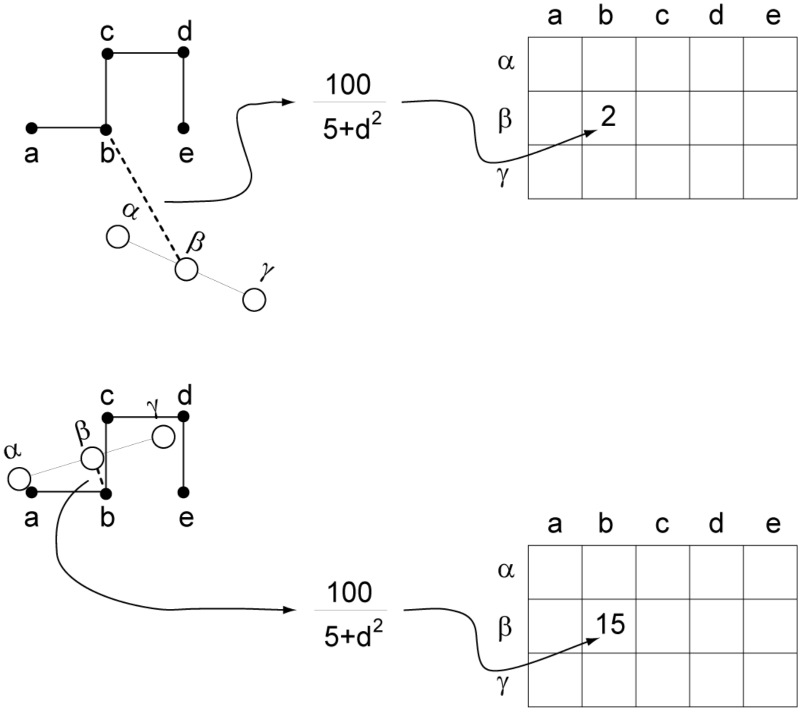
Find Best Score (8) and Trace Back

```
A B C N Y - R Q C L C R - P M
A Y C - Y N R - C K C R B P
```

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

In Structural Alignment, Not Yet Done (Step 6*)

- Use Alignment to LSQ Fit Structure B onto Structure A
  - ◊ However, movement of B will now change the Similarity Matrix
- This Violates Fundamental Premise of Dynamic Programming
  - ◊ Way Residue at i is aligned can now affect previously optimal alignment of residues (from 1 to i-1)



ACSQRP--LRV-SH -R S**E**NCV

A-SNKPQLVKLMTH VK **D**FCV-

Structural Alignment (7*), Iterate Until Convergence

1 Compute Sim. Matrix

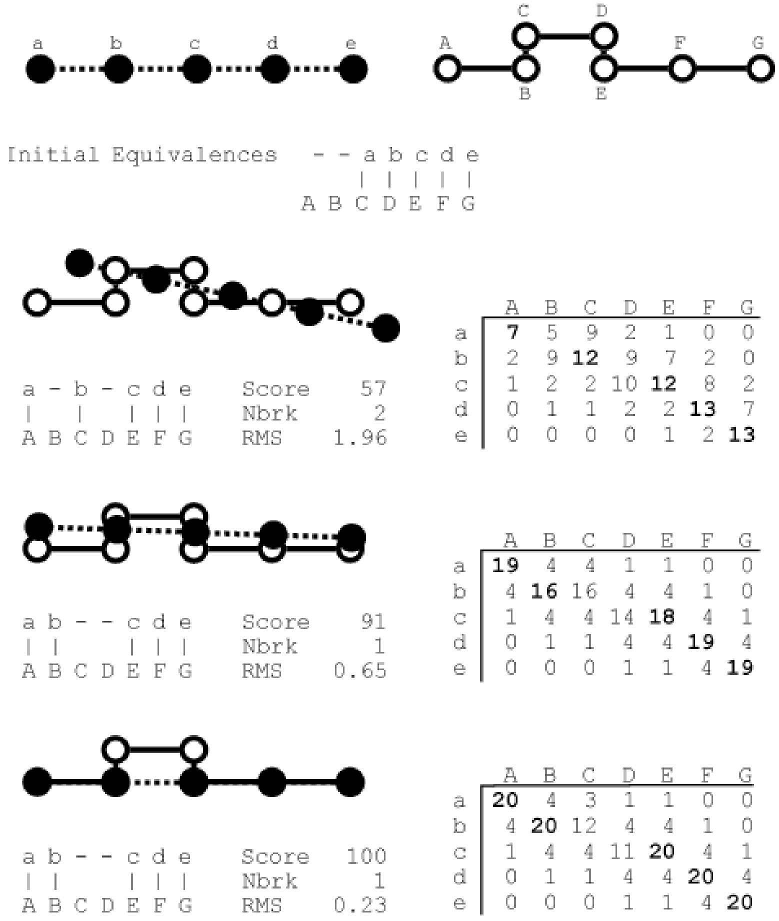
2 Align via Dyn. Prog.

3 RMS Fit Based on Alignment

4 Move Structure B

5 Re-compute Sim. Matrix

6 If changed from #1, GOTO #2



```
Initial Equivalences   - - a b c d e
                           | | | | |
                       A B C D E F G
```

```
a - b - c d e    Score    57
|   |   | | |    Nbrk      2
A B C D E F G    RMS    1.96
```

	A	B	C	D	E	F	G
a	**7**	5	9	2	1	0	0
b	2	9	**12**	9	7	2	0
c	1	2	2	10	**12**	8	2
d	0	1	1	2	2	**13**	7
e	0	0	0	0	1	2	**13**

```
a b - - c d e    Score    91
| |     | | |    Nbrk      1
A B C D E F G    RMS    0.65
```

	A	B	C	D	E	F	G
a	**19**	4	4	1	1	0	0
b	4	**16**	16	4	4	1	0
c	1	4	4	14	**18**	4	1
d	0	1	1	4	4	**19**	4
e	0	0	0	1	1	4	**19**

```
a b - - c d e    Score   100
| |     | | |    Nbrk      1
A B C D E F G    RMS    0.23
```

	A	B	C	D	E	F	G
a	**20**	4	3	1	1	0	0
b	4	**20**	12	4	4	1	0
c	1	4	4	11	**20**	4	1
d	0	1	1	4	4	**20**	4
e	0	0	0	1	1	4	**20**

Structure alignment - Scoring

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  - ◊ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  - ◊ Distance Matrix based methods
  - ◊ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity

- Protein Geometry
- Surface I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking

Score S at End Just Like SW Score, but also have final RMS

S = Total Score

S(i,j) = similarity matrix score for aligning i and j

Sum is carried out over all aligned i and j

n = number of gaps (assuming no gap ext. penalty)

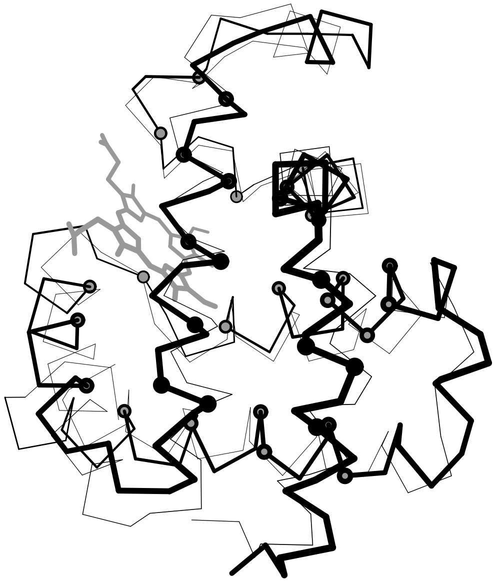
G = gap penalty

$$S = \sum_{i,j} S(i,j) - nG$$

Some Similarities are Readily Apparent others are more Subtle

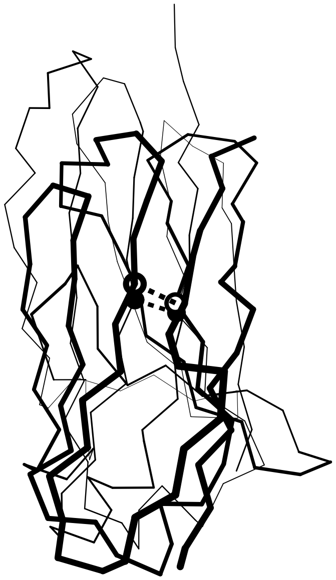
Easy:
Globins

125 res.,
~1.5 Å

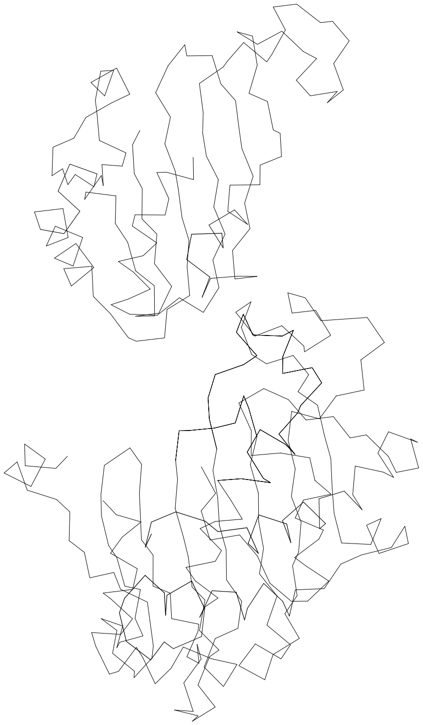


Tricky:
Ig C & V

85 res.,
~3 Å



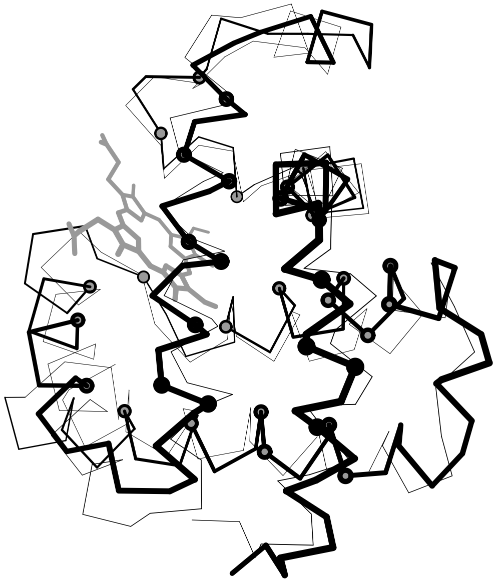
Very Subtle: G3P-dehydrogenase, C-term. Domain
>5 Å



Some Similarities are Readily Apparent others are more Subtle

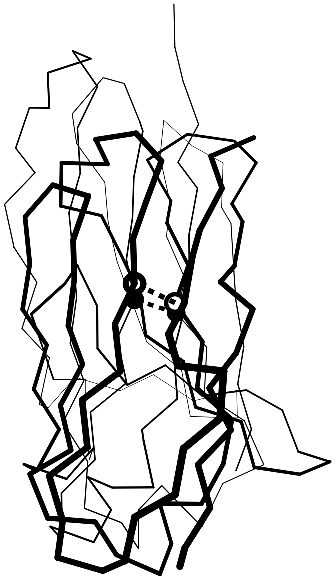
Easy:
Globins

125 res.,
~1.5 Å

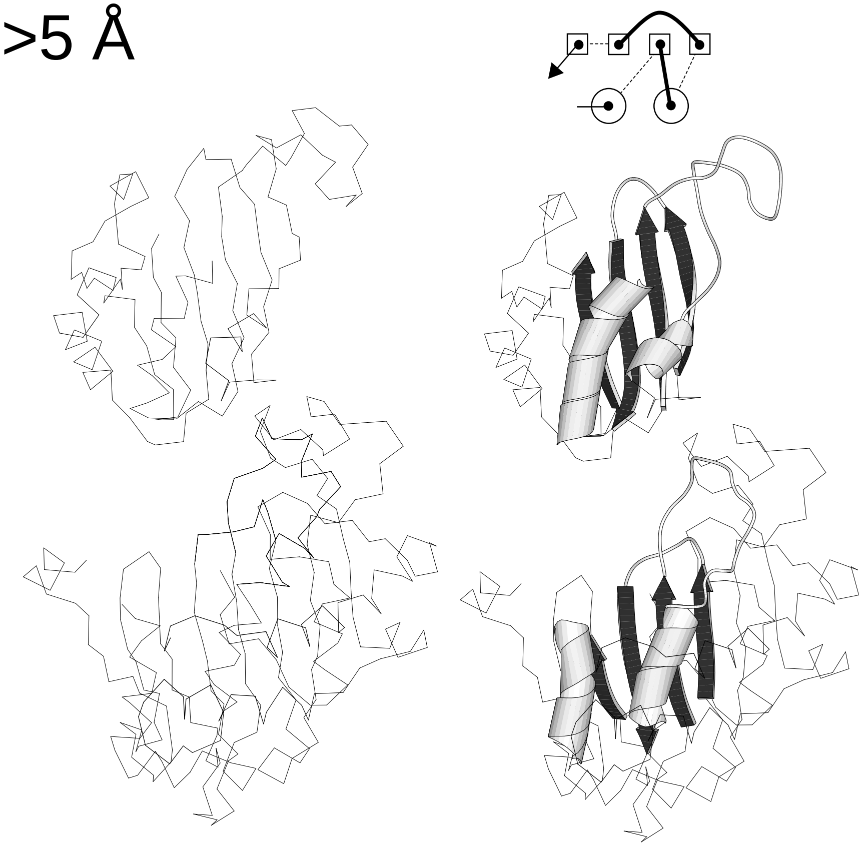


Tricky:
Ig C & V

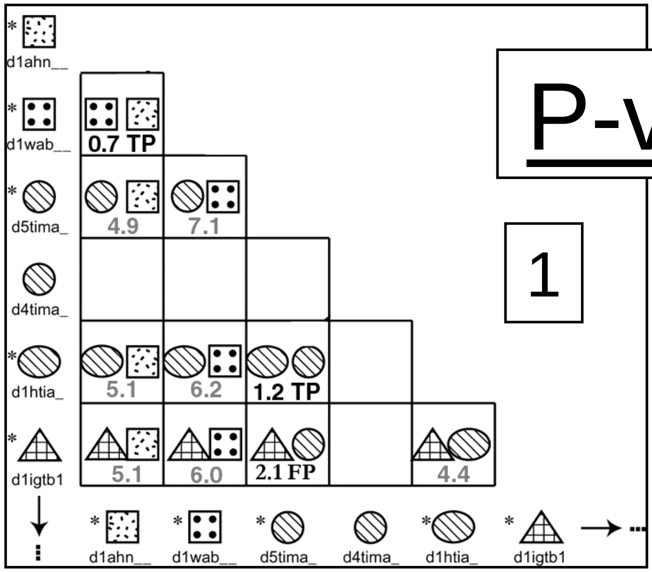
85 res.,
~3 Å



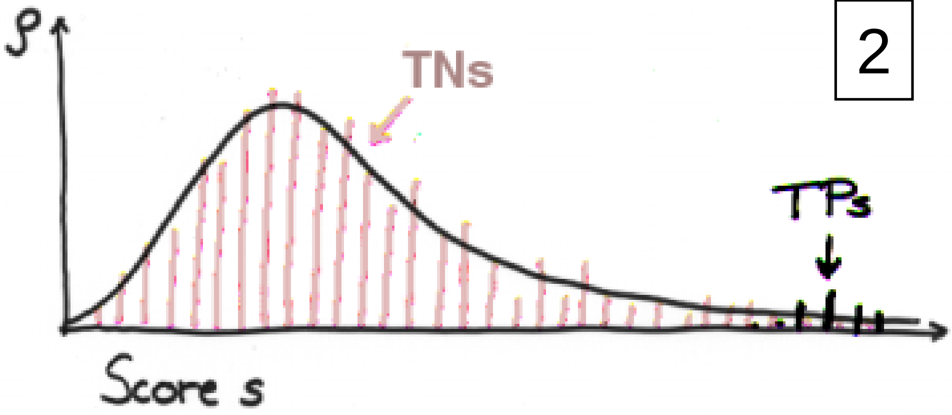
Very Subtle: G3P-dehydrogenase, C-term. Domain
>5 Å



P-values

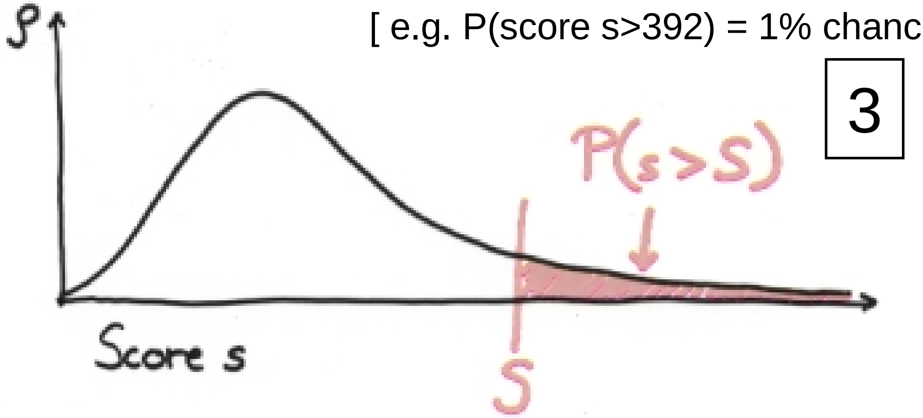


1



2

[e.g. P(score s>392) = 1% chance]

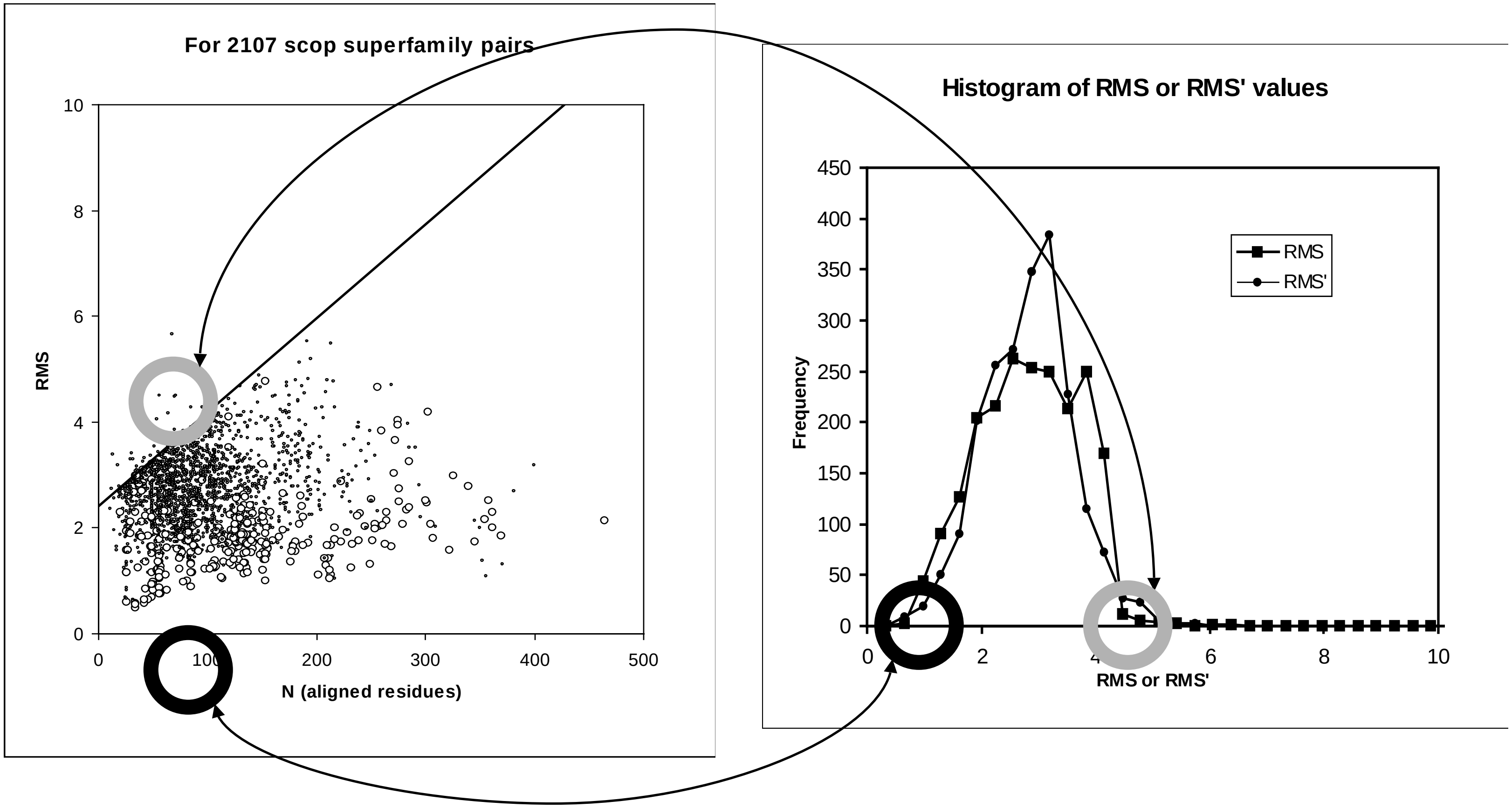


3

- •Significance Statistics
  - ◊ For sequences, originally used in Blast (Karlin-Altschul). Then in FASTA, &c.
  - ◊ Extrapolated Percentile Rank: How does a Score Rank Relative to all Other Scores?
- •Our Strategy: Fit to Observed Distribution
  - **1)** All-vs-All comparison
  - **2)** Graph Distribution of Scores in 2D (N dependence); 1K x 1K families -> ~1M scores; ~2K included TPs
  - **3)** Fit a function ρ(S) to TN distribution (TNs from scop); Integrating ρ gives P(s>S), the CDF, chance of getting a score better than threshold S randomly
  - 4) Use same formalism for sequence & structure

Statistics on Range of Similarities

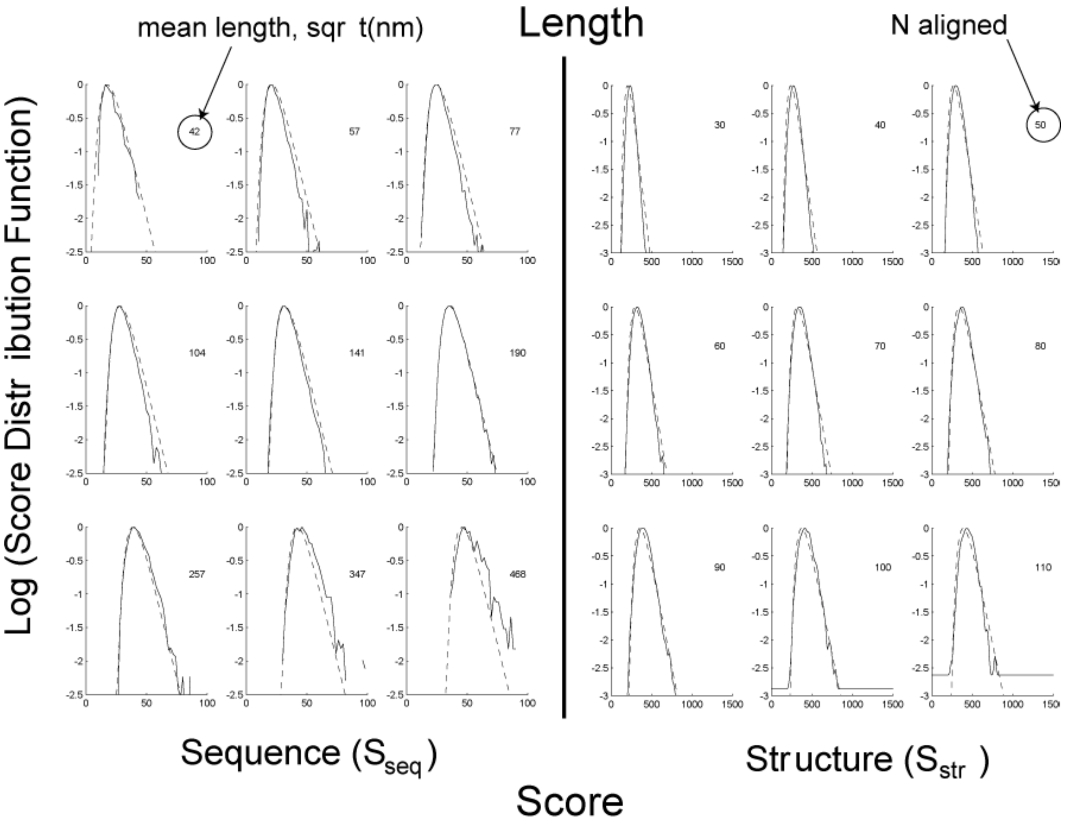
For 2107 pairs, only 2% Outliers (with subtle similarity)

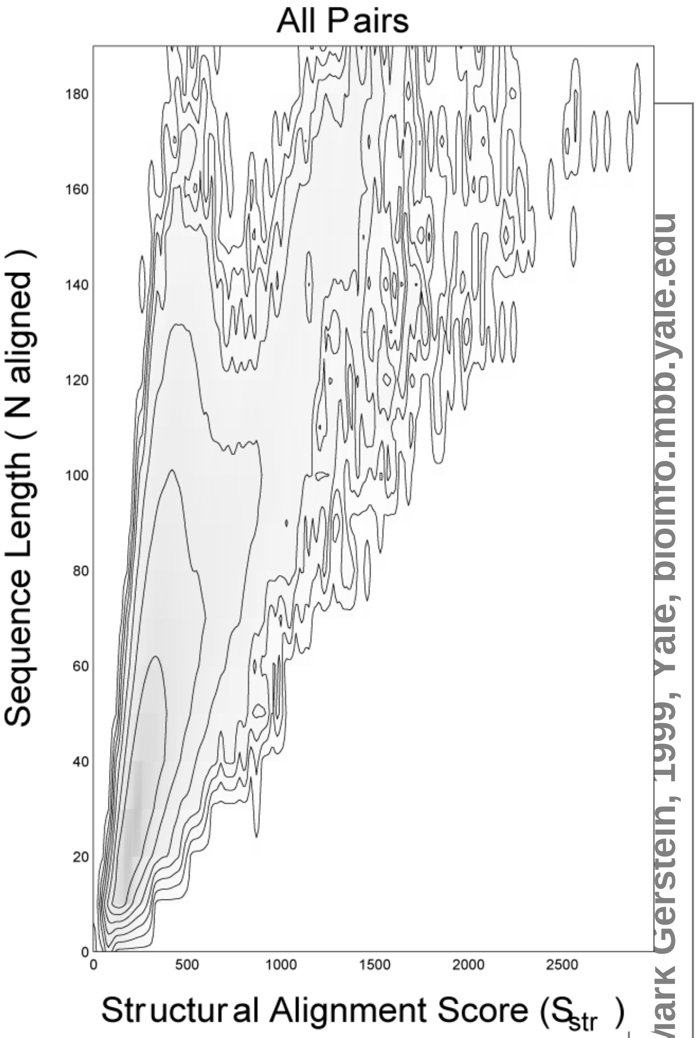


RMS

Num. Aligned

Scores from Structural Alignment Distributed Just Like Ones from Sequence Alignment (E.V.D.)





Same Results for Sequence & Structure

3 Free Parm. fit to EVD involving: ***a, b,σ***. These are the only difference betw. sequence and structure.

$$Z = \frac{S - (a \ln N + b)}{c}$$

$$S = \sum_{i,j} M(i,j) - G$$

$$\boxed{\rho(z) = \exp\left(-z - e^{-z}\right)}$$

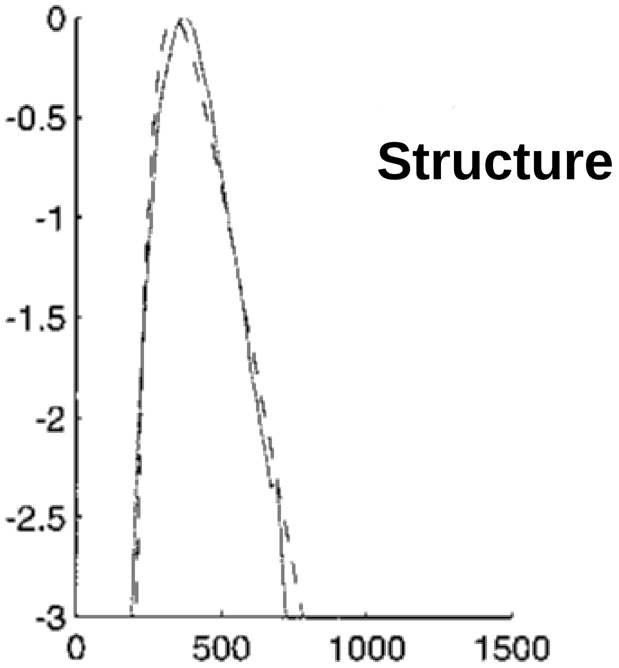
N, G, M also defined differently for sequence and structure.

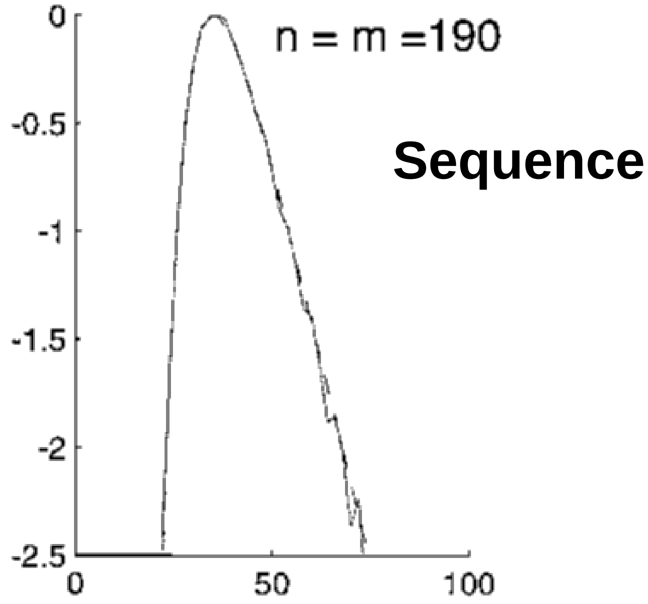
N = number of residues matched.

G = total gap penalty.

M(i,j) = similarity matrix

(Blossum for seq. or M_{str}(i,j), struc.)





Score Significance (P-value) derived from Extreme Value Distribution (just like BLAST, FASTA)

F(s) = E.V.D of scores

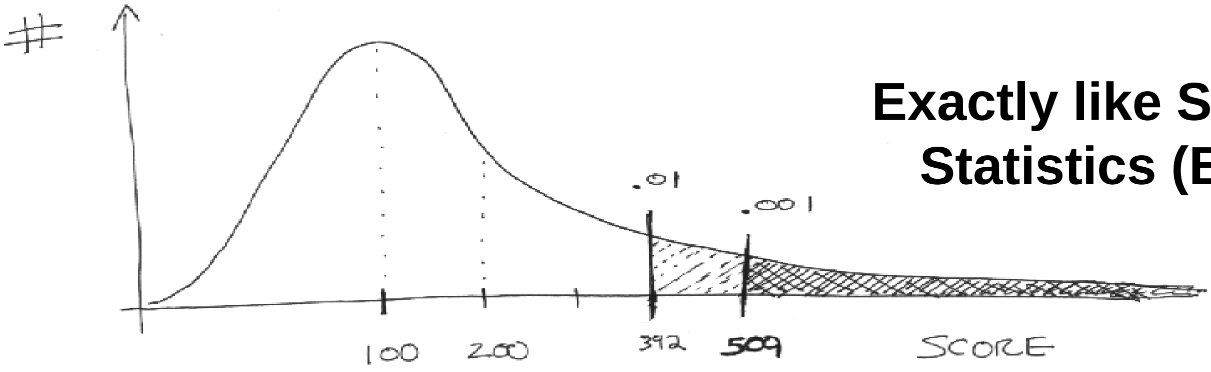
F(s) = exp(-Z(s) - exp(-Z(s)))

Z(s) = As + ln(N) + B

s = Score from random alignment

N length of sequence matched

A & B are fit parameters



P(s>S) = CDF = integral[F(s)]

P(s>S) = 1 - exp(-exp(-Z(s)))

Given Score S (1%), P (s > S) is the chance that a given random score **s** is greater than the threshold

i.e. P-value gives chance score would occur randomly

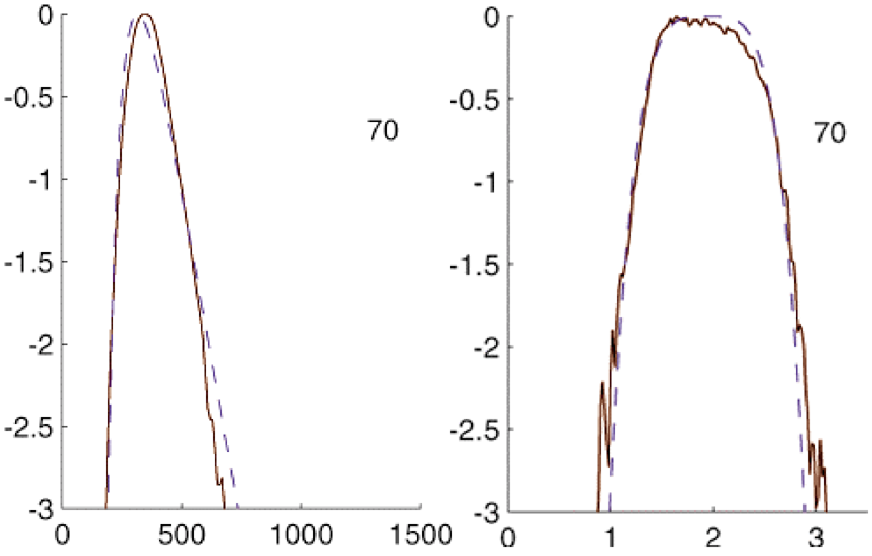
Exactly like Sequence Matching Statistics (BLAST and FASTA)

RMS is a similarity Score

- Also, RMS doesn't work instead of structural alignment (no EVD fit)
  - ◊ RMS penalizes worst fitting atoms, easily skewed

S_{str} RMS

$$\sum \frac{100}{5+\mathbf{d}_i^2} \, vs \quad \sqrt{\sum \mathbf{d}_i^2}$$



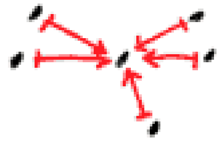
Structure alignment - Other methods

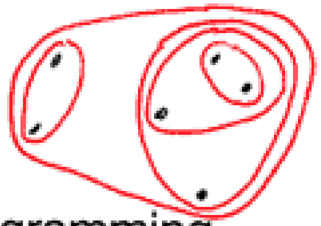
- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  - ◊ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  - ◊ Distance Matrix based methods
  - ◊ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity

- Protein Geometry
- Surface I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking

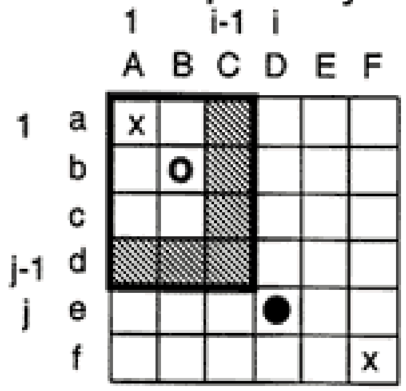
Refine Method

- Multiple Aligment by aligning to central structure





- More Complex Dynamic Programming



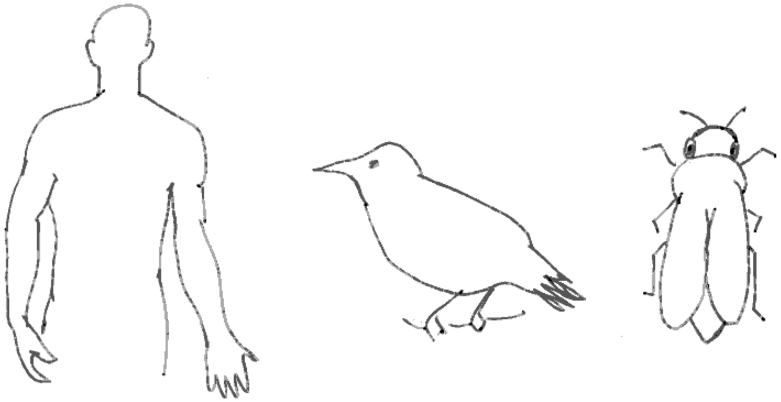
```
AB-C-DEF
abc-de-f
```

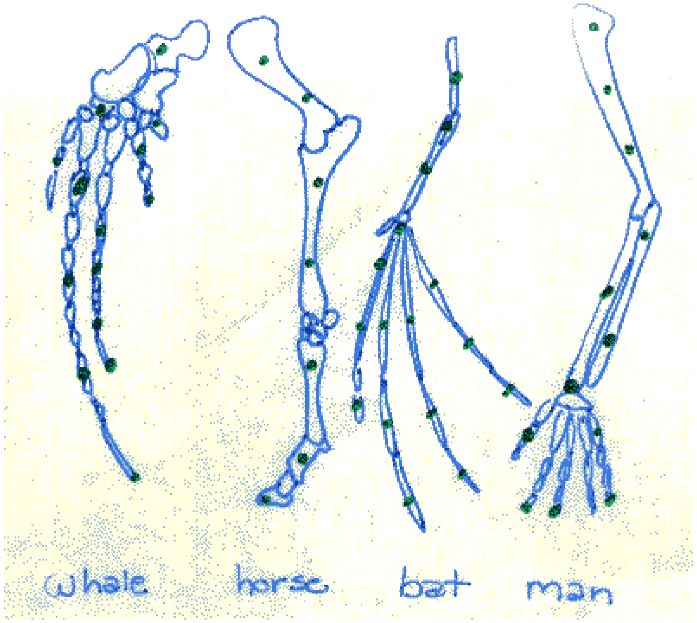
n^2 vs. n^4

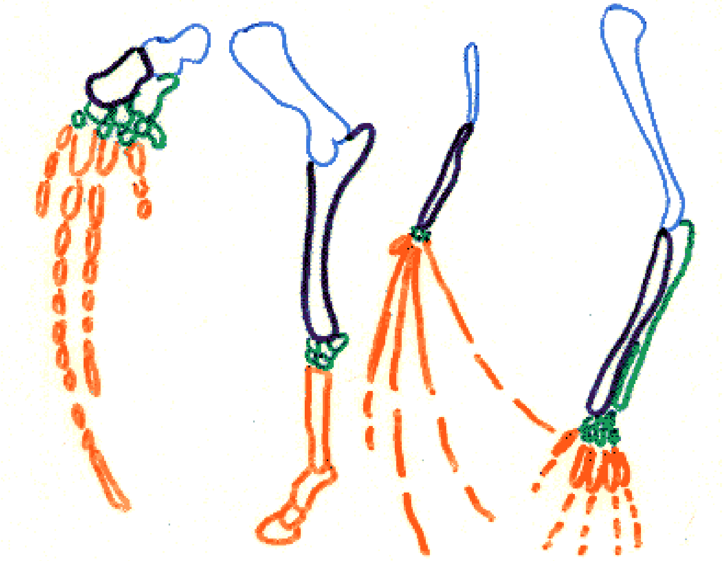
- Find "best" aligned regions
  - "Core-finding" to remove outliers
  - "Noisy" suboptimal paths

	A	y	B	C	D	E
A	2	0	0	0	0	0
B	0	2	3	1	1	1
C	0	1	2	5	2	2
x	0	1	1	2	5	4
D	0	1	1	2	6	5
E	0	1	1	2	4	8

Significance Ignoring Crucial Features in Structural Similarity

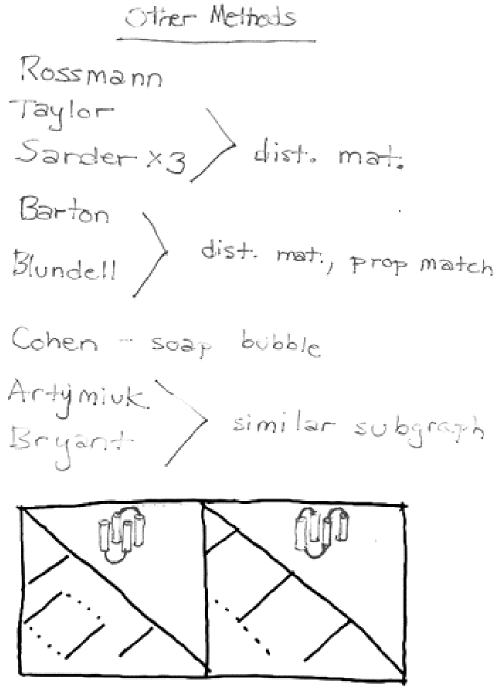






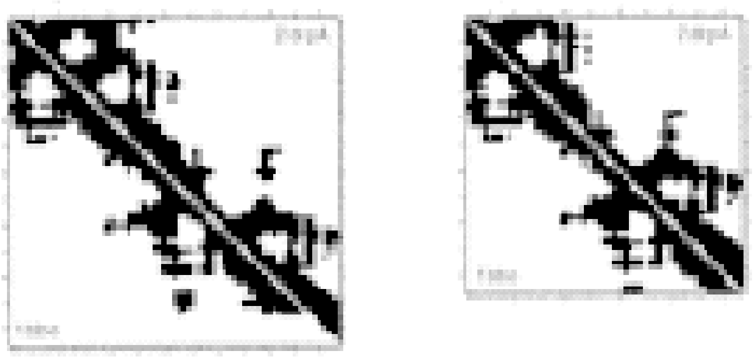
Other Methods of Structural Alignment

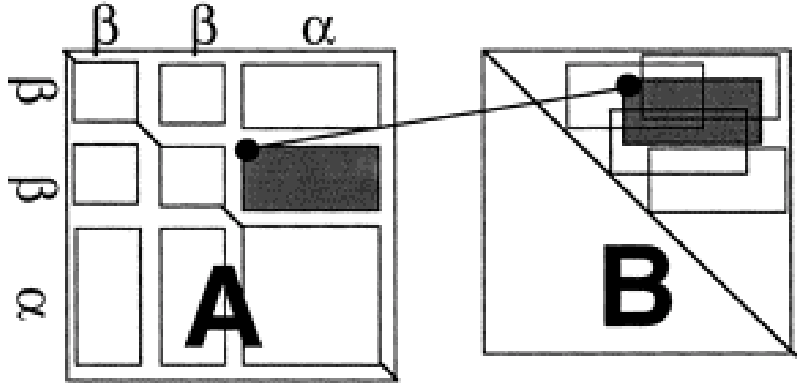
- RMS fitting used universally, but other alignment methods
- Comparison of Distance Matrices
  - ◊ Holm & Sander, DALI
  - ◊ Taylor & Orengo



Other Methods

Rossmann
Taylor
Sander ×3 > dist. mat.
Barton
Blundell > dist. mat.; prop match
Cohen – soap bubble
Artymiuk
Bryant > similar subgraph

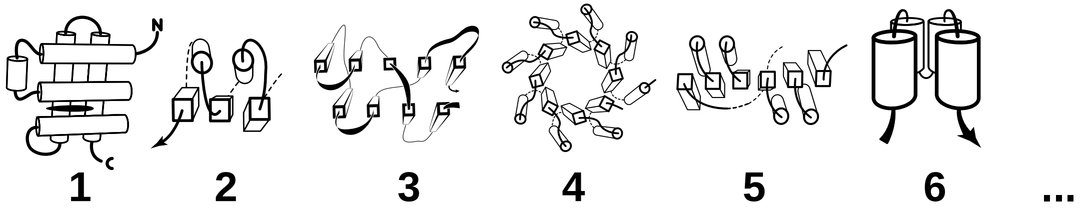




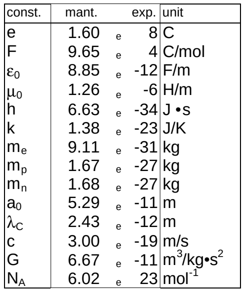
Structure Hashing
Bryant, VAST
Rice, Artymiuk
Others
Cohen (Soap)
Sippl
Godzik (Lattice)

Fold Library vs.
Other Fundamental Data structures

Parts List **Database**; **Statistical**, rather than mathematical relationships and conclusions



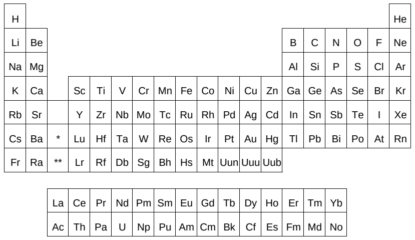
Folds in Molecular Biology **1000-10000**



const.	mant.		exp.	unit
e	1.60	e	8	C
F	9.65	e	4	C/mol
ε_0	8.85	e	-12	F/m
μ_0	1.26	e	-6	H/m
h	6.63	e	-34	J •s
k	1.38	e	-23	J/K
m_e	9.11	e	-31	kg
m_p	1.67	e	-27	kg
m_n	1.68	e	-27	kg
a_0	5.29	e	-11	m
λ_C	2.43	e	-12	m
c	3.00	e	-19	m/s
G	6.67	e	-11	$m^3/kg•s^2$
N_A	6.02	e	23	mol^{-1}

10

Physics

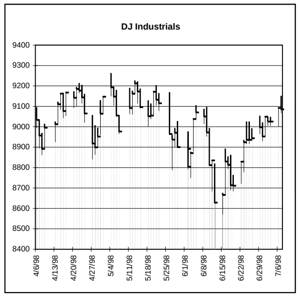


H																		He
Li	Be												B	C	N	O	F	Ne
Na	Mg												Al	Si	P	S	Cl	Ar
K	Ca		Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr		Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	*	Lu	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra	**	Lr	Rf	Db	Sg	Bh	Hs	Mt	Uun	Uuu	Uub						

La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb
Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No

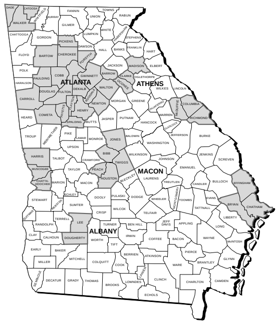
100

Chemistry



1000
-10000

Finance



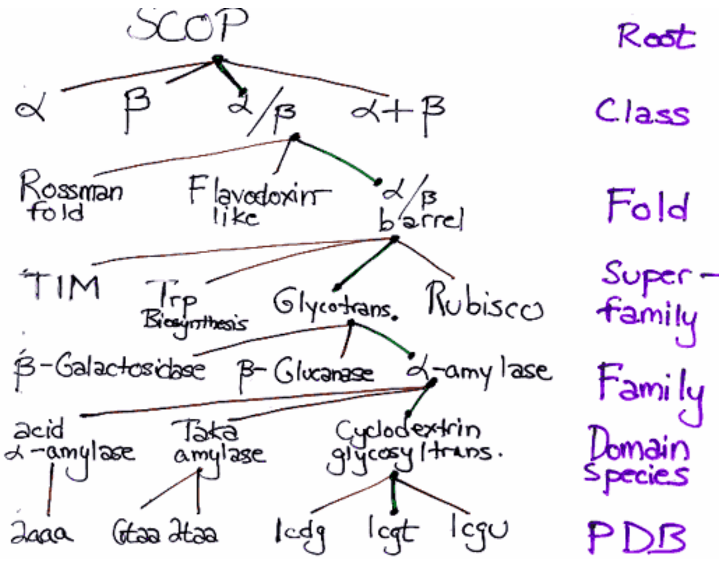
>1000000

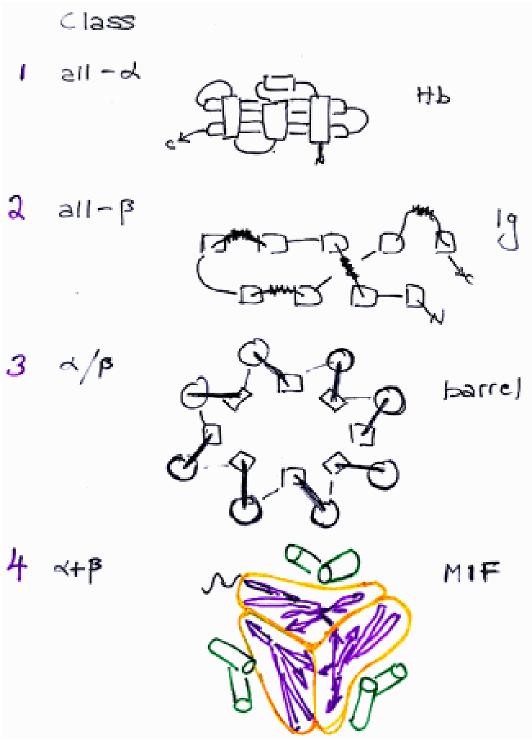
Politics

(Large than physics and chemistry, Similar to Finance (Exact Finite Number of Objects (3,056 on NYSE by 1/98), descrip. by Standardized Statistics (even abbrevs, INTC) and groups (sectors))
Smaller than Social Surveys, Indefinite Number of People, Not Well Defined Vocabulary and statistics.

Fold Classifications

- Scop
  - ◊ Chothia, Murzin (Cambridge)
  - ◊ Manual classification, auto-alignments available
  - ◊ Evolutionary clusters
- Cath
  - ◊ Thornton (London)
  - ◊ semi-automatic classification with alignments
  - ◊ class, arch, topo., homol.
- FSSP
  - ◊ Sander, Holm (Cambridge)
  - ◊ totally automatic with DALI
  - ◊ objective but not always interpretable clusters
- VAST



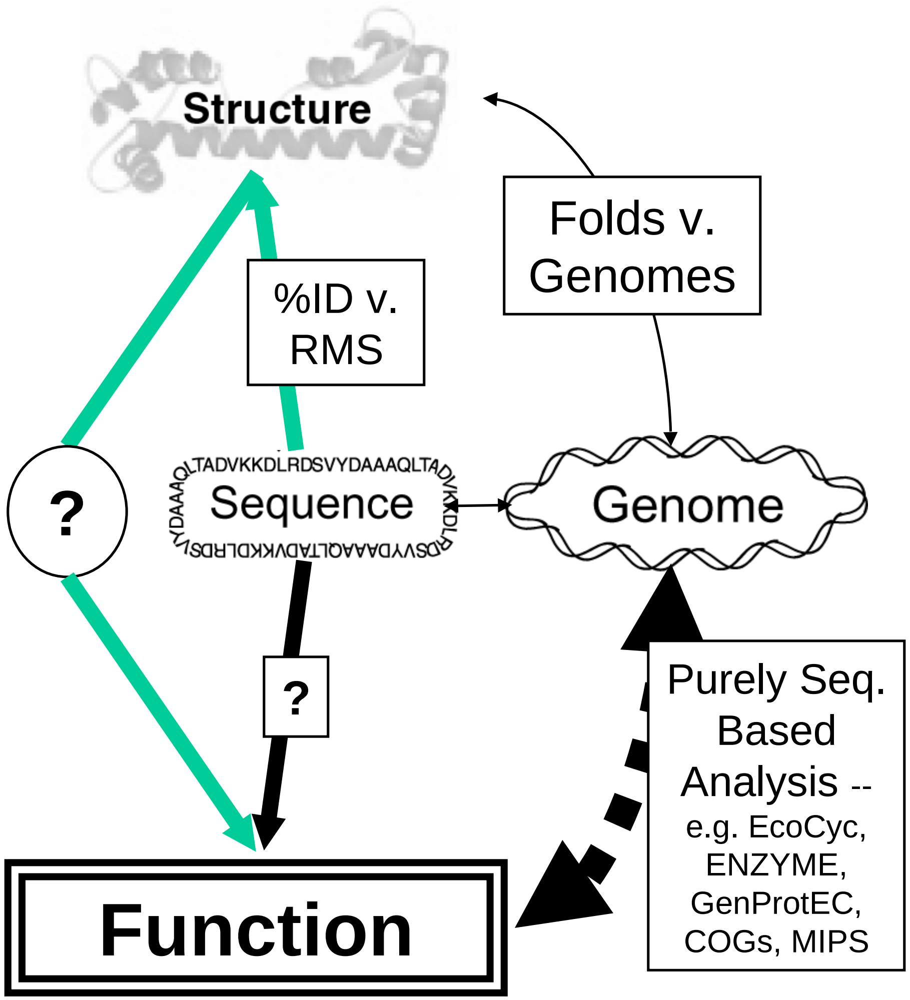


Sequence-structure Relationships

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  - ◊ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  - ◊ Distance Matrix based methods
  - ◊ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity

- Protein Geometry
- Surface I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking

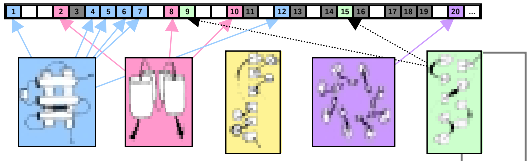
Adding Structure to Functional Genomics, Function to Structural Genomics



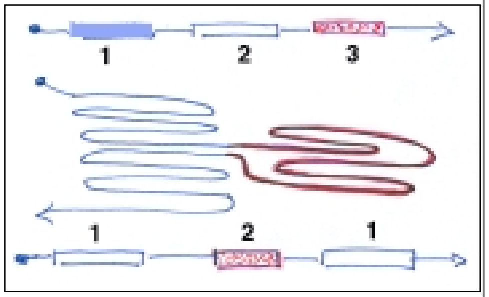
Purely Seq. Based Analysis -- e.g. EcoCyc, ENZYME, GenProtEC, COGs, MIPS

Why Structure? Do we really need it?

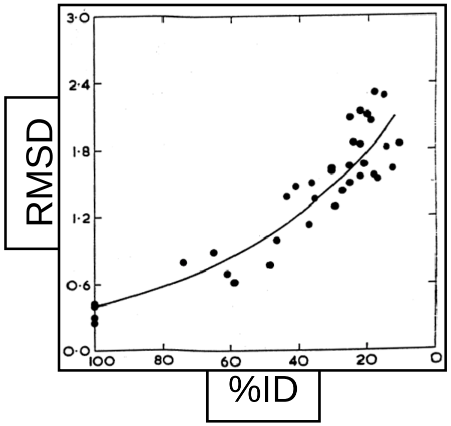
1 Most Highly Conserved



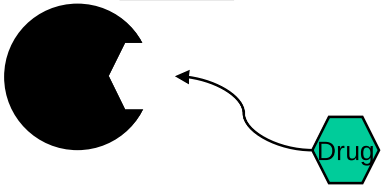
2 Precisely Defined Modules



3 Seq. ⇔ Struc. Clearer than Seq. ⇔Func.



4 Link to Chemistry, Drugs



Chothia & Lesk, 1986 -- 32 points

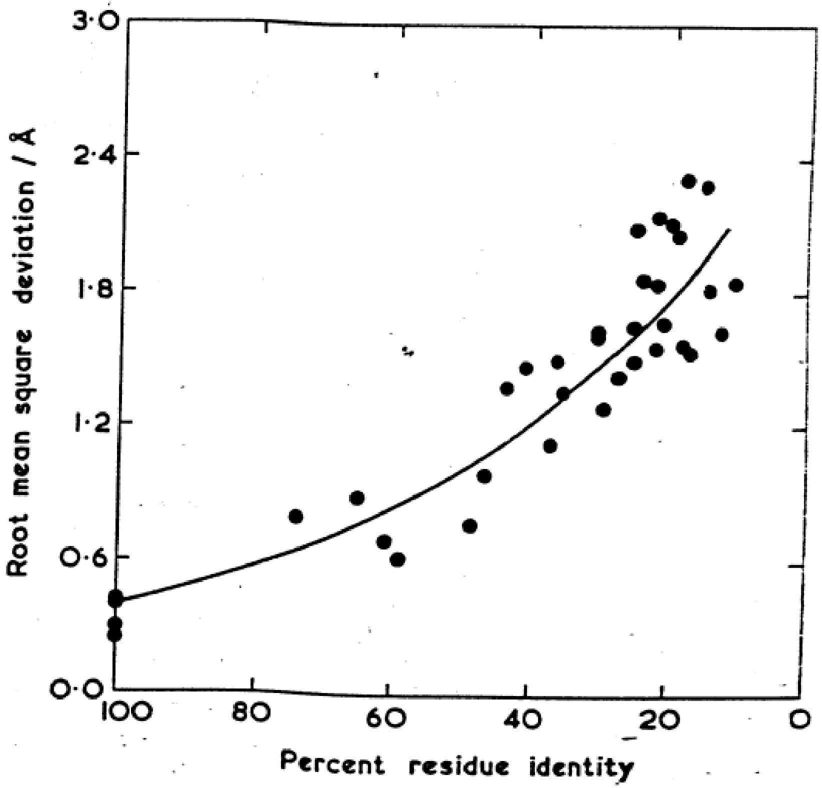


Fig. 2. The relation of residue identity and the r.m.s. deviation of the backbone atoms of the common cores of 32 pairs of homologous proteins (see Table II).

EMBO J **4**: 823 (1986)

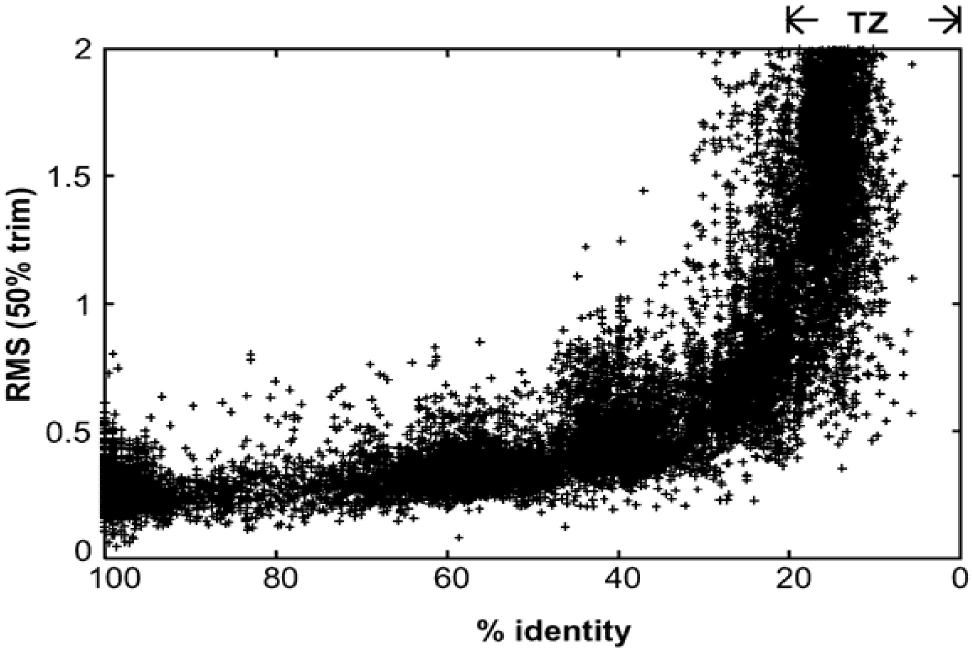
"The relation between the divergence of sequence and structure in proteins"

32 pairs of homologous proteins

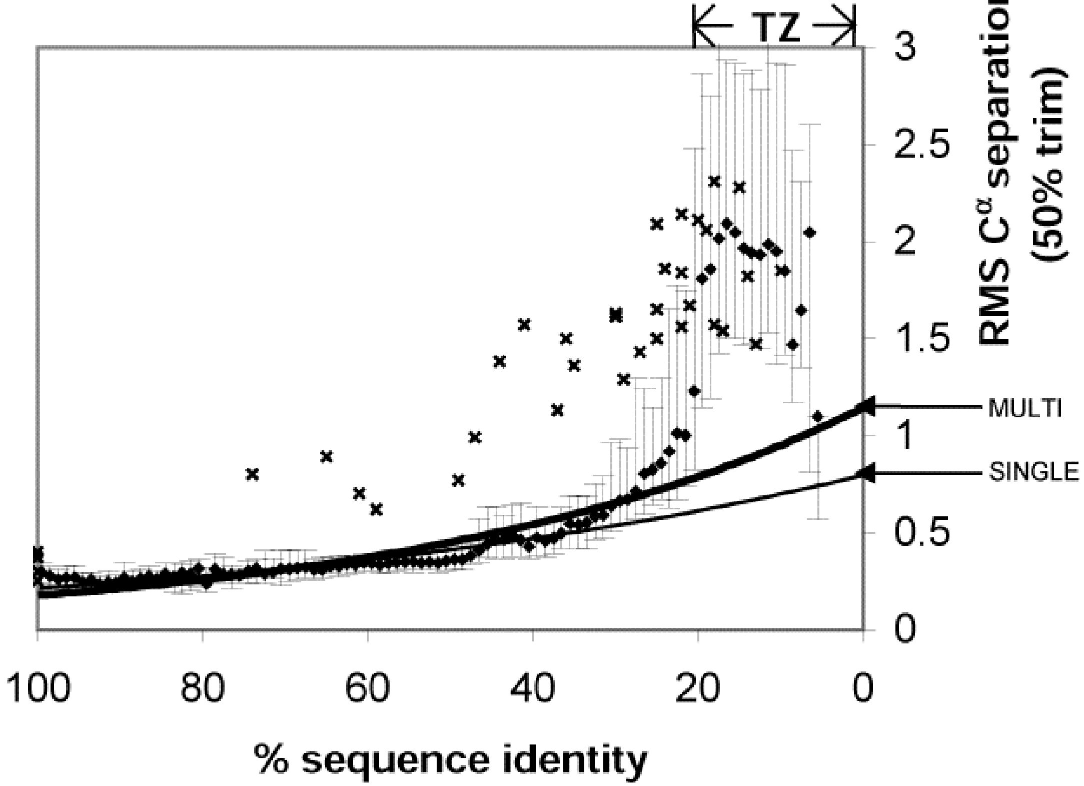
RMS, percent identity

$\Delta = 0.40\ e^{1.87H}$

Now redo with >16,000 pairs in scop +
auto-alignments (pdb95d)....



Chothia and Lesk, revisited 16K points



C&L ‘86:

Δ=.4 exp(1.9 H)

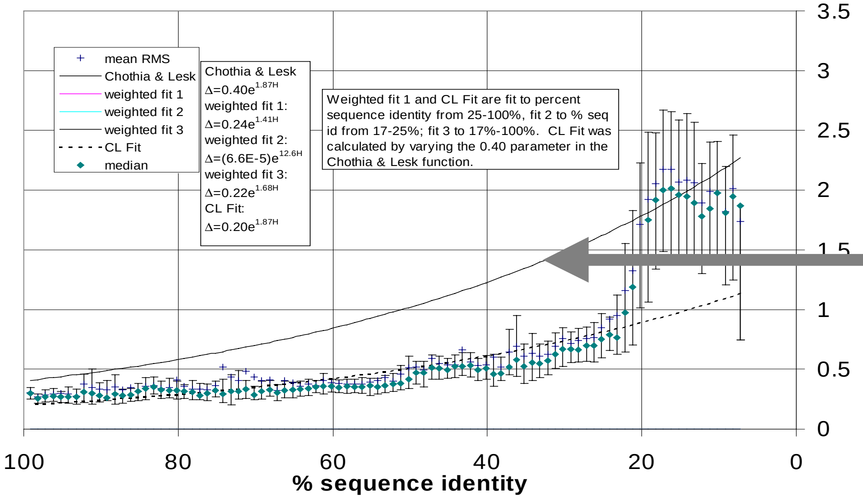
Here:

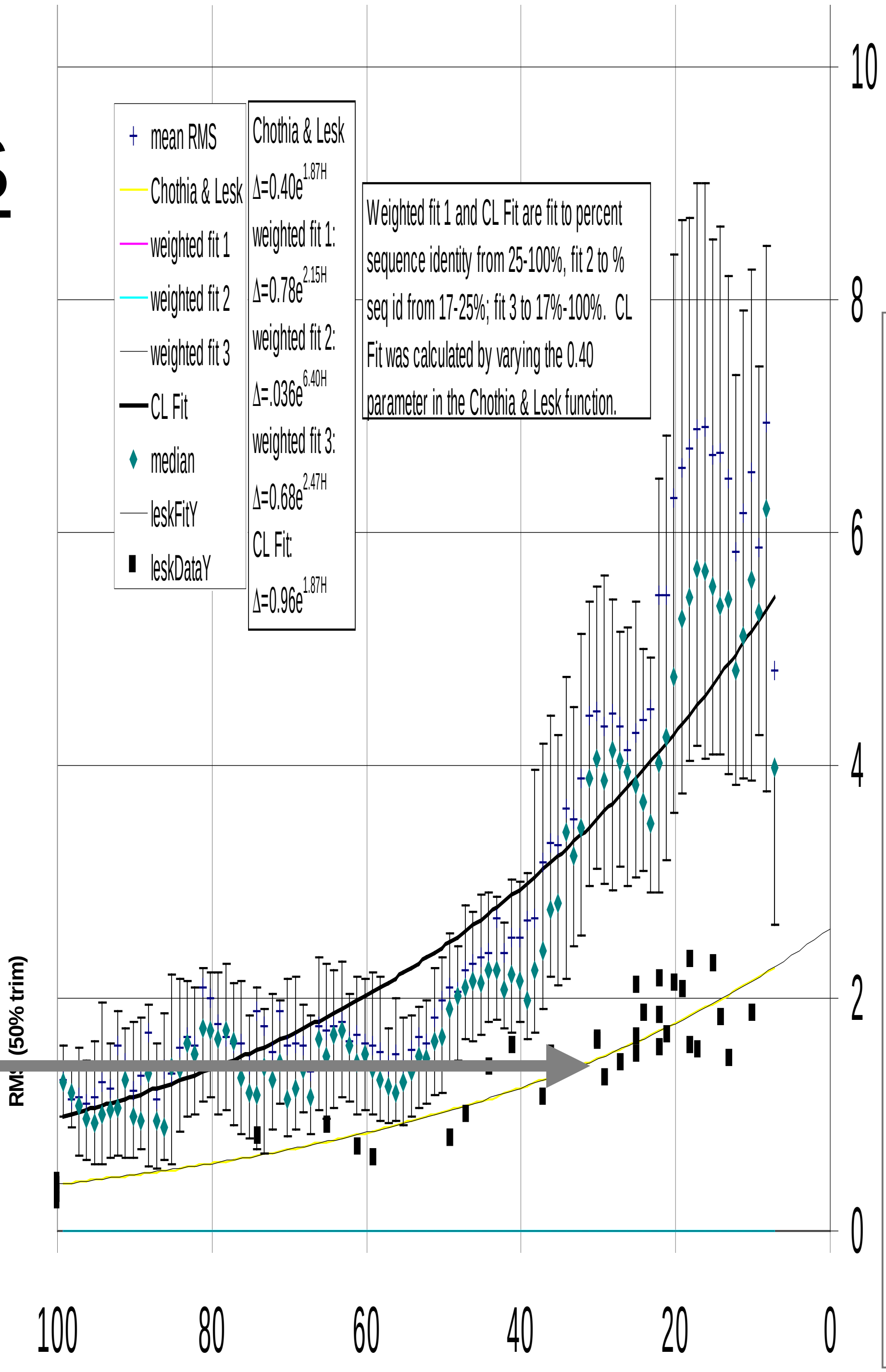
Δ= .2 exp(1.3 H)

Δ= .2 exp(1.9 H)

Problems with RMS

- Dominated by worst-fitting atoms
- Trimming is arbitrary (50%)
- “Bunching up” between 20% and 0% identity

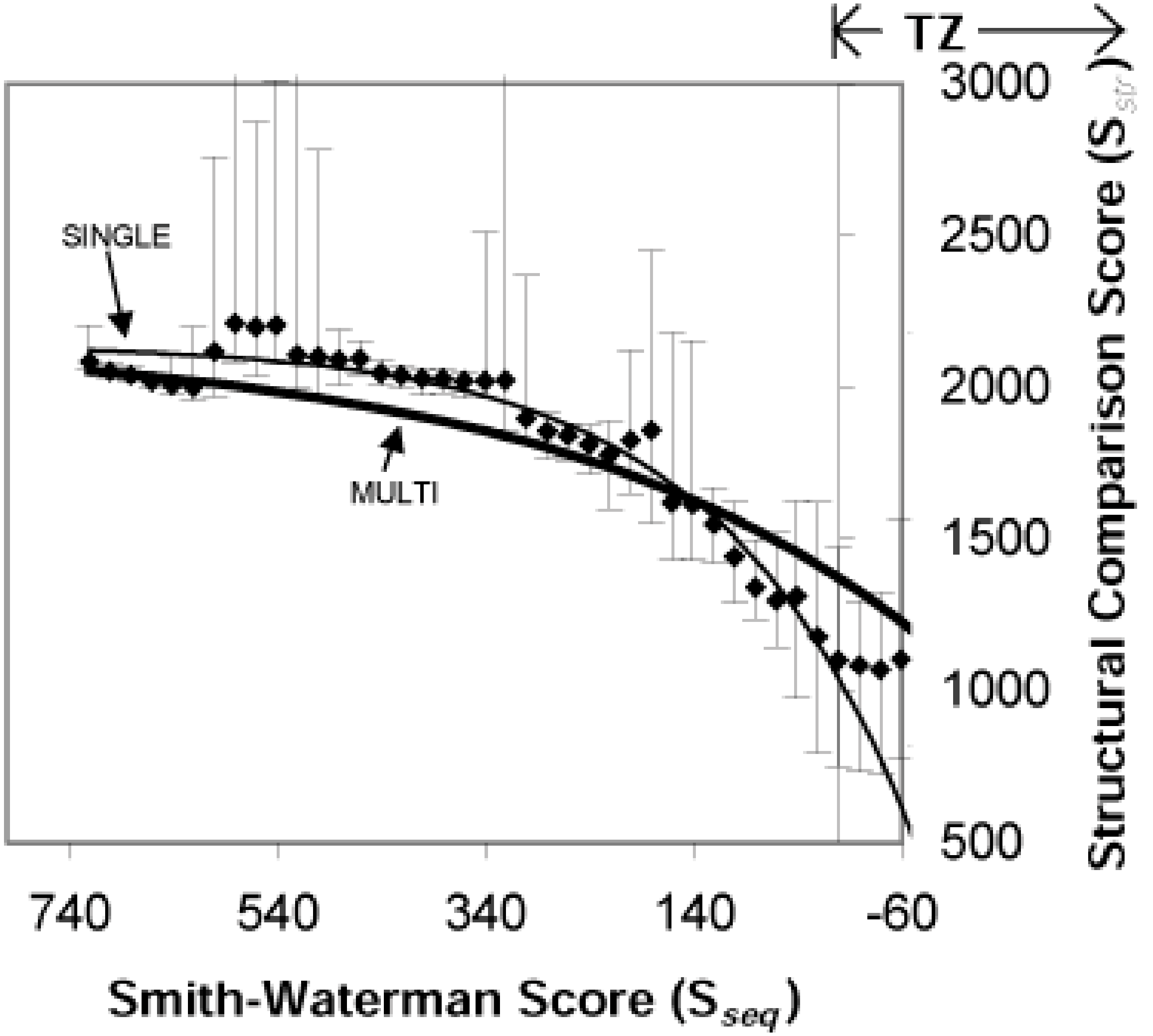




Structural Comp. Score vs.Smith-Waterman Score

overcomes zero bunching, trimming problem

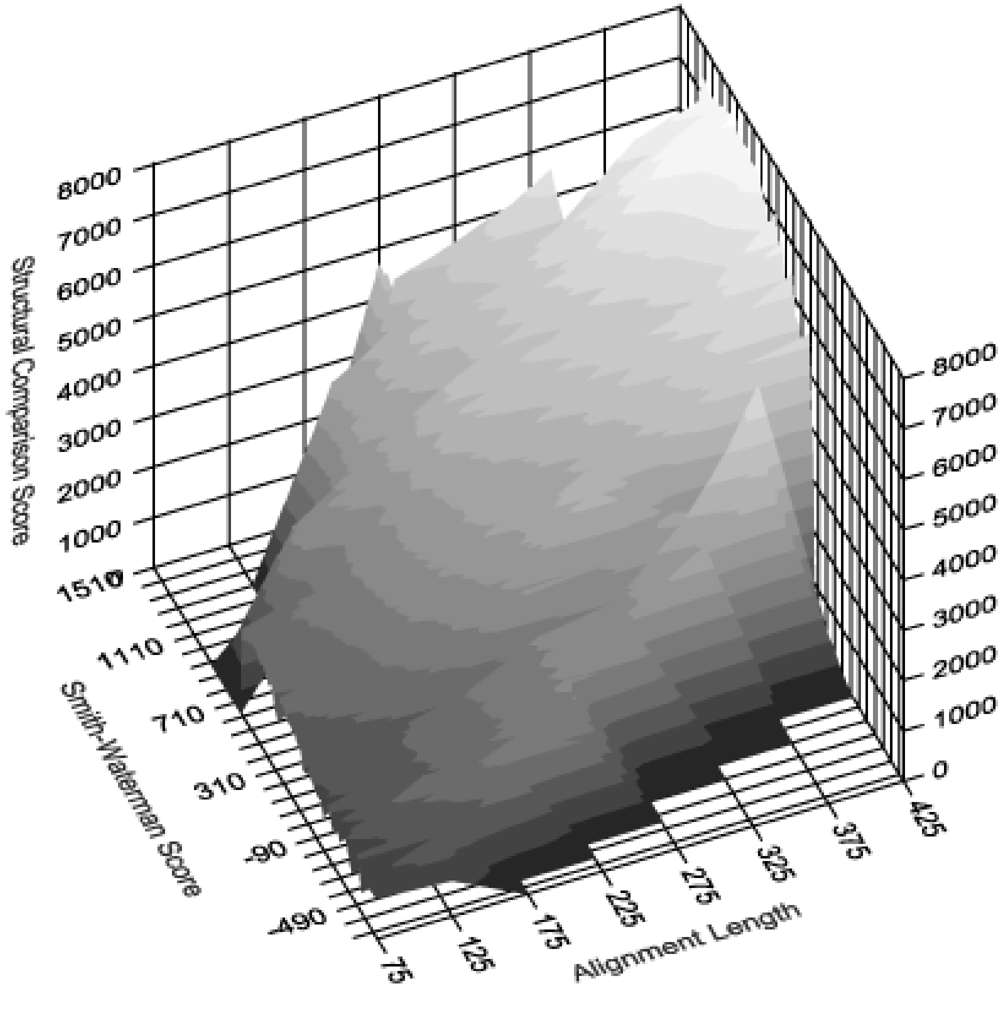
Sstr =
100(21
- 11 exp (-0.0054 SWS)



Problems with Structural Alignment Score

Different Lengths give different scores.

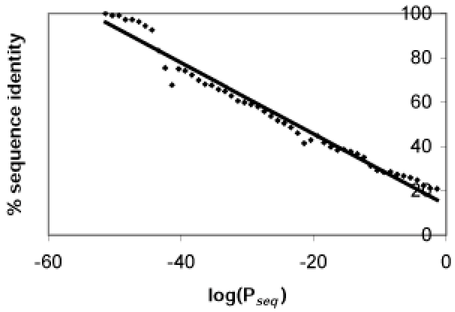
Scores follow equation of the form: $y=A\mathbf{n}+M\mathbf{x}+B$

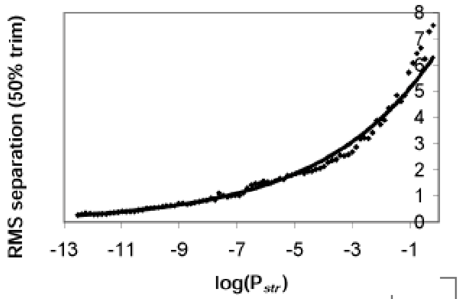


Modern statistical language

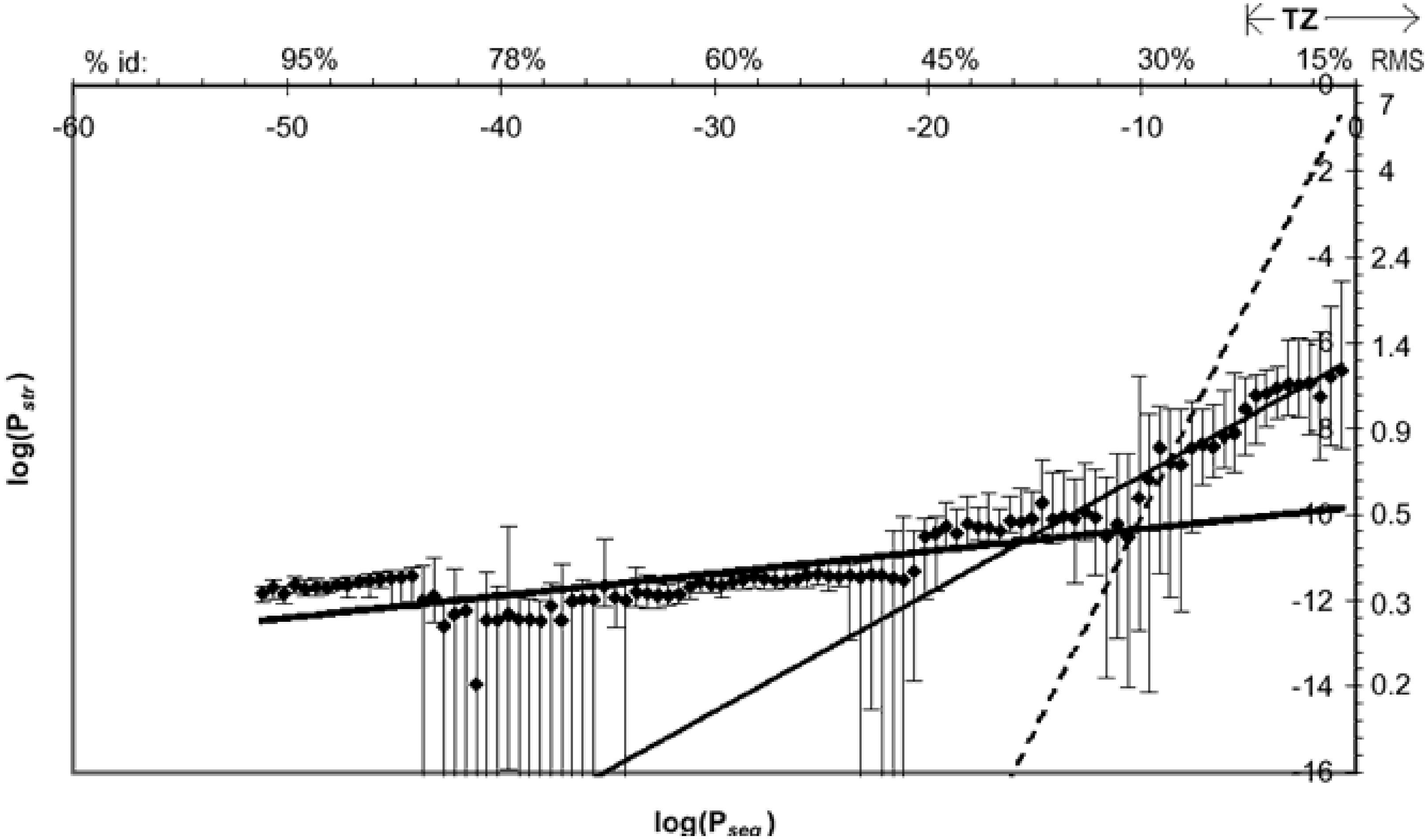
~in TZ:
$P_{str}=10^{-10}P_{seq}^{.05}$

in TZ
$P_{str}=10^{-6}P_{seq}^{.274}$



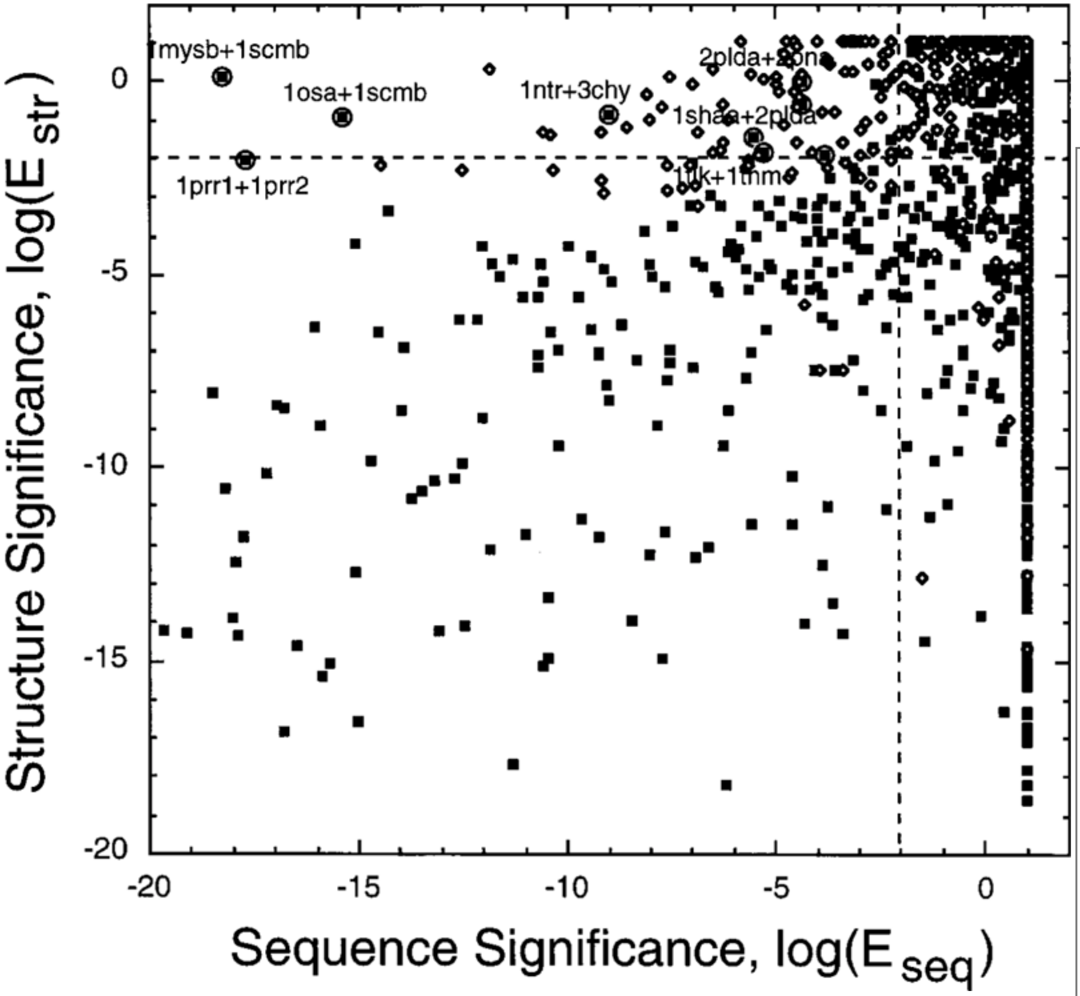


overcomes length dependency

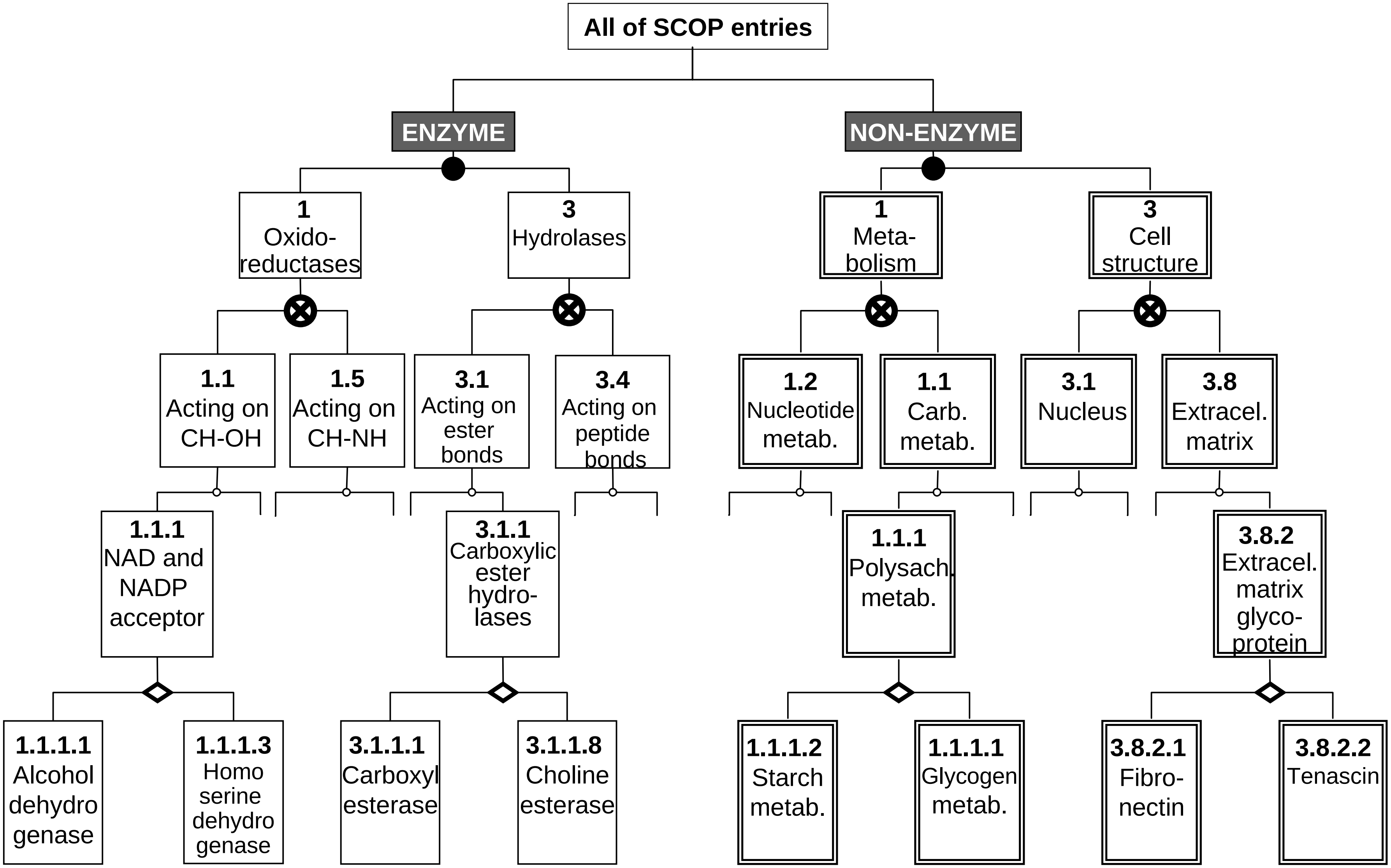


Focus on Twilight Zone

- Sequence Sig. without structure signif.
  - ◊ Protein motions
  - ◊ small proteins
  - ◊ low-res, NMR
- Struc. Sig. without Seq. signif.
  - ◊ More in bottom-right than top-left

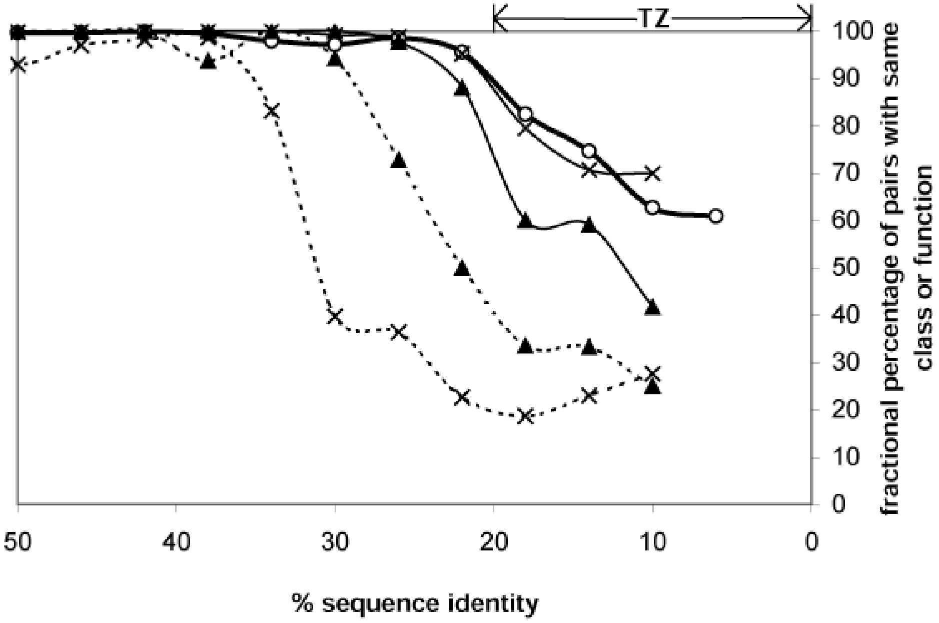


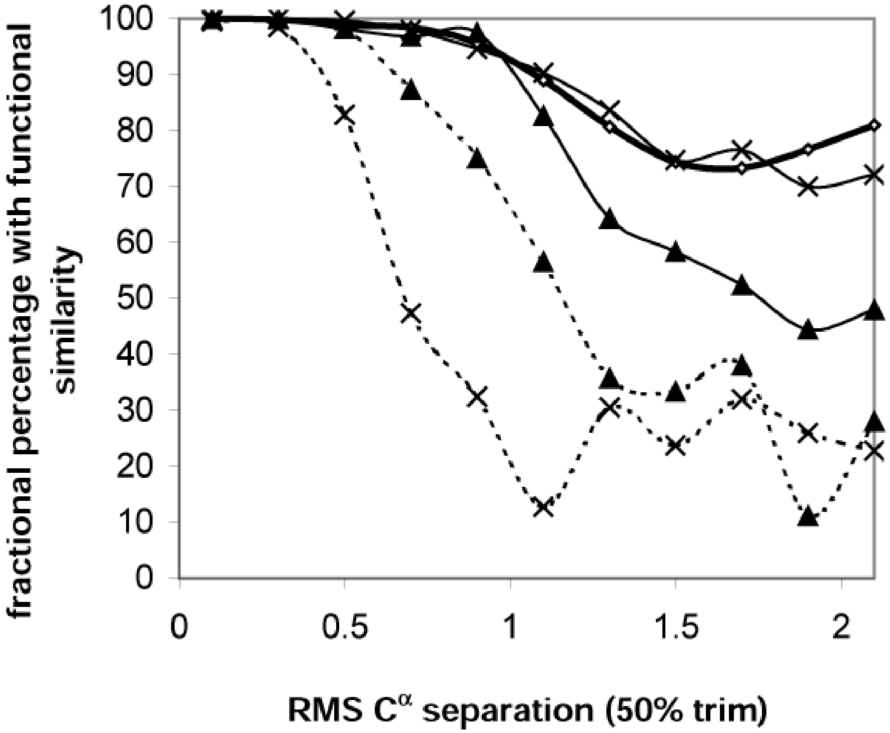
Hierarchy of Protein Functions

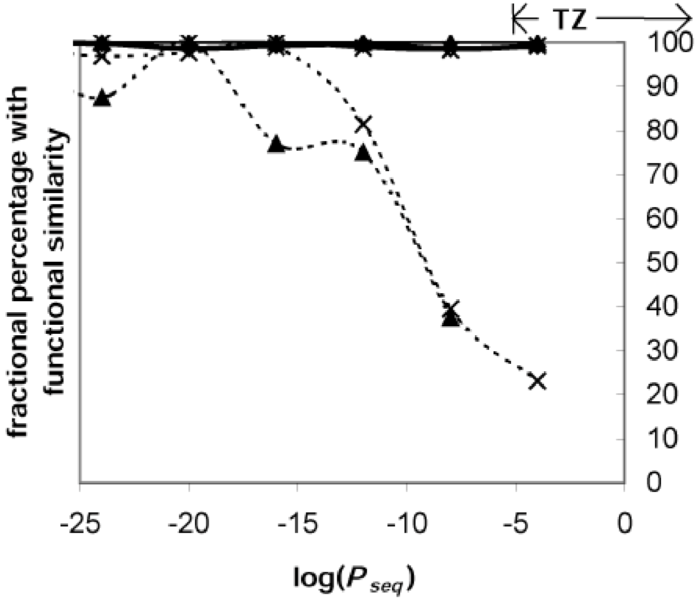


◇ **Precise functional similarity** ● **General similarity** ⊕ **Functional class similarity**

Relationship of Similarity in Sequence & Structure to that in Function







(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Relationship of Similarity in Sequence & Structure, & Function - Summary

	Sequence Similarity	Structural Similarity	Features	Limitations
Traditional Scores	Percent sequence identity	RMS C^{α} separation	Well understood, in use	RMS depends most highly on worst matches, requiring arbitrary trimming
Aligment Similarity Scores	S_{seq}	S_{str}	Analogous similarity scores, S_{str} depends most highly on best matches	Dependence on alignment length
Modern Probabilistic Scores	P_{seq}	P_{str}	Statistical significance, unified framework for different comparisons	Not as familiar as RMS and percent identity, some residual length-dependency

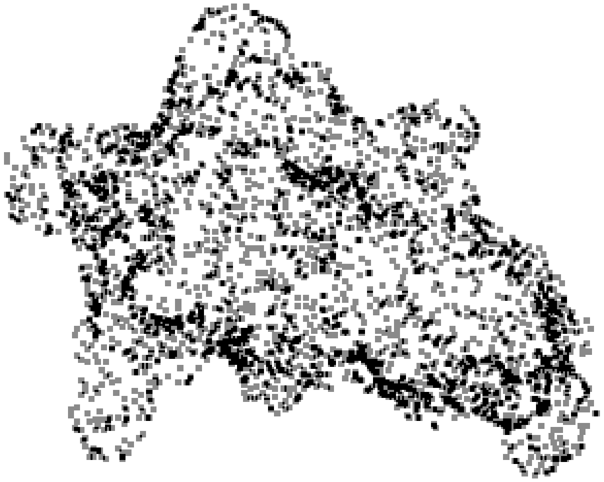
Surfaces I

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  - RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  - Distance Matrix based methods
  - Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity

- Protein Geometry
- Surfaces I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking

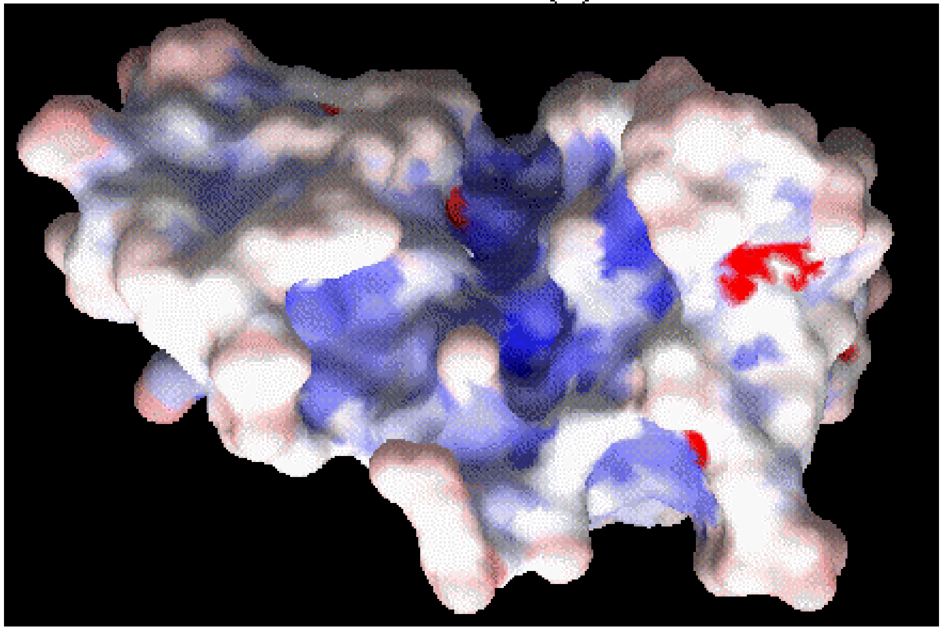
Protein Surfaces -- Accessible Surface

- Why calculate?
  - ◊ Protein is solid object. Surface is where action takes place.
  - ◊ Surface useful for docking and drug-design
  - ◊ Hydrophobic energy proportional to surface area
- Various Types of Protein Surfaces
  - ◊ Accessible Surface
  - ◊ Molecular Surface
  - ◊ Hydration Surface
- Accessible Surface
  - ◊ Roll sphere (water) on surface and look at locus of sphere centers.
  - ◊ Usually represented as a dot surface
  - ◊ Not smooth and continuously differentiable (relevant for energy calculations). It has sharp cusps.



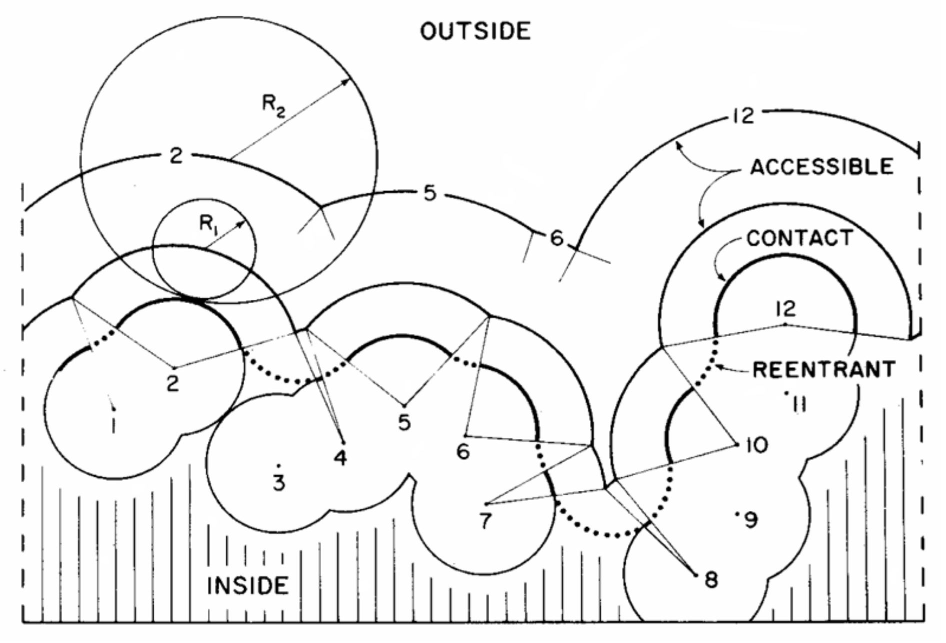
Molecular Surface

08 Molecular Surface (2)



- Cusps in the Accessible Surface
- Solution: the smooth molecular surface.
  - ◊ M.S. = contact surface + re-entrant surface
  - ◊ C.S. = points of tangency between probe sphere and protein when probe sphere is only touching one atom
  - ◊ R.S. = solid angle of probe sphere when tangent to two protein atoms
  - ◊ First proposed by Richards, but hard to calculate. First numeric calc. by Connelly. Later analytic calculation by Connelly.
  - ◊ Analytic version is continuously differentiable.

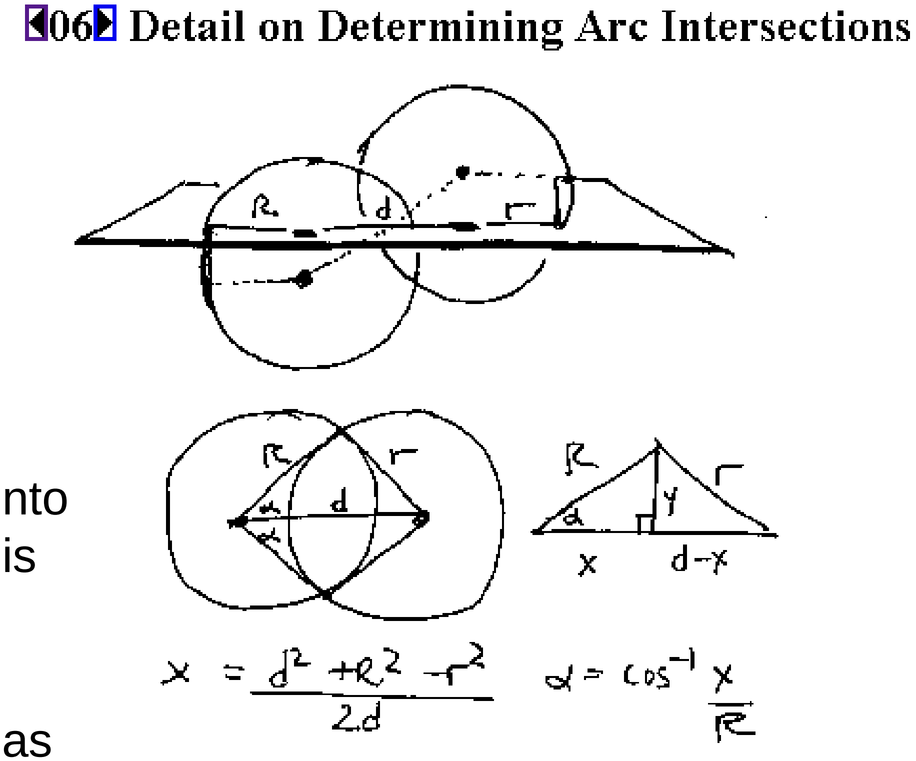
Richards' Molecular and Accessible Surfaces



Probe Radius	Part of Probe Sphere	Type of Surface
0	Center (or Tangent)	Van der Waals Surface (vdWS)
1.4 Å	Center	Solvent Accessible Surface (SAS)
""	Tangent (1 atom)	Contact Surface (CS, from parts of atoms)
""	Tangent (2 or 3 atoms)	Reentrant Surface (RS, from parts of Probe)
""	Tangent (1,2, or 3 atoms)	Molecular Surface (MS = CS + RS)
10 Å	Center	A Ligand or Reagent Accessible Surface
∞	Tangent	Minimum limit of MS (related to convex hull)
""	Center	Undefined

How to Calculate Accessible Surface Area

- Lee & Richards algorithm (first method, 1970)
  - ◊ Pick an arbitrary direction from which to view the protein. Slice it into many sections perpendicular to this direction.
  - ◊ In each section, cycle over all the atoms. Each atom is represented as a sphere with a radius that is the sum of its VDW radius plus that of a probe solvent -- i.e. 1.4 for water.


06 Detail on Determining Arc Intersections

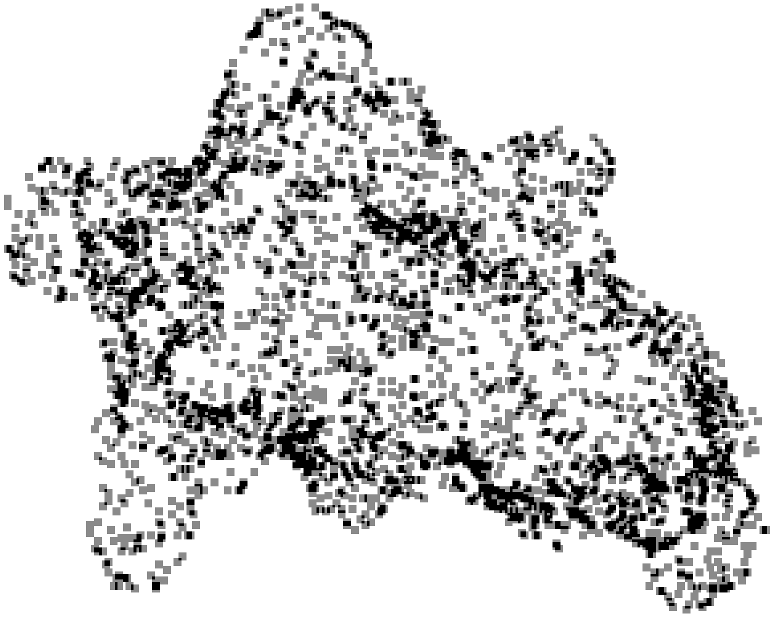
$$x = \frac{d^2 + R^2 - r^2}{2d} \qquad \alpha = \cos^{-1}\frac{x}{R}$$

•For each atom determine the circle corresponding to the intersection of this sphere with the sectioning plane. Remove all parts (i.e. arcs) of this circle occluded by the circles of other atoms.

•Multiply the total amount of non-occluded arc length by the sectioning width to get the surface area for atom. Sum over all atoms and all sections to get total area.

Shrake & Rupley algorithm (easier)

- Surround each atom with sphere of uniformly spaced dots (e.g. 92).
- Remove dots contained in other atoms spheres. Total number of remaining dots is accessible surface.



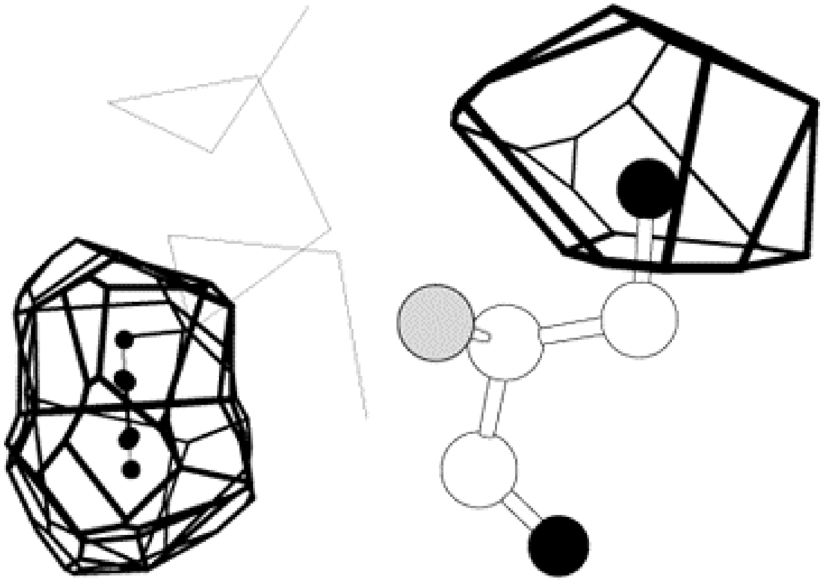
Calculation of Volumes

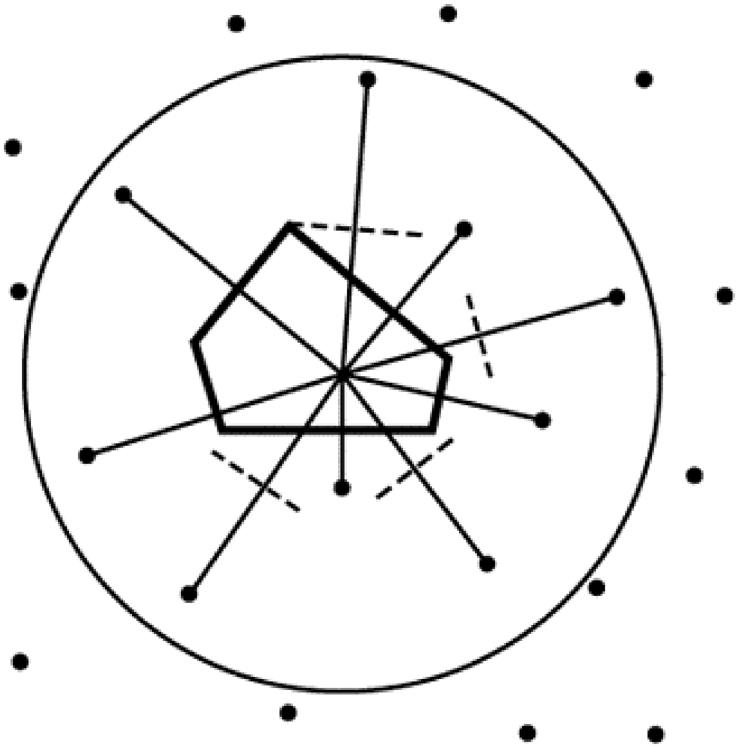
- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  - ◊ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  - ◊ Distance Matrix based methods
  - ◊ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity

- Protein Geometry
- Surfaces I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking

Voronoi Volumes

- Each atom surrounded by a single convex polyhedron and allocated space within it
  - Allocation of all space (large V implies cavities)
- 2 methods of determination
  - Find planes separating atoms, intersection of these is polyhedron
  - Locate vertices, which are equidistant from 4 atoms





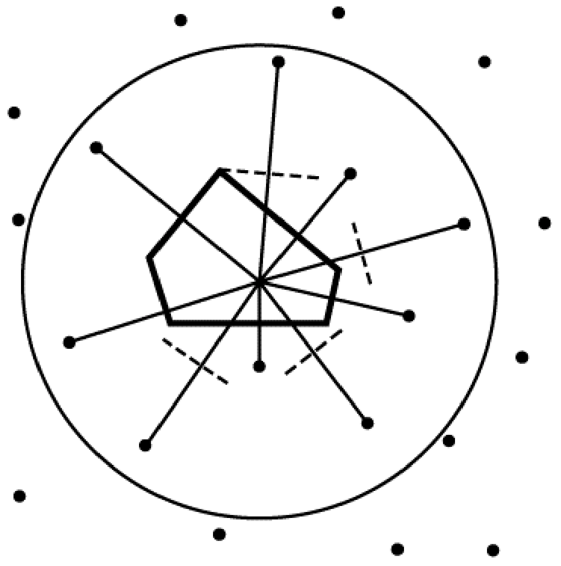
Classic Papers

- Lee, B. & Richards, F. M. (1971). “The Interpretation of Protein Structures: Estimation of Static Accessibility,”
  J. Mol. Biol. **55**, 379-400.
- Richards, F. M. (1974). “The Interpretation of Protein Structures: Total Volume, Group Volume Distributions and Packing Density,”
  J. Mol. Biol. **82**, 1-14.
- Richards, F. M. (1977). “Areas, Volumes, Packing, and Protein Structure,”
  Ann. Rev. Biophys. Bioeng. 6, 151-76.

Calculating Volumes with Voronoi polyhedra

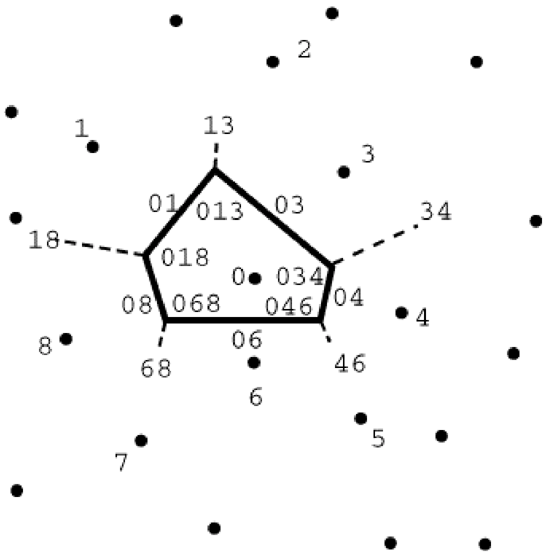
- In 1908 Voronoi found a way of partitioning all space amongst a collection of points using specially constructed polyhedra. Here we refer to a collection of "atom centers" rather than "points.”
- In 3D, each atom is surrounded by a unique limiting polyhedron such that all points within an atom's polyhedron are closer to this atom than all other atoms.
- Likewise, points equidistant from 2 atoms form planes (lines in 2D). Those equidistant from 3 atoms form lines, and those equidistant form 4 centers form vertices.

Determining Voronoi Volumes



- Integrating on a Grid
  - ◊ The simplest method for calculating volumes with Voronoi polyhedra is to put all atoms in the system on a fine grid. Then go to each grid-point (i.e., voxel) and add its infinitesimal volume to the atom center closest to it. This is prohibitively slow for a real protein structure, but it can be made somewhat faster by randomly sampling grid-points. It is, furthermore, a useful approach for high-dimensional integration.
- Solving for the Vertices
  - ◊ In the basic Voronoi construction, each atom is surrounded by a unique limiting polyhedron such that all points within an atom's polyhedron are closer to this atom than all other atoms. Points equidistant from 2 atoms lie on a dividing plane; those equidistant from 3 atoms are on a line, and those equidistant from 4 centers form a vertex.
  - ◊ It is straightforward to solve for possible vertex coordinates using the equation of a sphere. (That is, one uses four sets of coordinates (x,y,z) and the equation $(x\text{-}a)^2 + (y\text{-}b)^2 + (z\text{-}c)^2 = r^2$ to solve for the center (a,b,c) and radius (r) of the sphere.) One then checks whether this putative vertex is closer to these four atoms than any other atom; if so, it is a real vertex.

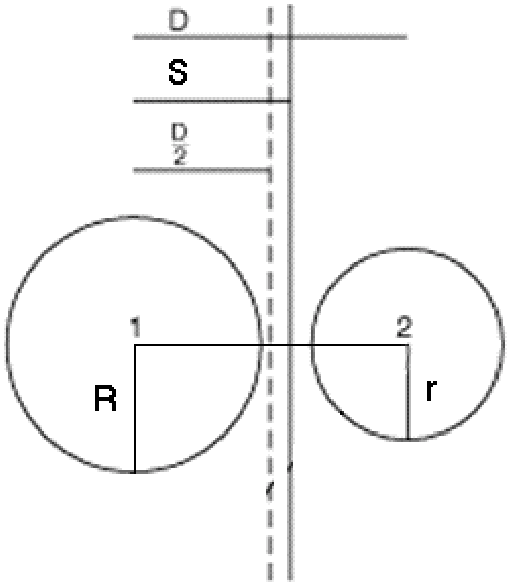
Collecting Vertices and Calculating Volumes



- To systematically collect the vertices associated with an atom, label each one by the indices of the four atoms with which it is associated. To traverse the vertices on one face of a polyhedron, find all vertices that share two indices and thus have two atoms in common — e.g., a central atom (atom 0) and another atom (atom 1). Arbitrarily pick a vertex to start and walk around the perimeter of the face. One can tell which vertices are connected by edges because they will have a third atom in common (in addition to atom 0 and atom 1). This sequential walking procedure also provides a way to draw polyhedra on a graphics device. More importantly, with reference to the starting vertex, the face can be divided into triangles, for which it is trivial to calculate areas and volumes.

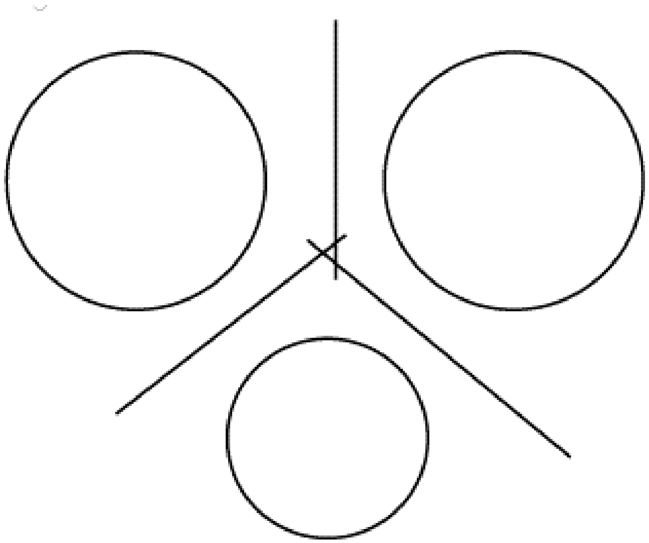
Atoms have different sizes

- Difficulty with Voronoi Meth. Not all atoms created equal
- Solutions
  - ◊ Bisection -- plane midway between atoms
  - ◊ Method B (Richards) Positions the dividing plane according to ratio
  - ◊ Radical Plane
- VDW Radii Set

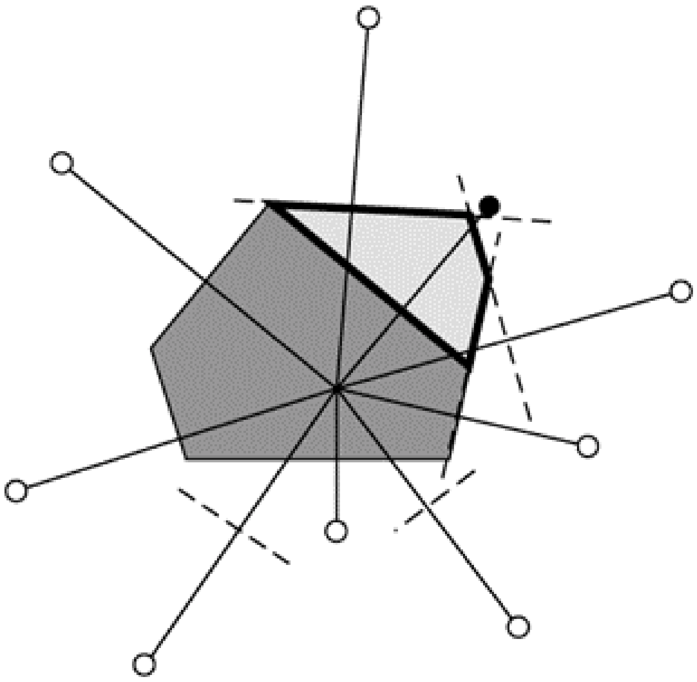


Complexity from different atom sizes requires new ways to calculate polyhedra

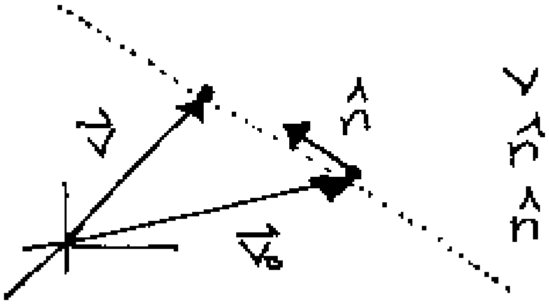
Vertex Error



Chopping Down Method of Calculating Polyhedra

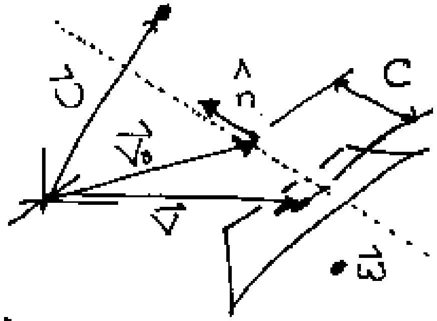


Representing Lines, Planes, and their Intersection



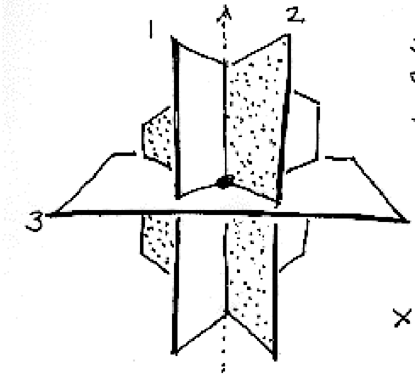
$v = (x\ y\ z)$
$\hat{n} = (\ell\ m\ n)$
$\hat{n}^2 = 1$

Line $\vec{v} - \vec{v}_0 = t\hat{n}$



$(\vec{v}-\vec{v}_0)\cdot\hat{n} = C$ $\quad(\vec{u}-\vec{v}_0)\cdot\hat{n} > C$
$\hat{n}=(\ell\ m n)$ $\vec{v}=(xyz)$ $\quad(w-v_0)\cdot n < C$
$\ell x + my + nz = C'$

◀34▶ **Intersection of Planes and Lines**



$\vec{v}\cdot\hat{n}_1 = C_1$
$\vec{v}\cdot\hat{n}_2 = C_2$
$\vec{v}\cdot\hat{n}_3 = C_3$

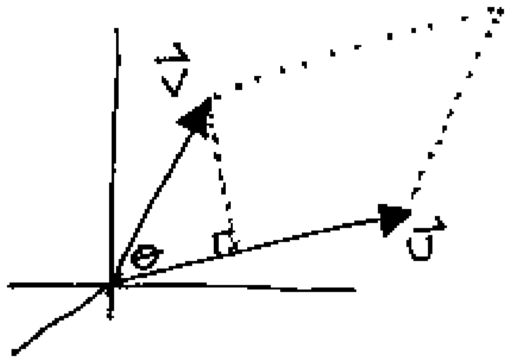
$v = (x\ y z)$
$\hat{n} = (\ell\ m n)$

$$x = \frac{\begin{vmatrix} C_1 & m_1 & n_1 \\ C_2 & m_2 & n_2 \\ C_3 & m_3 & n_3 \end{vmatrix}}{\begin{vmatrix} \ell_1 & m_1 & n_1 \\ \ell_2 & m_2 & n_2 \\ \ell_3 & m_3 & n_3 \end{vmatrix}}$$

Solve for vertex $\vec{v}$
Line (12) then is
$u - \vec{v} = t\ \hat{n}_1 \times \hat{n}_2$

Calculating Areas and Volumes from Vectors

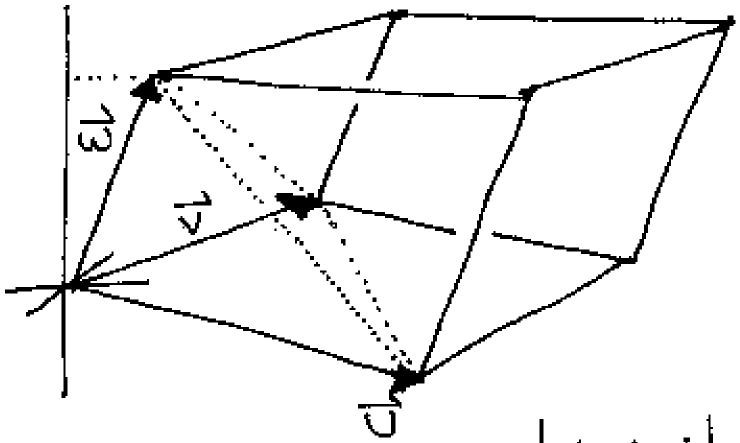
35 Calculating Areas from Vectors



$$A = |\vec{u}||\vec{v}|\sin\theta = |\vec{u} \times \vec{v}|$$

$$A = \begin{vmatrix} u_x & v_x \\ u_y & v_y \end{vmatrix} = \det M$$

36 Calculating Volumes from Vectors

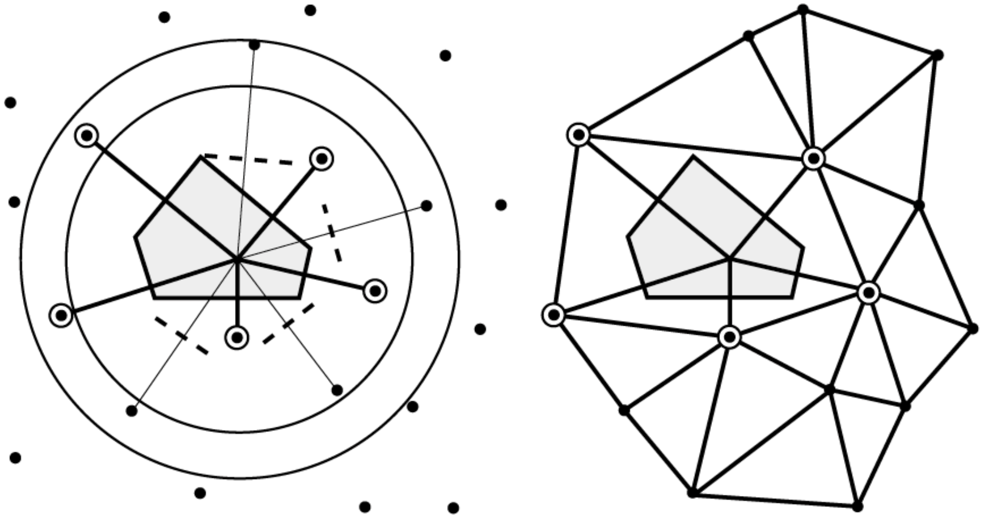


$$V = \vec{w} \cdot (\vec{u} \times \vec{v}) = \begin{vmatrix} \vec{w} & \vec{u} & \vec{v} \end{vmatrix}$$

$$V_{tet} = \frac{1}{6} V_{parallel}$$

Delauney Triangulation, the Natural Way to Define Packing Neighbors

- Related to Voronoi polyhedra (dual)
- What “coordination number” does an atom have? Doesn’t depend on distance
- alpha shape
- threading



Properties of Voronoi Polyhedra

- If Voronoi polyhedra are constructed around atoms in a periodic system, such as in a crystal, all the volume in the unit cell will be apportioned to the atoms. There will be no gaps or cavities as there would be if one, for instance, simply drew spheres around the atoms.
- Voronoi volume of an atom is a weighted average of distances to all its neighbors, where the weighting factor is the contact area with the neighbor.

Voronoi diagrams are generally useful, beyond proteins

- Border of D.T. is Convex Hull
- D.T. produces "fatest" possible triangles which makes it convenient for things such as finite element analysis.
- Nearest neighbor problems. The nearest neighbor of a query point in center of the Voronoi diagram in which it resides
- Largest empty circle in a collection of points has center at a Voronoi vertex
- Voronoi volume of "something" often is a useful weighting factor. This fact can be used, for instance, to weight sequences in alignment to correct for over or under-representation

Voronoi Volumes & Packing

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  - ◊ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  - ◊ Distance Matrix based methods
  - ◊ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity

- Protein Geometry
- Surfaces I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking

Voronoi Volumes,
the Natural Way to Measure Packing

Packing Efficiency

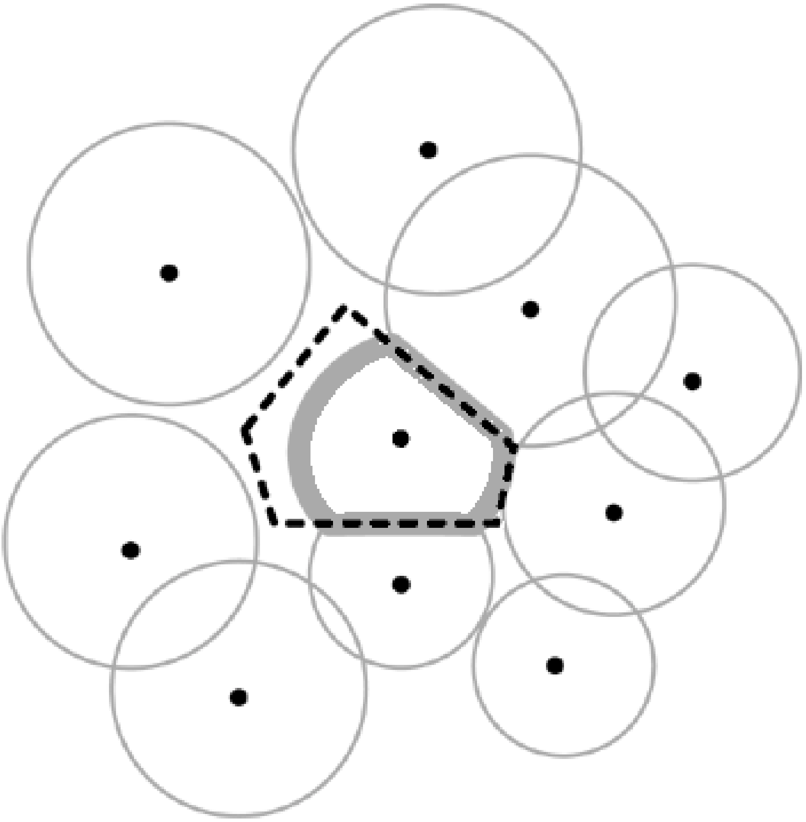
```
= Volume-of-Object
  -----------------
  Space-it-occupies
```

= V(VDW) / V(Voronoi)

- Absolute v relative eff.

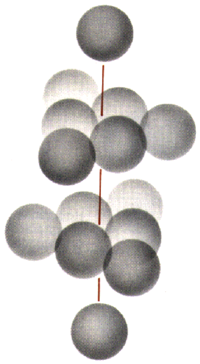
  V1 / V2

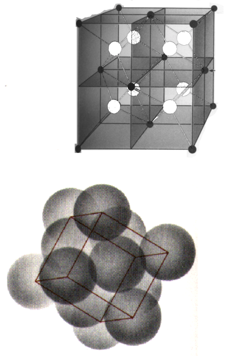
- Other methods
  - ◊ Measure Cavity Volume (grids, constructions, &c)

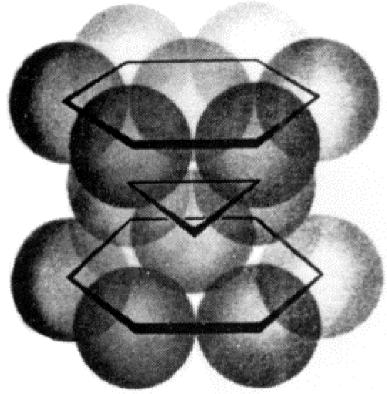


Close-Packing of Spheres

- Efficiency
  - ◊ Volume Spheres / Volume of space
- Close packed spheres
  - ◊ 74% volume filled
  - ◊ Coordination of 12
  - ◊ Two Ways of laying out
- Fcc
  - ◊ cubic close packing
  - ◊ ABC layers
- hcp
  - ◊ Hexagonally close packed
  - ◊ ABABAB







ABCABC . . . , the spheres are cubic close-packed (ccp). The ccp structure gives rise to face-centred unit cells, and so may also be denoted cubic F (or

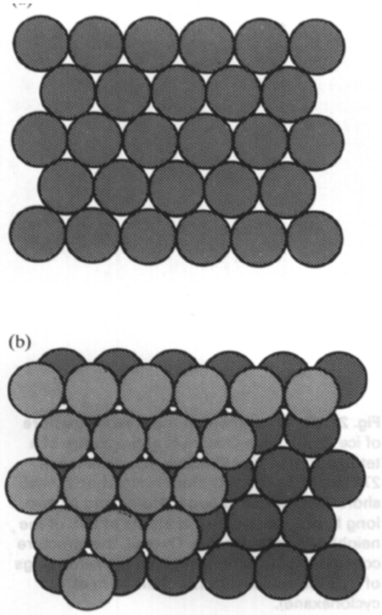


Fig. 21.20 The close-packing of identical spheres. (a) The first layer of close-packed spheres. (b) The second layer of close-packed spheres occupies the dips of the first layer. The two layers are the AB component of the structure.

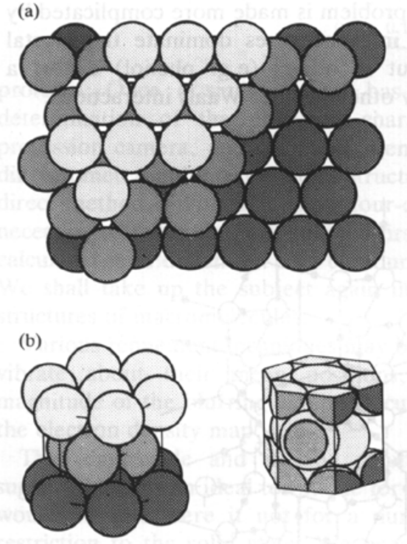


Fig. 21.21 The third layer of close-packed spheres might occupy the dips lying directly above the spheres in the first layer, resulting in an ABA structure (a) which corresponds to hexagonal close-packing (b). This hcp structure is possessed by the elements Be, Cd, Co, He, Mg, Ti, and Zn.

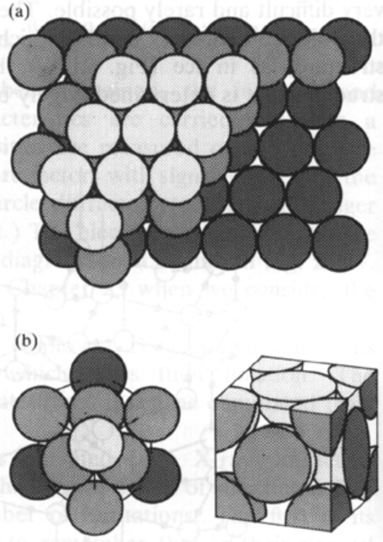
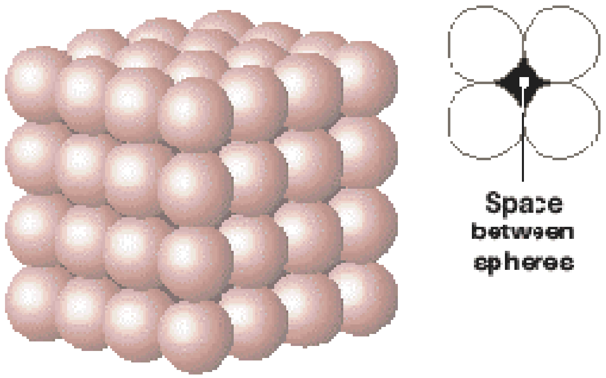


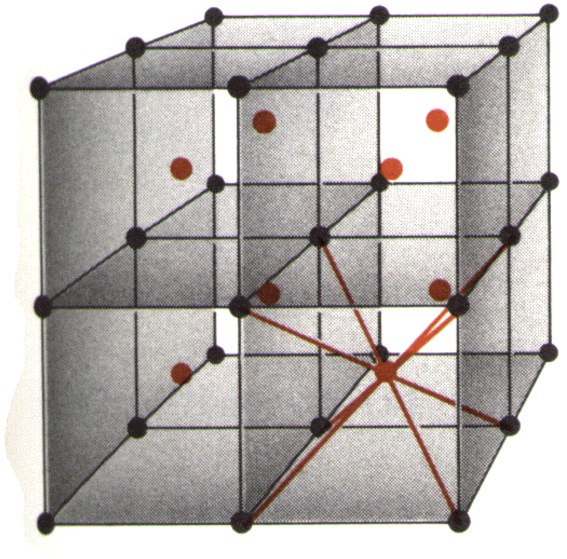
Fig. 21.22 Alternatively, the third layer might lie in the dips that are not above the spheres in the first layer, resulting in an ABC structure (a) which correspond to cubic close-packing (b). This ccp (or fcc) structure is possessed by the elements Ag, Al, Ar, Au, Ca, Cu, Ne, Ni, Pb, Pt, and Xe.

Illustration Credits: Atkins, Pchem, 634

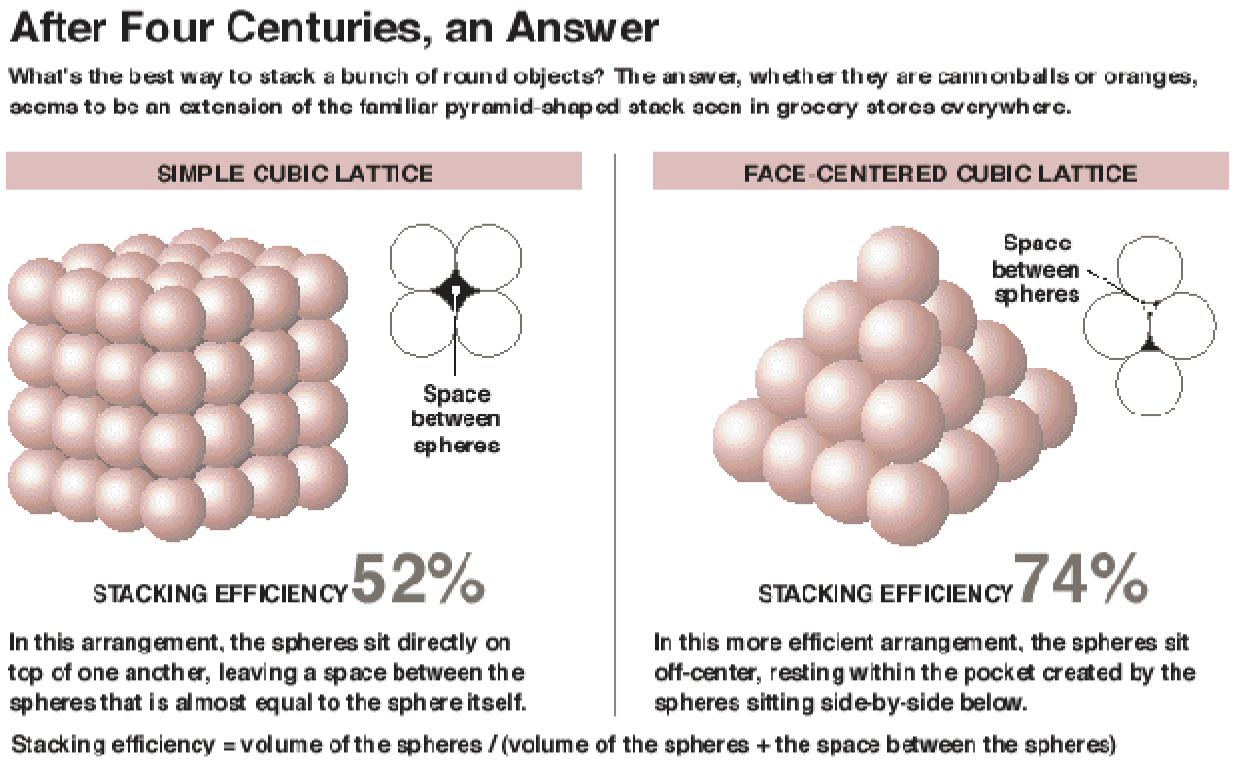
Other Well Known Sphere Arrangements

- Simple cubic packing
  - 8 nbrs
  - 52% efficiency
- bcc cubic packing
  - one sphere sits in middle of 8 others (body-centered)
  - 8 nbrs
  - 68% efficiency
- fcc -> bcc -> simple
  - apx 3/4, 2/3, 1/2





Optimal Packing Finally Proved



After Four Centuries, an Answer

What's the best way to stack a bunch of round objects? The answer, whether they are cannonballs or oranges, seems to be an extension of the familiar pyramid-shaped stack seen in grocery stores everywhere.

SIMPLE CUBIC LATTICE

STACKING EFFICIENCY 52%

In this arrangement, the spheres sit directly on top of one another, leaving a space between the spheres that is almost equal to the sphere itself.

FACE-CENTERED CUBIC LATTICE

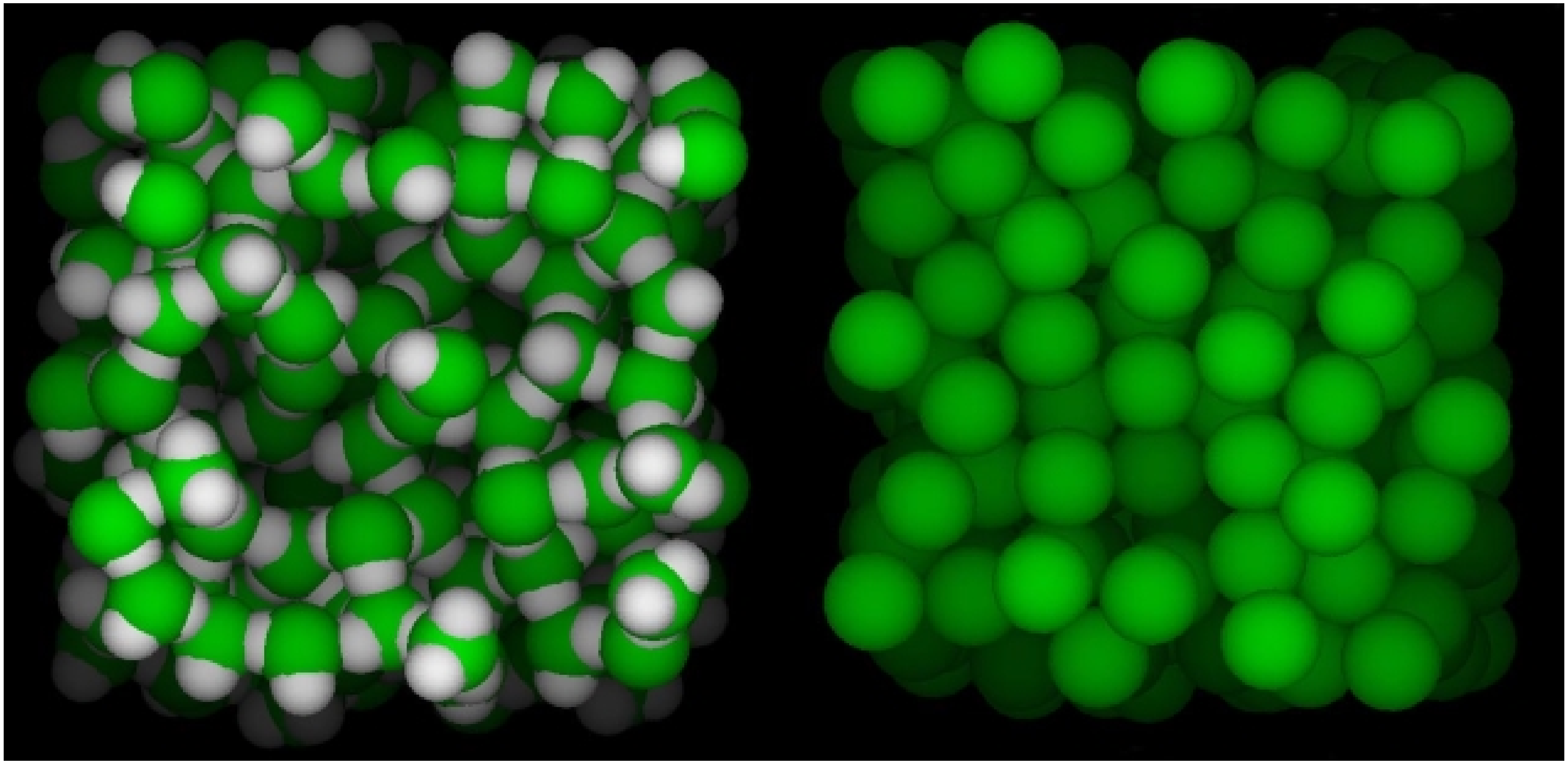
STACKING EFFICIENCY 74%

In this more efficient arrangement, the spheres sit off-center, resting within the pocket created by the spheres sitting side-by-side below.

Stacking efficiency = volume of the spheres / (volume of the spheres + the space between the spheres)

Illustration Credits: Singh, New York Times

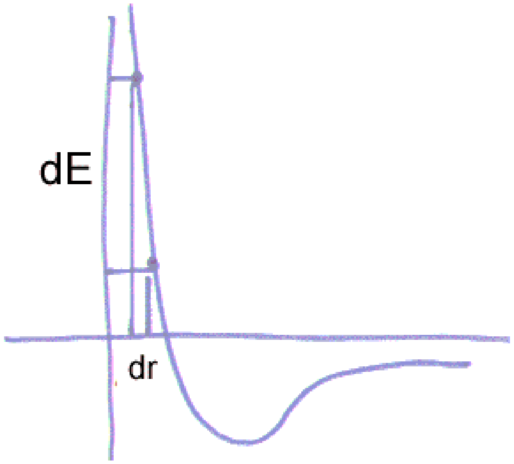
Water v. Argon

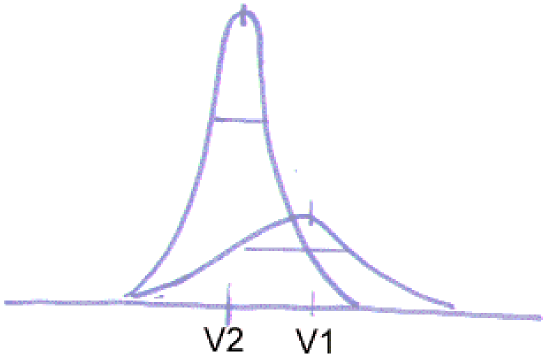


More Complex Systems -- what to do?

Small Packing Changes Significant

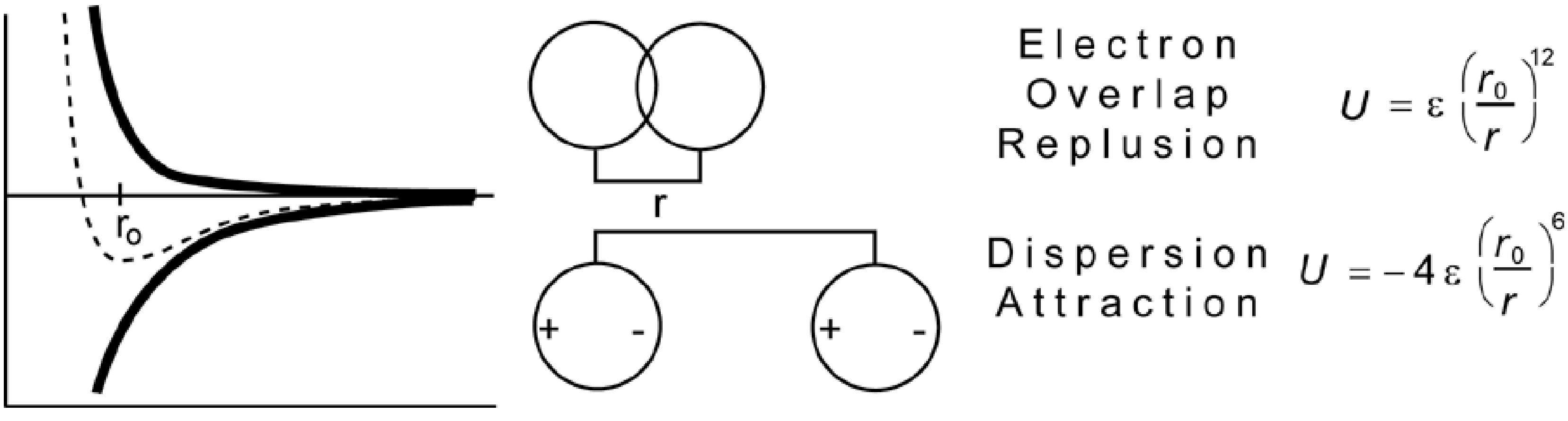
- Exponential dependence
- Bounded within a range of 0.5 (.8 and .3)
- Many observations in standard volumes gives small error about the mean (SD/sqrt(N))





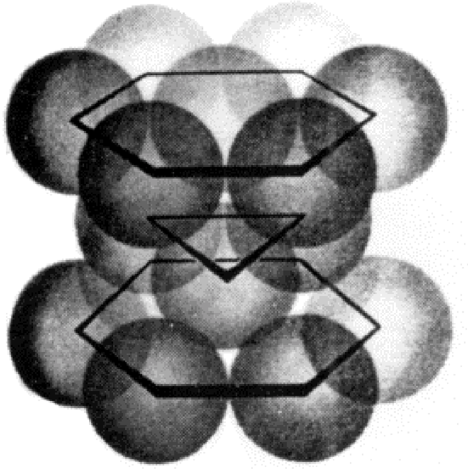
Packing ~ VDW force

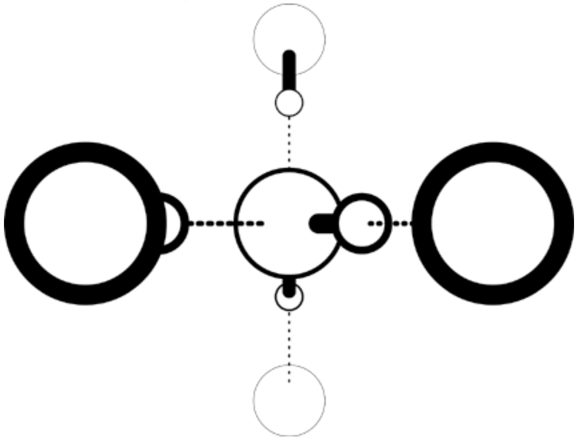
- Longer-range isotropic attractive tail provides general cohesion
- Shorter-ranged repulsion determines detailed geometry of interaction
- Billiard Ball model, WCA Theory


r_o
r
+
-
+
-
Electron Overlap Replusion
$U = \varepsilon \left(\frac{r_0}{r}\right)^{12}$
Dispersion Attraction
$U = -4\varepsilon \left(\frac{r_0}{r}\right)^{6}$

Close-packing is Default

- No tight packing when highly directional interactions (such as H-bonds) need to be satisfied
- Packing spheres (.74), hexagonal
- Water (~.35), "Open" tetrahedral, H-bonds





Standard Radii & Volumes

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
    - ◊ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
    - ◊ Distance Matrix based methods
    - ◊ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity

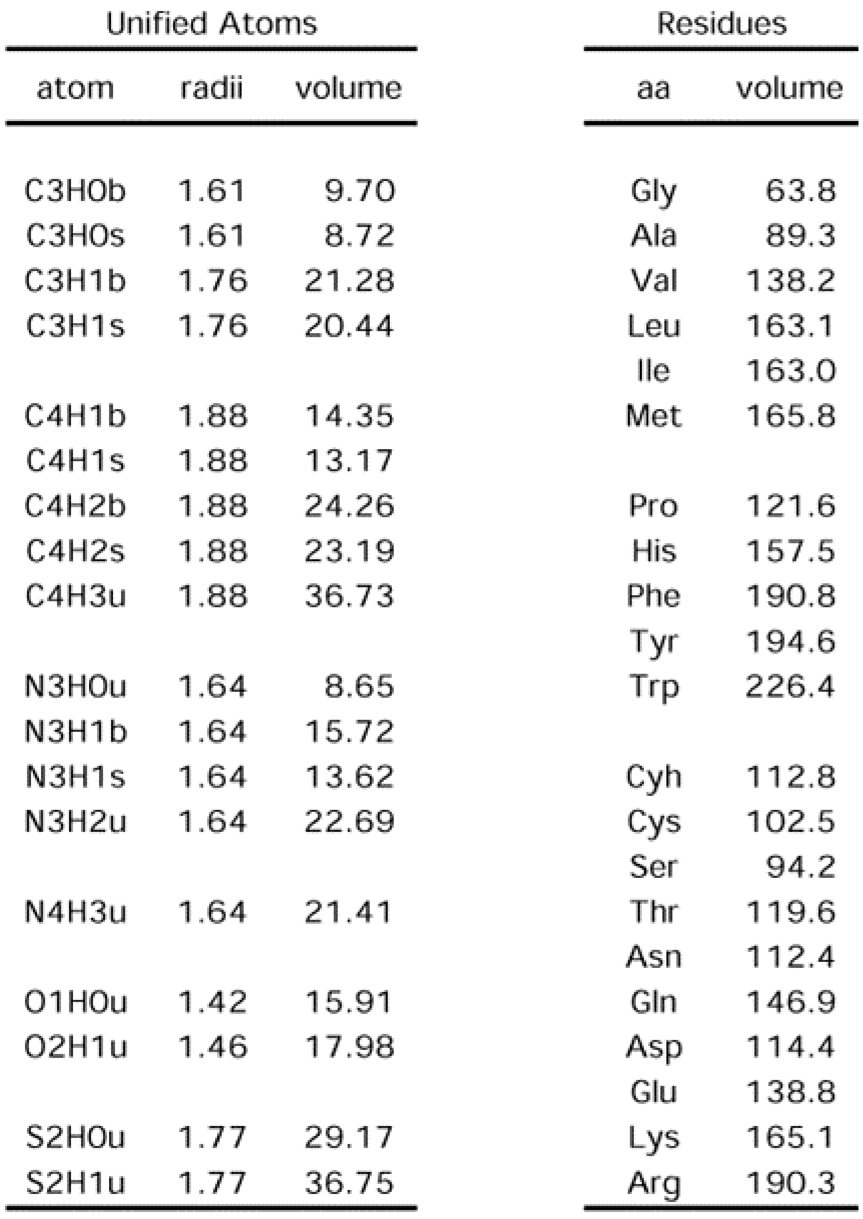
- Protein Geometry
- Surfaces I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking

Different Sets of Radii

Atom Type & Symbol		Bondi 1968	Lee & Richards 1971	Shrake & Rupley 1973	Richards 1974	Chothia 1975	Richmond & Richards 1978	Gelin & Karplus 1979	Dunfield et al. 1979	ENCAD derived 1995	CHARMM derived 1995	Tsai et al. 1998
-CH$_3$	Aliphatic, methyl	2.00	1.80	2.00	2.00	1.87	1.90	1.95	2.13	1.82	1.88	1.88
-CH$_2$-	Aliphatic, methyl	2.00	1.80	2.00	2.00	1.87	1.90	1.90	2.23	1.82	1.88	1.88
>CH-	Aliphatic, CH	-	1.70	2.00	2.00	1.87	1.90	1.85	2.38	1.82	1.88	1.88
=CH	Aromatic, CH	-	1.80	1.85	*	1.76	1.70	1.90	2.10	1.74	1.80	1.76
>C=	Trigonal, aromatic	1.74	1.80	*	1.70	1.76	1.70	1.80	1.85	1.74	1.80	1.61
-NH$_3$+	Amino, protonated	-	1.80	1.50	2.00	1.50	0.70	1.75		1.68	1.40	1.64
-NH$_2$	Amino or amide	1.75	1.80	1.50	-	1.65	1.70	1.70		1.68	1.40	1.64
>NH	Peptide, NH or N	1.65	1.52	1.40	1.70	1.65	1.70	1.65	1.75	1.68	1.40	1.64
=O	Carbonyl Oxygen	1.50	1.80	1.40	1.40	1.40	1.40	1.60	1.56	1.34	1.38	1.42
-OH	Alcoholic hydroxyl	-	1.80	1.40	1.60	1.40	1.40	1.70		1.54	1.53	1.46
-OM	Carboxyl Oxygen	-	1.80	1.89	1.50	1.40	1.40	1.60	1.62	1.34	1.41	1.42
-SH	Sulfhydryl	-	1.80	1.85	-	1.85	1.80	1.90		1.82	1.56	1.77
-S-	Thioether or –S-S-	1.80	-	-	1.80	1.85	1.80	1.90	2.08	1.82	1.56	1.77

ProtOr Parameter Set

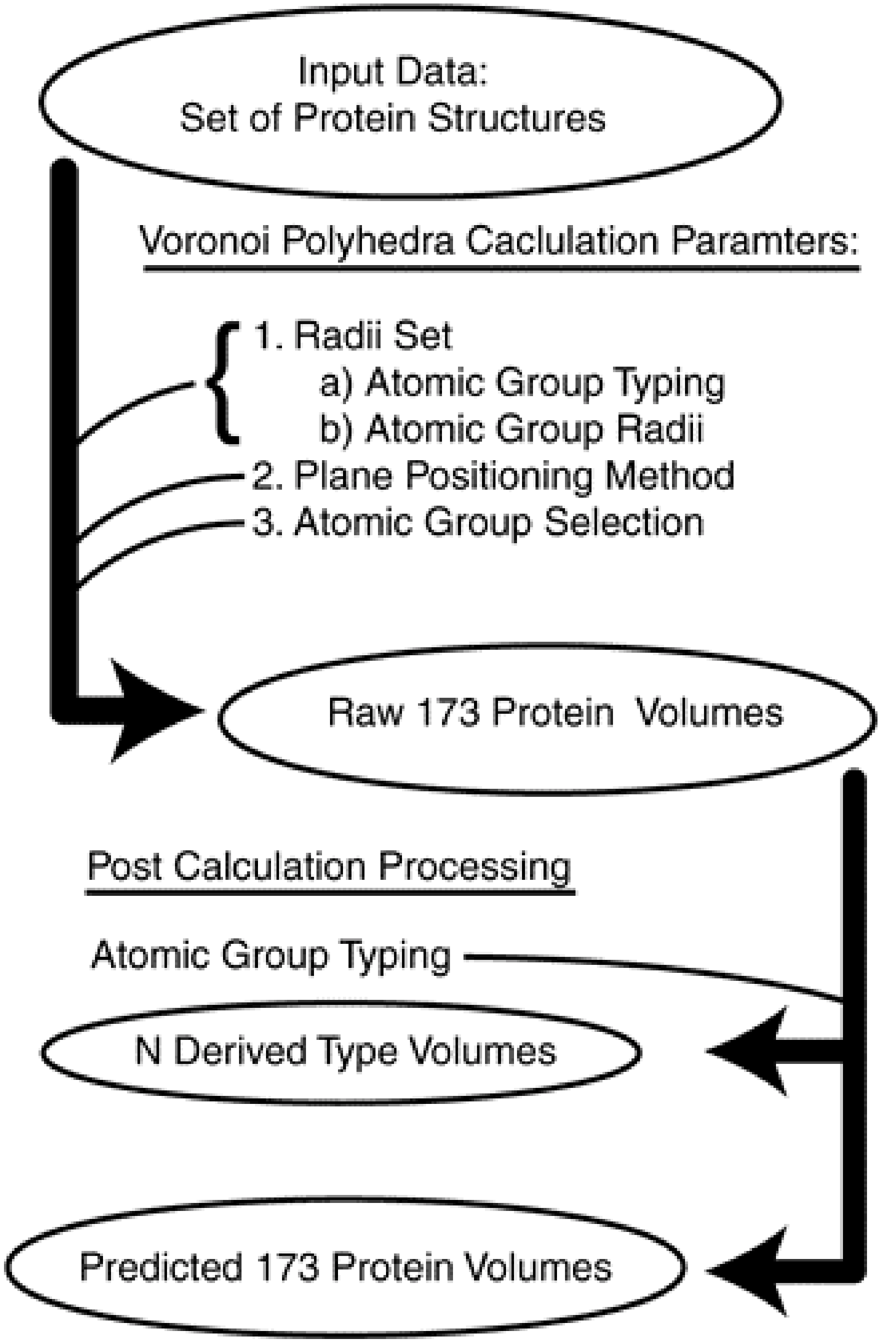
- Consistent Radii, Typing, and Volumes for Packing Calculations



Unified Atoms		
atom	radii	volume
C3H0b	1.61	9.70
C3H0s	1.61	8.72
C3H1b	1.76	21.28
C3H1s	1.76	20.44
C4H1b	1.88	14.35
C4H1s	1.88	13.17
C4H2b	1.88	24.26
C4H2s	1.88	23.19
C4H3u	1.88	36.73
N3H0u	1.64	8.65
N3H1b	1.64	15.72
N3H1s	1.64	13.62
N3H2u	1.64	22.69
N4H3u	1.64	21.41
O1H0u	1.42	15.91
O2H1u	1.46	17.98
S2H0u	1.77	29.17
S2H1u	1.77	36.75

Residues	
aa	volume
Gly	63.8
Ala	89.3
Val	138.2
Leu	163.1
Ile	163.0
Met	165.8
Pro	121.6
His	157.5
Phe	190.8
Tyr	194.6
Trp	226.4
Cyh	112.8
Cys	102.5
Ser	94.2
Thr	119.6
Asn	112.4
Gln	146.9
Asp	114.4
Glu	138.8
Lys	165.1
Arg	190.3

Factors Affecting Volume Calculations



Input Data:
Set of Protein Structures

Voronoi Polyhedra Caclulation Paramters:

1. Radii Set
   a) Atomic Group Typing
   b) Atomic Group Radii
2. Plane Positioning Method
3. Atomic Group Selection

Raw 173 Protein Volumes

Post Calculation Processing

Atomic Group Typing

N Derived Type Volumes

Predicted 173 Protein Volumes

Parameters used in Protor Volume Derivation

Typing Scheme	Hybrid chemical and numerical typing with 18 basic types
Radii Set	ProtOr Radii, Tsai et al. (1999)
Plane-Positioning Method	Ratio
Atom Selection Criteria	BL+
Structure Set	SCOP (87 structures)

Set of VDW Radii

- Great differences in a sensitive parameter (Radii for carbon 1.87 vs 2.00)
- Complex calculation: minimizing SD, iterative procedure, from protein structures
- Look for common distances in CCD
- Preliminary Solution

Atom	Bondi	New
C4__	1.87	1.88
C3H1	1.76	1.76
C3H0	1.76	1.61
O1HO	1.40	1.42
O2H1	1.40	1.46
N___	1.65	1.64
S___	1.85	1.77

Standard Residue Volumes

- Database of many hi-res structures (~100, 2 Å)
- Volumes statistics for buried residues (various selections, resample, &c)
- Standard atomic volumes harder… parameter set development...

G	64	c	105	T	120	V	139	H	159	M	168	R	194
A	90	C	113	P	124	E	140	L	165	K	170	Y	198
S	94	D	117	N	128	N	150	I	165	F	193	W	233

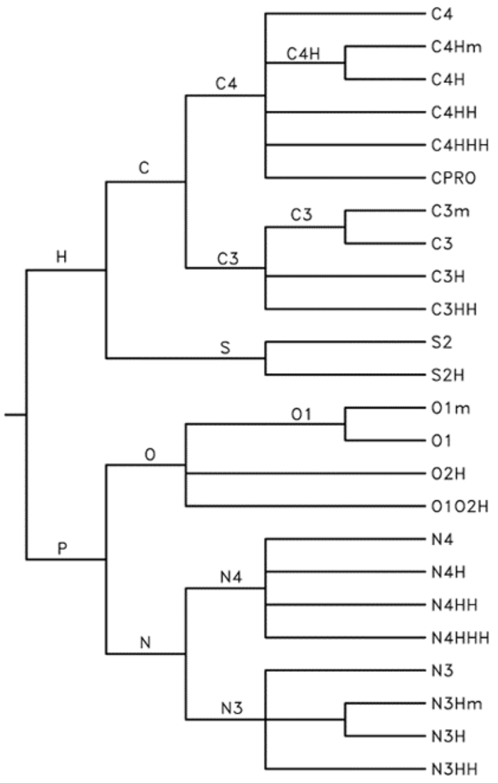
Standard Core Volumes (Prelim.)

Atom Types		Num.	Volume ($\AA^3$)	Error (%)
Mainchain Atoms				
carbonyl carbon (except G)	C	8361	9.2	.08
alpha carbon (except G)	CA	7686	13.4	.09
nitrogen (except P)	N	9042	13.9	.09
carbonyl oxygen	O	7831	15.8	.10
Gly C		811	10.2	.27
Gly CA		522	23.5	.39
Pro N		334	8.6	.39
Sidechain atoms				
trigonal or aromatic carbon	>C=	3026	10.3	.13
aromatic CH (H,F,W,Y)	-CH=	4333	21.1	.14
aliphatic CH	>CH-	3411	14.6	.14
methylene group	-CH2-	5427	23.7	.12
methyl group (A,V,L,I)	–CH3	5273	36.7	.11
hydroxyl oxygen (S,T)	–OH	851	17.2	.36
carbonyl oxygen (N,Q)	=O	272	16.8	.76
carboxyl oxygen (D,E)	–O	517	16.0	.53
2° amine (R,H,W)	-NH-	530	15.6	.53
1° amine or amide (R,N,Q)	–NH2	355	23.4	.52
tetrahedral nitrogen (K)	–NH3	31	20.0	1.40
thioether or disulfide (C,M)	-S-	1242	19.3	1.22
sulfhydryl (C)	–SH	67	37.8	1.33

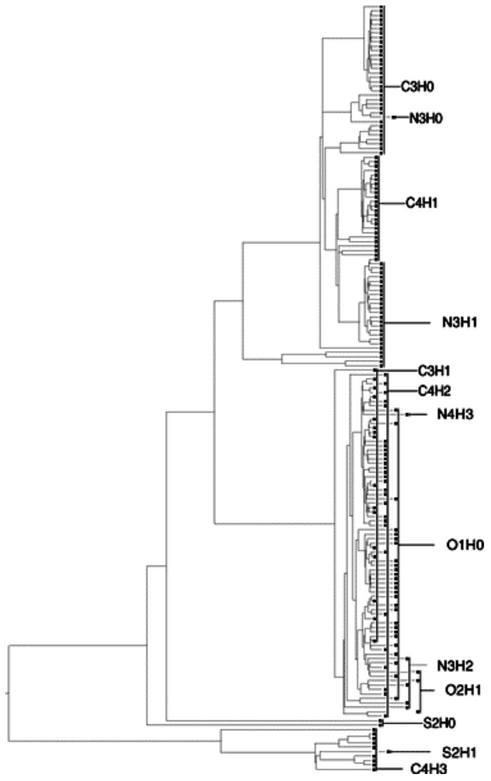
Clustering into a set of Atom Types I

- Which atoms are equivalent? How many types valid?
- 18 types, [CNOS][34]H[123][bsu]

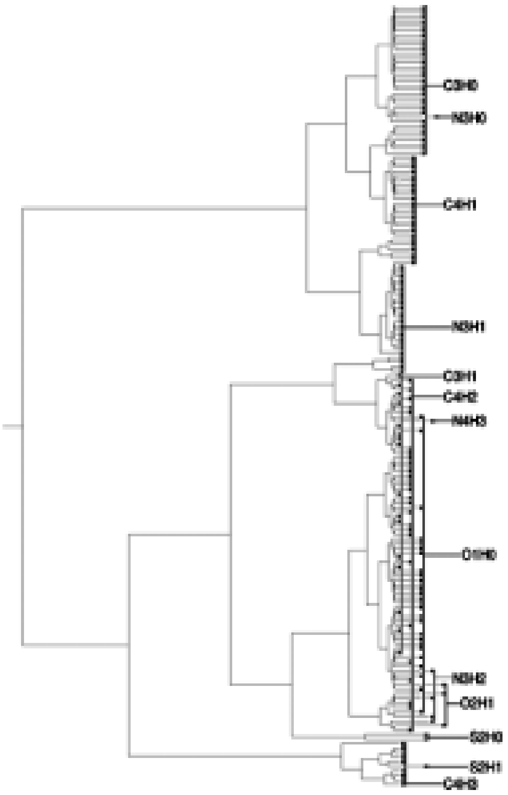
Chemical



Single-link

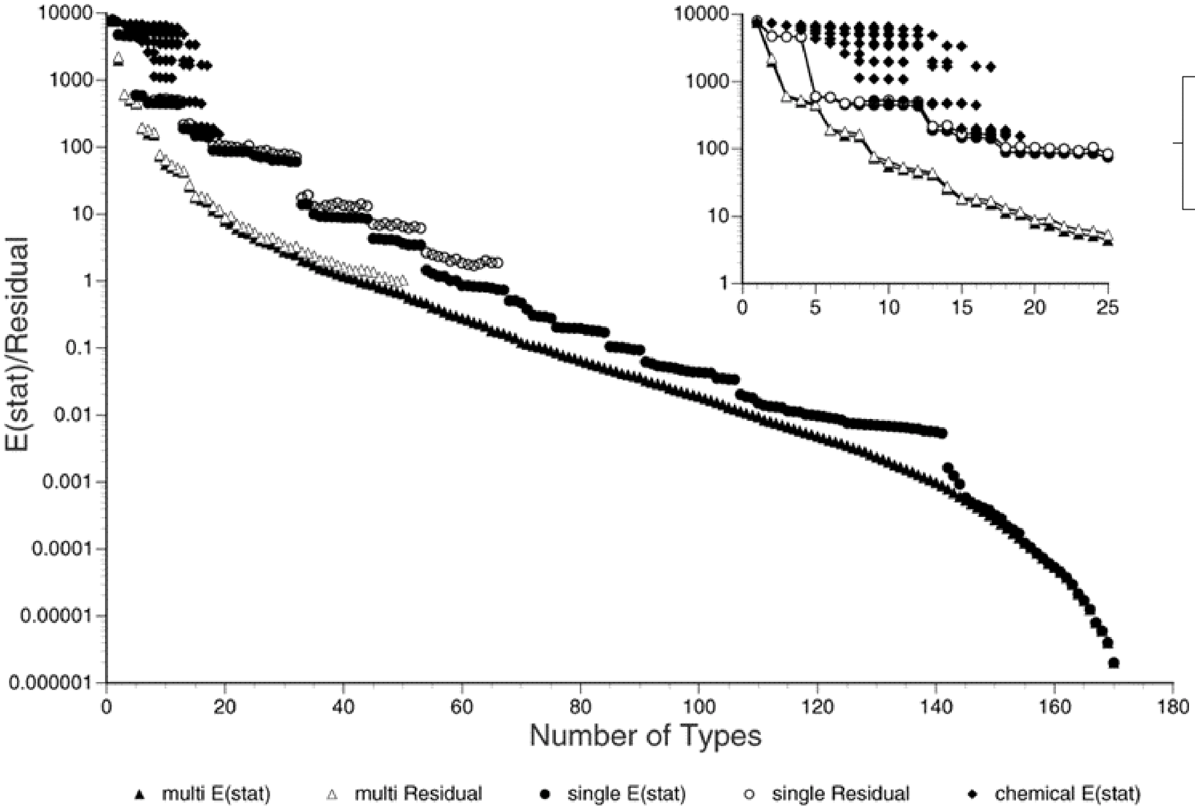


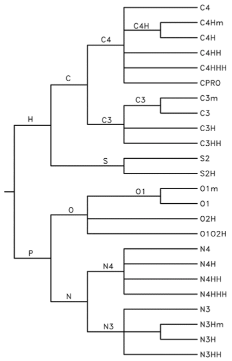
Multi-link

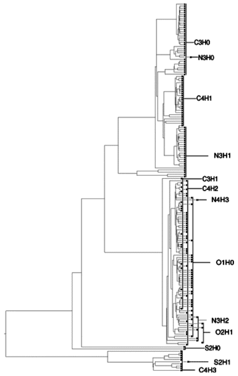


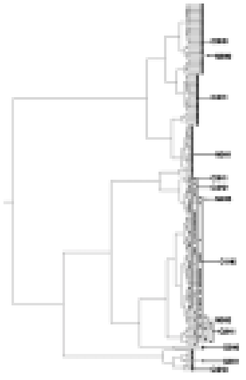
Clustering into a set of Atom Types II

- Which atoms are equivalent? How many types valid?
- 18 types, [CNOS][34]H[123][bsu]
- E statistic to tell apart







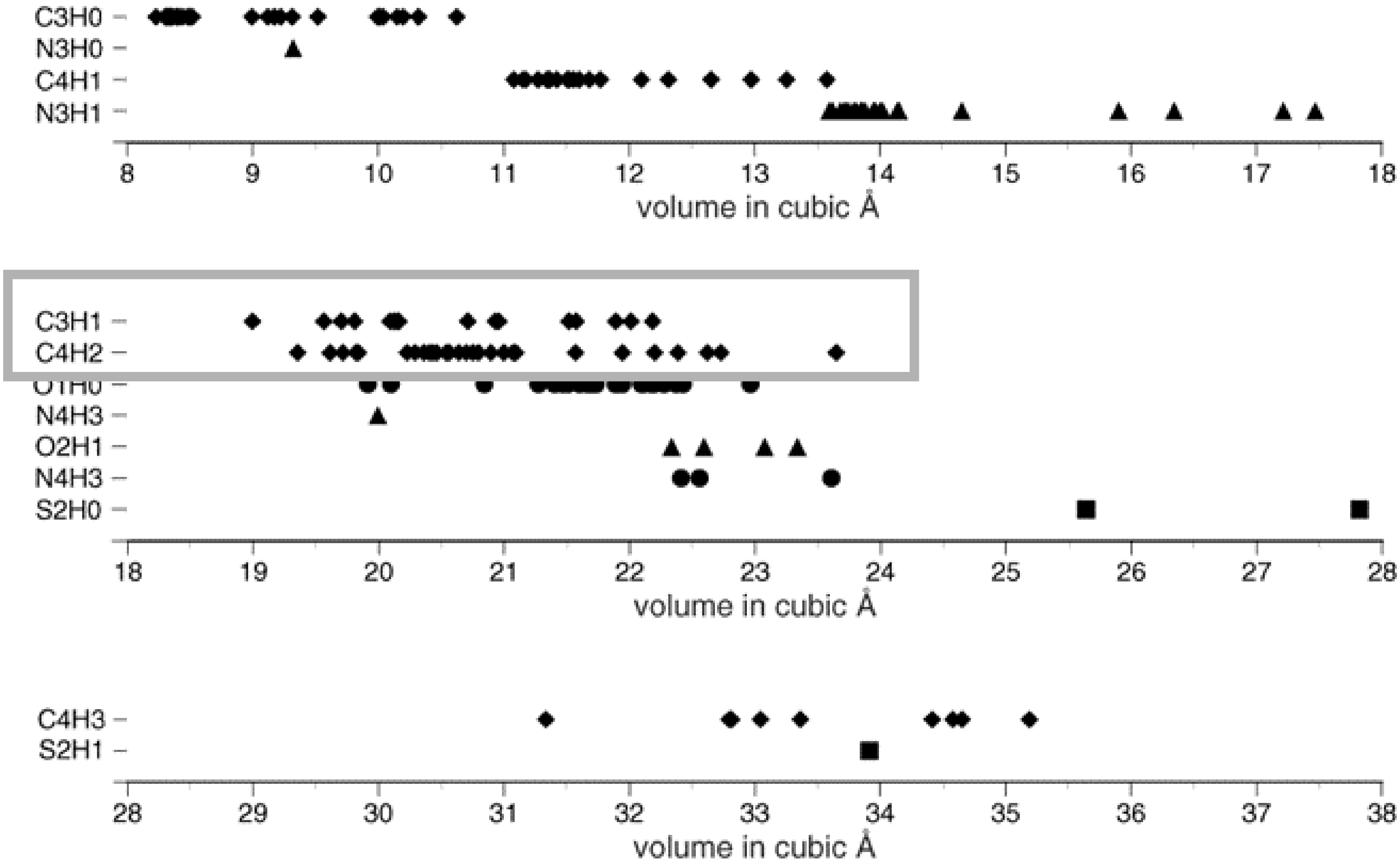


Compare Different Structure Sets

ProtOr atom type	PDB Sets[a]													
	SCOP		Standard		High		Low		NMR		New		Obsolete	
	Vol.[b]	SD	Vol.[b]	SD	Vol.[b]	SD	Vol.[b]	SD	Vol.[b]	SD	Vol.[b]	SD	Vol.[b]	SD
C3H0b	9.64	0.72	9.67	0.68	9.65	0.68	9.68	0.69	9.53	1.05	9.78	0.79	9.83	0.86
C3H0s	8.66	0.58	8.68	0.59	8.65	0.57	8.70	0.60	8.65	0.80	8.77	0.69	8.84	0.76
C3H1b	21.33	1.87	21.38	1.89	21.36	1.85	21.39	1.91	19.40	2.73	21.26	2.11	20.96	2.30
C3H1s	20.45	1.76	20.41	1.77	20.27	1.72	20.50	1.80	18.48	2.78	20.42	2.02	20.43	2.21
C4H3u	14.35	1.35	14.41	1.22	14.38	1.20	14.43	1.23	13.89	1.55	14.40	1.48	14.42	1.59
C4H1b	13.14	0.97	13.17	0.96	13.20	0.94	13.15	0.97	13.20	1.27	13.11	1.11	13.18	1.20
C4H1s	24.14	2.07	24.25	2.13	24.11	1.95	24.33	2.21	20.48	5.89	24.26	2.43	24.07	2.76
C4H2b	23.17	2.35	23.29	1.94	23.28	1.96	23.29	1.93	19.13	6.40	23.14	2.23	22.92	2.46
C4H2s	36.84	3.24	36.94	2.99	36.93	3.00	36.94	2.98	30.38	8.26	36.43	3.75	35.76	3.95
N3H0u	8.62	0.59	8.57	0.65	8.60	0.70	8.56	0.6						
N3H1b	15.65	1.55	15.73	1.70	15.55	1.48	15.80	1.7						
N3H1s	13.54	0.99	13.53	1.00	13.52	0.97	13.53	1.0						
N3H2u	22.61	2.36	22.07	2.13	22.12	2.22	22.04	2.0						
N4H3u	21.56	1.28	21.03	1.29	20.30	0.55	21.76	1.40						
O1H0u	15.91	1.29	15.92	1.28	15.87	1.23	15.94	1.30						
O2H1u	18.11	1.78	18.09	1.86	18.10	1.97	18.09	1.7						
S2H0u	29.29	2.68	28.79	2.67	28.66	2.68	28.90	2.66						
S2H1u	36.82	3.48	35.93	2.44	37.15	2.46	35.71	2.38						

Set	Number	Identifier
Standard	130	135l, 1aaj, 1aap, 1ake, 1arb, 1bbh, 1bp2, 1ccr, 1cdp, 1cmb, 1cpc, 1crn, 1cse, 1ctf, 1cus, 1dfn, 1dr1, 1eco, 1ezm, 1fkf, 1fus, 1fxd, 1gct, 1gd1, 1gpr, 1hbg, 1hel, 1hne, 1ifc, 1igd, 1lmb, 1lz1, 1lz3, 1mba, 1mbd, 1ofv, 1omd, 1paz, 1pgx, 1pk4, 1plc, 1ppn, 1ppt, 1ptx, 1rcf, 1rdg, 1rms, 1rop, 1rpg, 1rpo, 1rro, 1sar, 1sgt, 1snc, 1st3, 1thm, 1ubq, 1ycc, 256b, 2act, 2alp, 2apr, 2aza, 2cba, 2ccy, 2cdv, 2cpp, 2ctc, 2cyp, 2er7, 2fb4, 2fcr, 2fx2, 2gbp, 2hhb, 2ihl, 2ltn, 2mcm, 2mhr, 2msb, 2ovo, 2por, 2prk, 2rhe, 2rn2, 2sga, 2sn3, 2trx, 2utg, 2wrp, 2zta, 3app, 3b5c, 3bcl, 3c2c, 3cla, 3dfr, 3ebx, 3est, 3fxn, 3grs, 3lzm, 3rp2, 3sgb, 451c, 4dfr, 4enl, 4icb, 4ins, 4ptp, 5cpa, 5cyt, 5p21, 5pal, 5pti, 5rub, 5rxn, 5tim, 6ebx, 6rlx, 6rxn, 6xia, 7aat, 7rsa, 8dfr, 8fab, 8rxn, 9pti, 9rnt, 9wga
SCOP	87	1cbn, 1lkk, 2erl, 8rxn, 1bpi, 1ctj, 1igd, 1rge, 1amm, 1arb, 1cse, 1jbc, 2sn3, 1cus, 7rsa, 1rro, 1aac, 193l, 1utg, 5p21, 1hms, 1xyz, 256b, 2olb, 2phy, 3ebx, 3sdh, 2end, 1xso, 1cka, 1cyo, 1edm, 1ezm, 1isu, 1mla, 1poa, 1rie, 1whi, 2ctb, 2eng, 2ovo, 2cba, 3grs, 1lit, 1ra9, 1tca, 1csh, 1epn, 1mrj, 1phc, 1ptf, 1smd, 1vcc, 2dri, 2ilk, 2sil, 3pte, 4fgf, 2cpl, 1kap, 1lcp, 1php, 1snc, 1sri, 2wrp, 1krn, 2trx, 1ctf, 1fnb, 1gai, 1gof, 1knb, 1llp, 1mol, 1pdo, 1rop, 1tad, 1tfe, 1vhh, 1vsd, 2act, 1fkd, 1chd, 1kpt, 1thw, 2bbk, 3cla
NMR	125	1aab, 1aaf, 1aca, 1acp, 1afp, 1ahd, 1ale, 1alf, 1bbo, 1bus, 1bw3, 1bw4, 1cdb, 1cdn, 1cis, 1clb, 1crp, 1crq, 1crr, 1csy, 1csz, 1ctl, 1dhm, 1erg, 1erh, 1fht, 1fkr, 1fks, 1fkt, 1ftz, 1gb1, 1gbr, 1gfc, 1gfd, 1hcc, 1hdn, 1hme, 1hmf, 1hom, 1hrq, 1hrr, 1hsm, 1hsn, 1hue, 1hum, 1hun, 1il8, 1iml, 1irp, 1kb7, 1kb8, 1ldl, 1ldr, 1lip, 1lpt, 1mbe, 1mbf, 1mbg, 1mbj, 1mbk, 1mef, 1ncp, 1neh, 1neq, 1ner, 1nhm, 1nhn, 1nil, 1nim, 1nmf, 1nmg, 1noe, 1odp, 1odq, 1odr, 1oef, 1oeg, 1pan, 1pao, 1pcp, 1pdc, 1pih, 1pij, 1pmc, 1pog, 1pra, 1prr, 1prs, 1pse, 1psf, 1qwe, 1qwf, 1rht, 1rip, 1rod, 1rpv, 1san, 1sap, 1srl, 1srm, 1stu, 1sxl, 1tam, 1tiv, 1tvs, 1tvt, 1ums, 1umt, 1utr, 1vnd, 1zer, 2abd, 2bus, 2gb1, 2gva, 2gvb, 2hid, 2hmx, 2hoa, 2igg, 2igh, 2il8, 2ptl, 2znf, 3ci2
Current	69	116l, 1act, 1alp, 1alr, 1anh, 251c, 156b, 1apd, 2bcl, 1abk, 1abp, 1abx, 1afg, 1ace, 1afn, 1ak3, 1asi, 1aza, 1baa, 1bjl, 2grs, 1cab, 1cae, 1cd4, 1ci2, 1cpk, 1cln, 1dhb, 1dri, 1eip, 1end, 7atc, 1fnr, 1gap, 1gbp, 1gcr, 1gmf, 1gn5, 2hvt, 1gsr, 1gyi, 1hft, 1hid, 1hmg, 1hmx, 1lrd, 3fab, 1mev, 1omf, 1ora, 1pab, 1pel, 1pgk, 1phy, 1ptc, 1r04, 1r1e, 1rsl, 1sod, 1srt, 1tbs, 1tct, 1trt, 1yhx, 2adk, 1vaa, 1ts1, 1ada
Obsolete	69	1abe, 1cdh, 1eri, 1fnb, 1lmb, 216l, 256b, 2abk, 2abx, 2ace, 2act, 2ada, 2afg, 2afn, 2ak3, 2alp, 2alr, 2anh, 2apd, 2asi, 2aza, 2baa, 2cab, 2cae, 2ci2, 2cpk, 2cyh, 2dhb, 2dri, 2eip, 2end, 2gmf, 2gn5, 2gsr, 2gyi, 2hft, 2hid, 2hmg, 2hmx, 2mev, 2omf, 2ora, 2pab, 2pel, 2phy, 2ptc, 2r04, 2rsl, 2sod, 2srt, 2tbs,2tct, 2trt, 2ts1, 2vaa, 2yhx, 351c, 3adk, 3bcl, 3bjl, 3cln, 3gap, 3gbp, 3grs, 3hvt, 3pgk, 4gcr, 5at1, 7fab

Overlap of Volumes of Aromatic C3H1 and Aliphatic C4H2



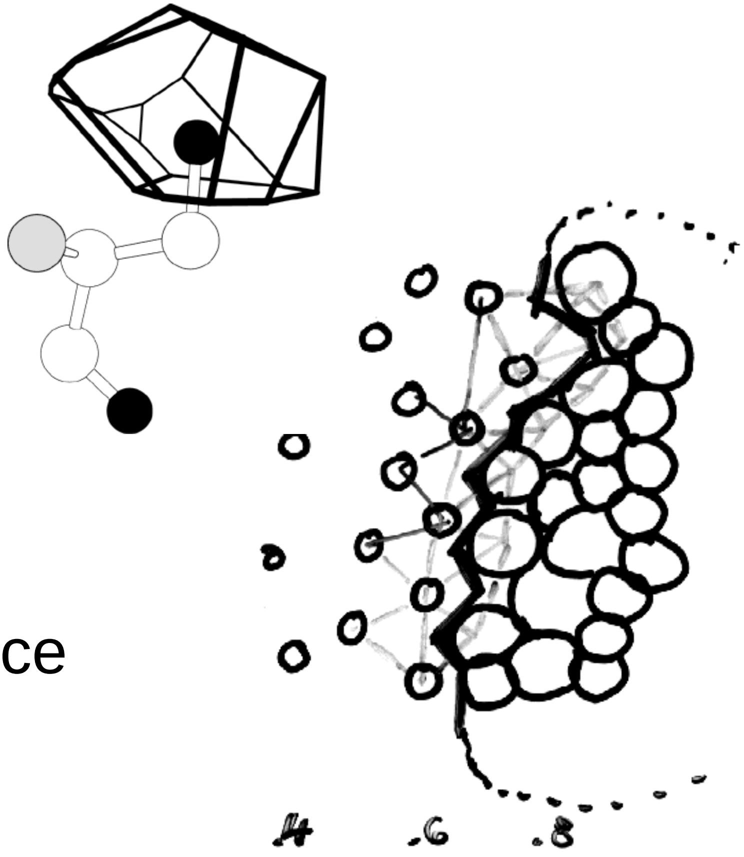
Surfaces II

- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  - ◊ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  - ◊ Distance Matrix based methods
  - ◊ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity

- Protein Geometry
- Surfaces I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking

Packing at Interfaces

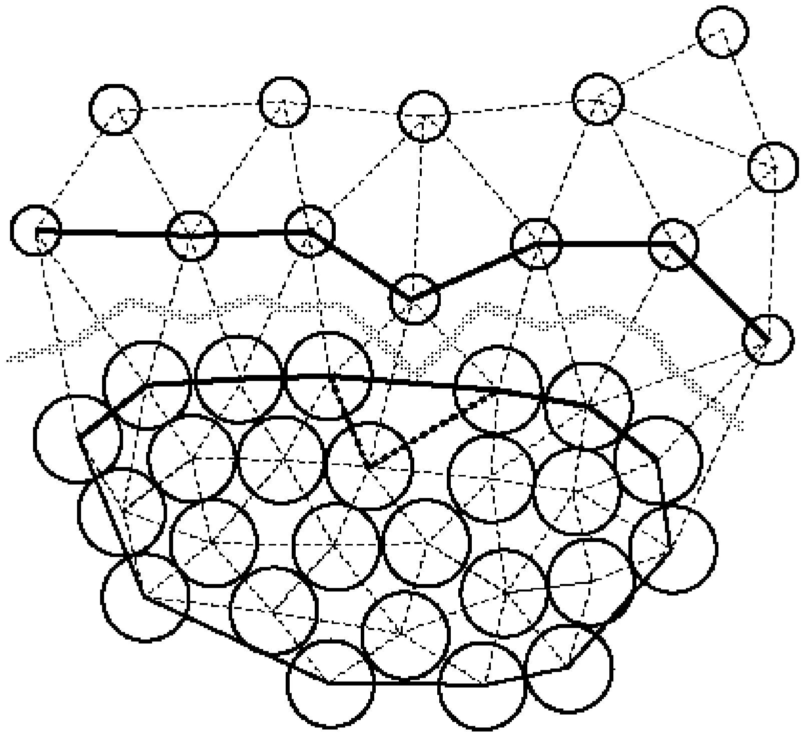
- Voronoi volumes
  (and D. triangulation)
  to measure packing
- Tight core packing v.
  Loose surface packing
- Grooves & ridges: close-packing v. H-bonding
- How packing defines a surface
  (hydration surface)
- Implications for Motions



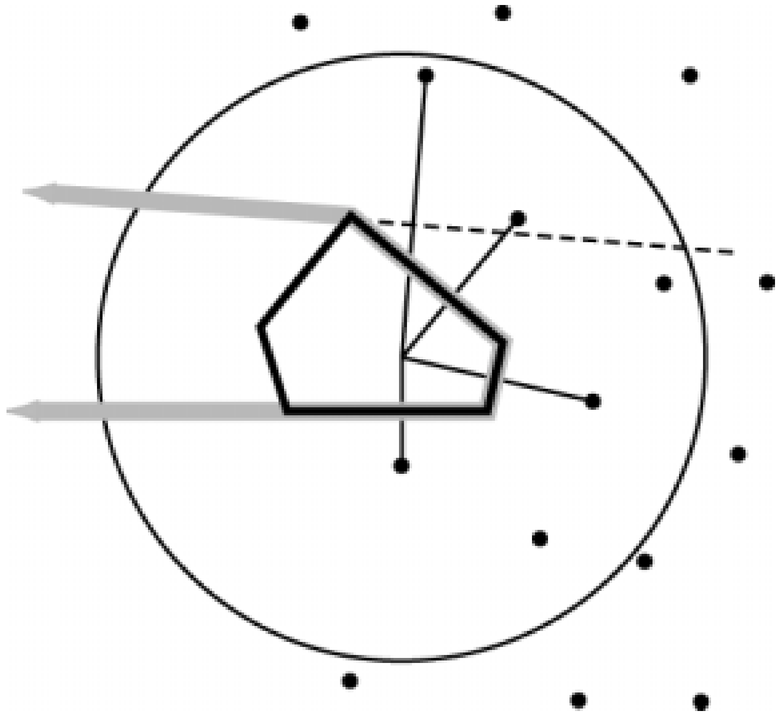
Packing defines the “Correct Definition” of the Protein Surface

- Voronoi polyhedra are the *Natural* way to study packing!
- How reasonable is a geometric definition of the surface in light of what we know about packing
- The relationship between
  - ◊ accessible surface
  - ◊ molecular surface
  - ◊ Delauney Triangulation (Convex Hull)
  - ◊ polyhedra faces
  - ◊ hydration surface

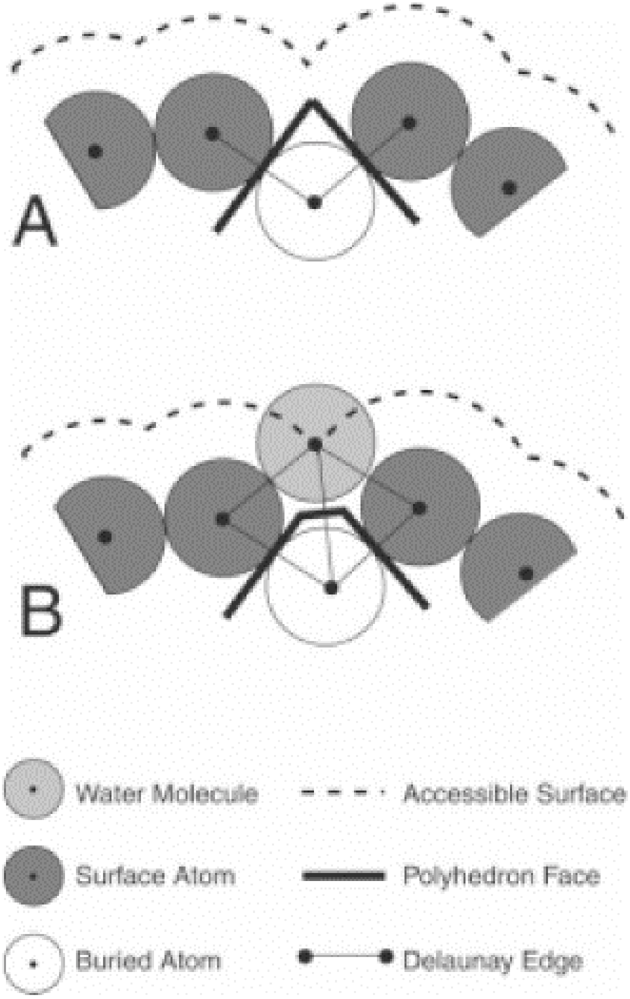
Surface and Volume Definitions Linked

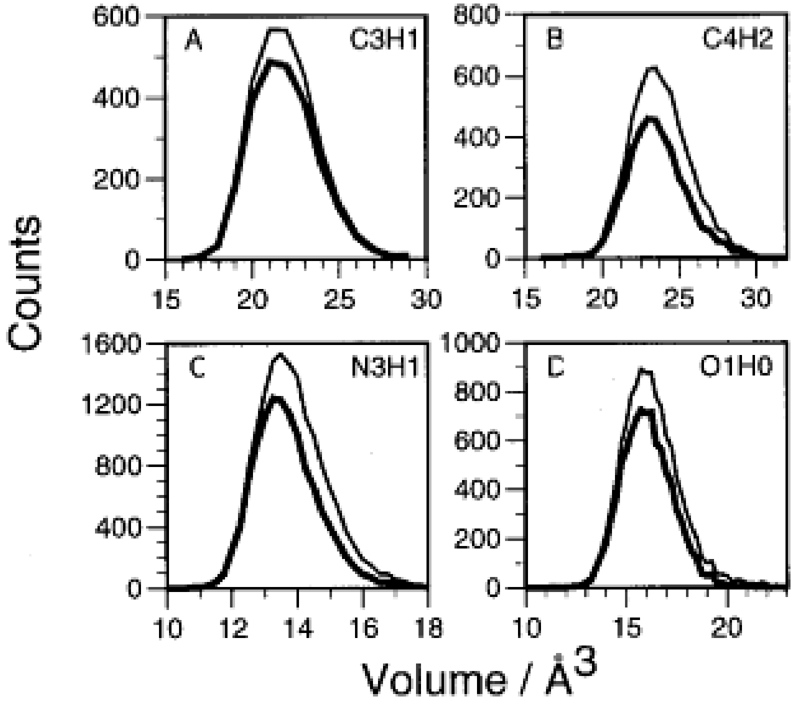


Problem of Protein Surface for Voronoi Construction



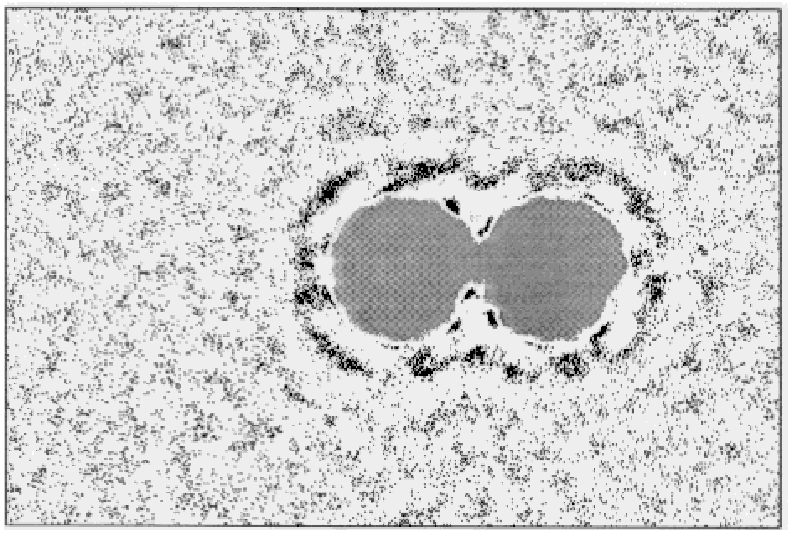
Sensitivity of Voronoi Construction to Surface Structure

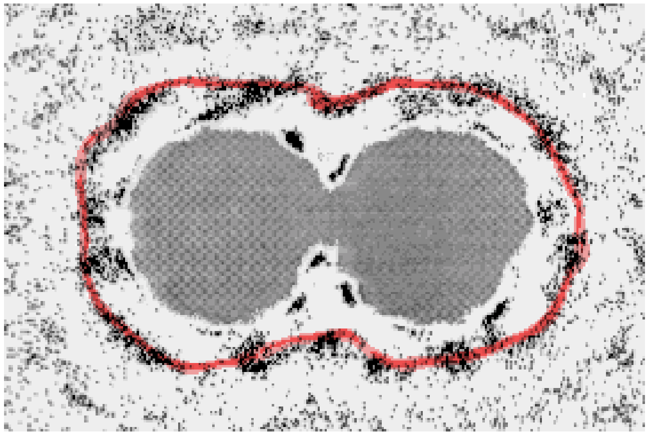




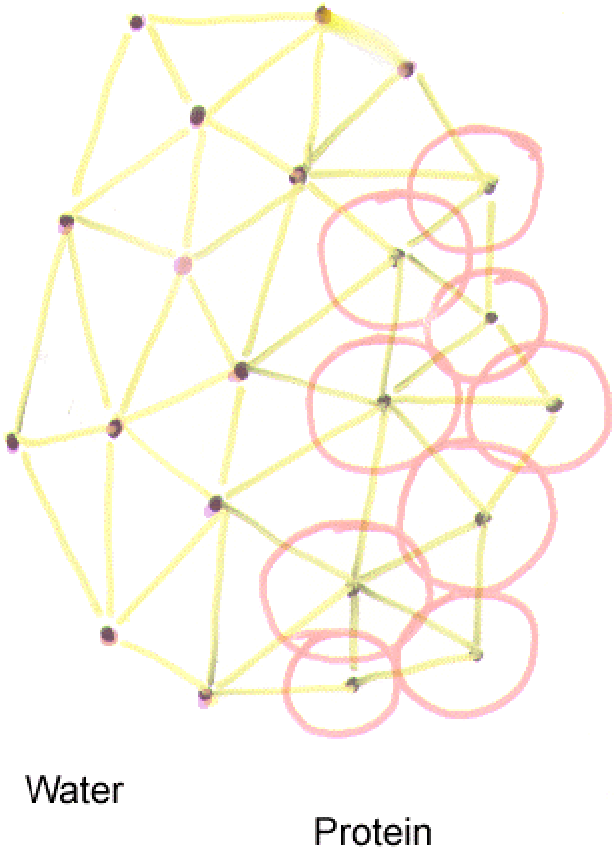
Hydration Surface

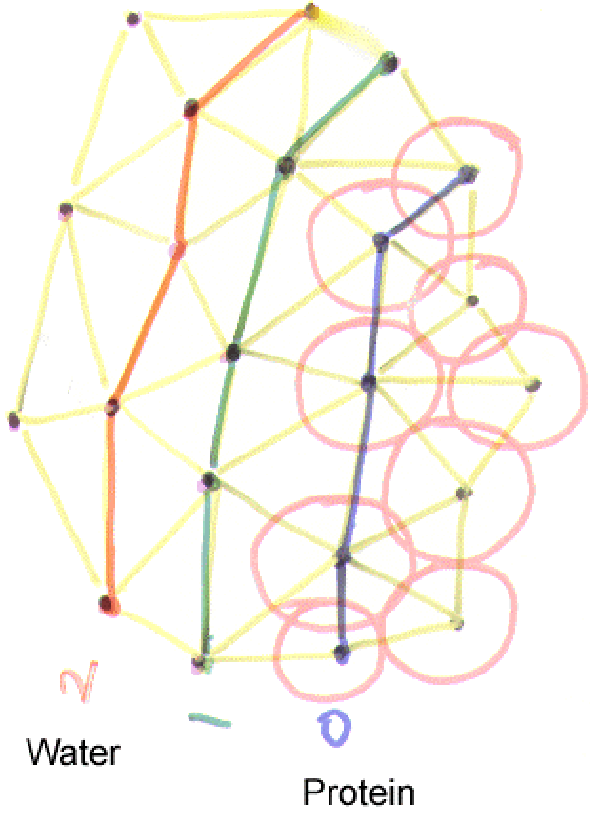
- Bring together two helices
  - ◊ Unusually low water density in grooves and crevices — especially, as compared to uncharged water
  - ◊ Fit line through second shell



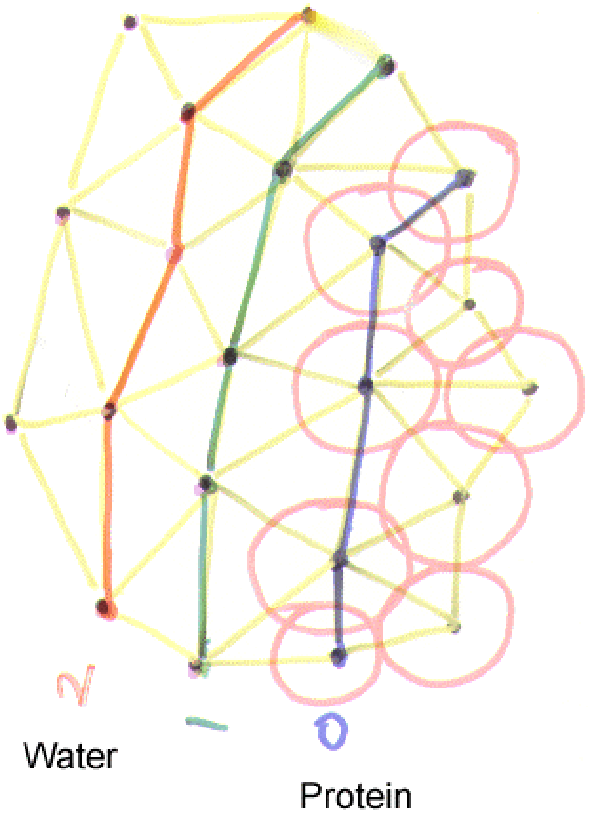


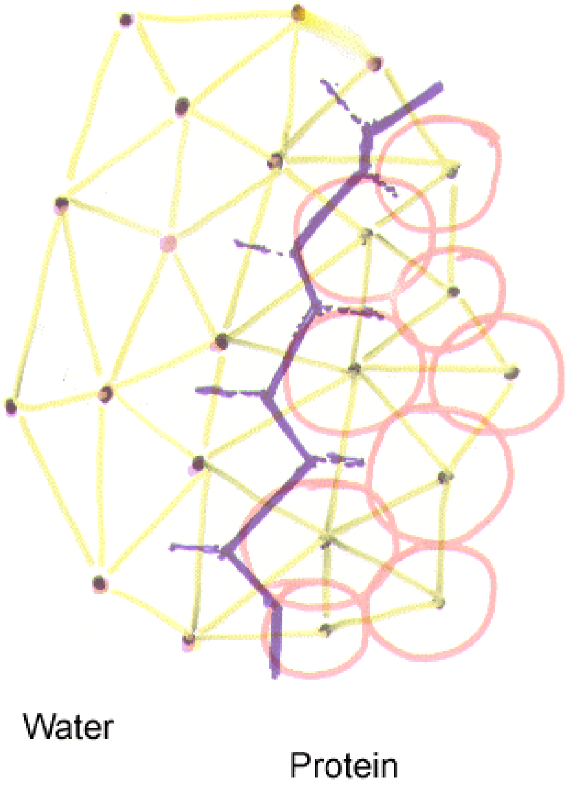
Defining Surfaces from Packing: Convex Hull and Layers of Waters



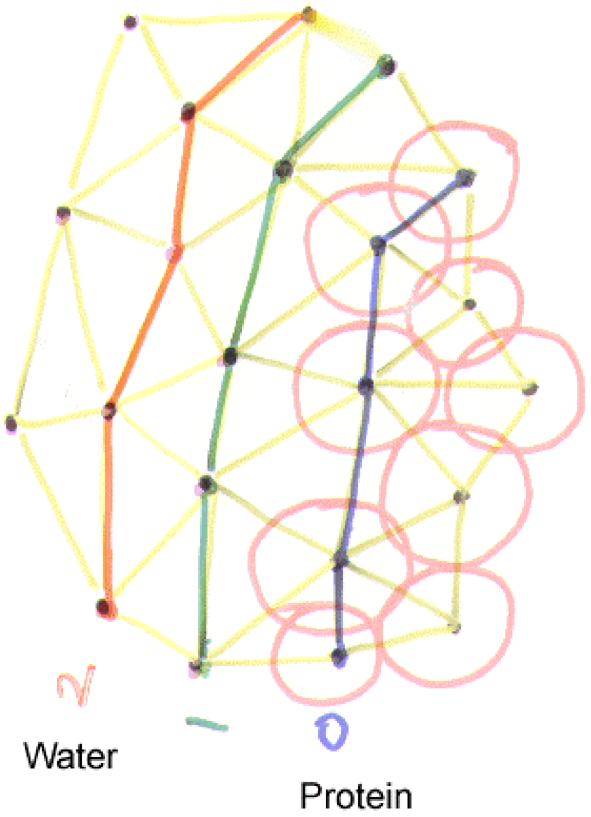


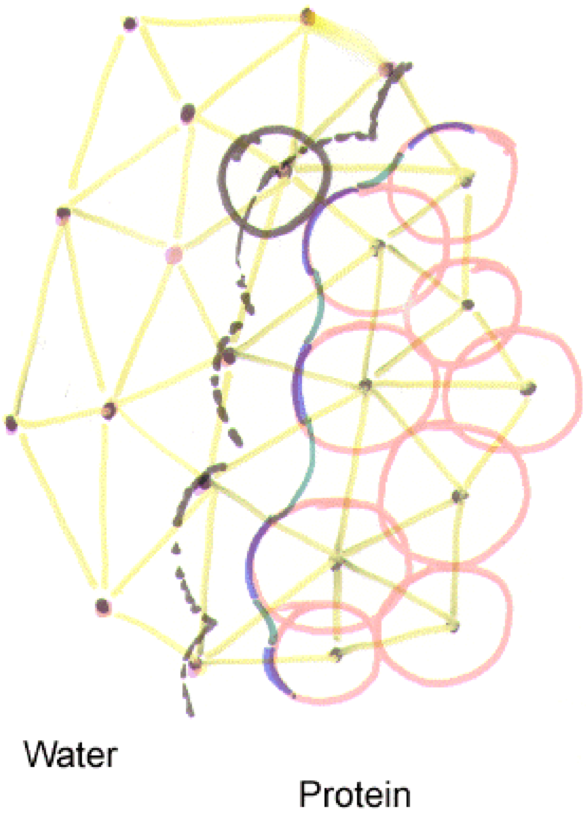
Defining a Surface from the Faces of Voronoi Polyhedra



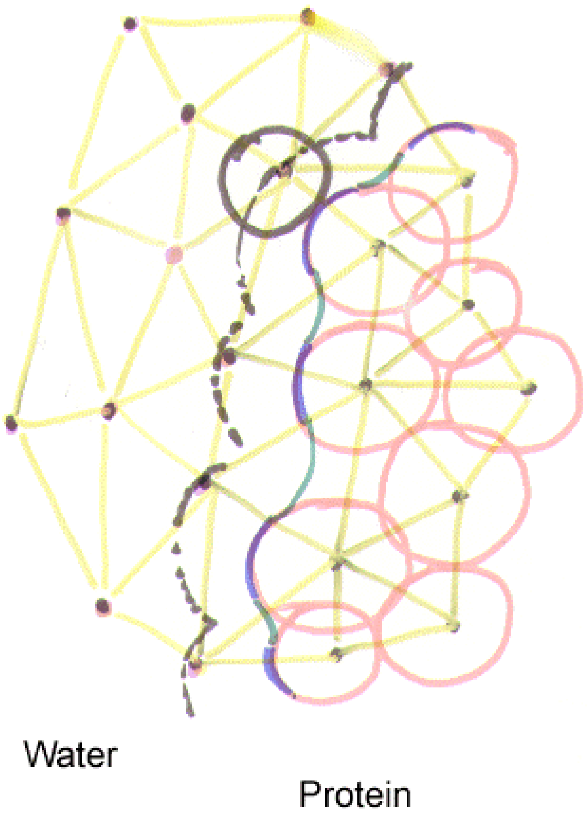


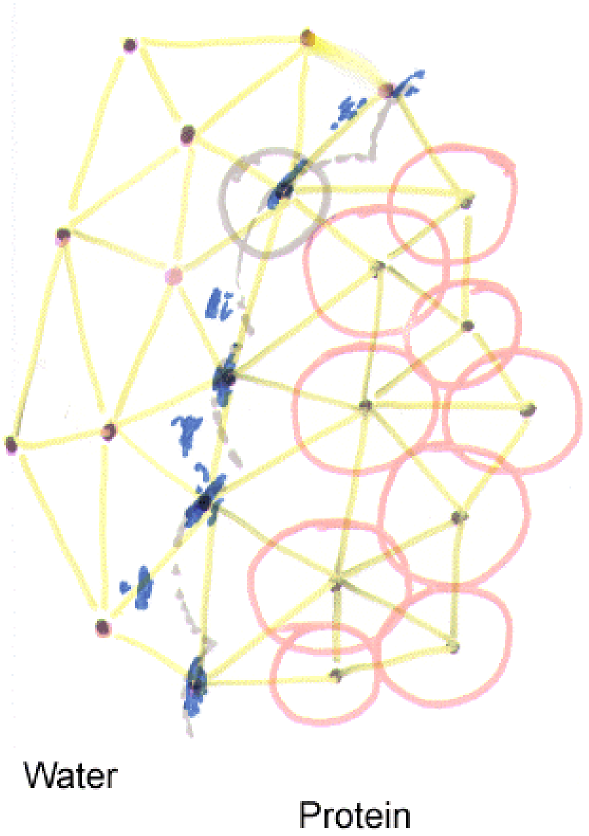
Accessible Surface as a Time-averaged Water Layer





The Hydration Surface: Trying to Model Real Water



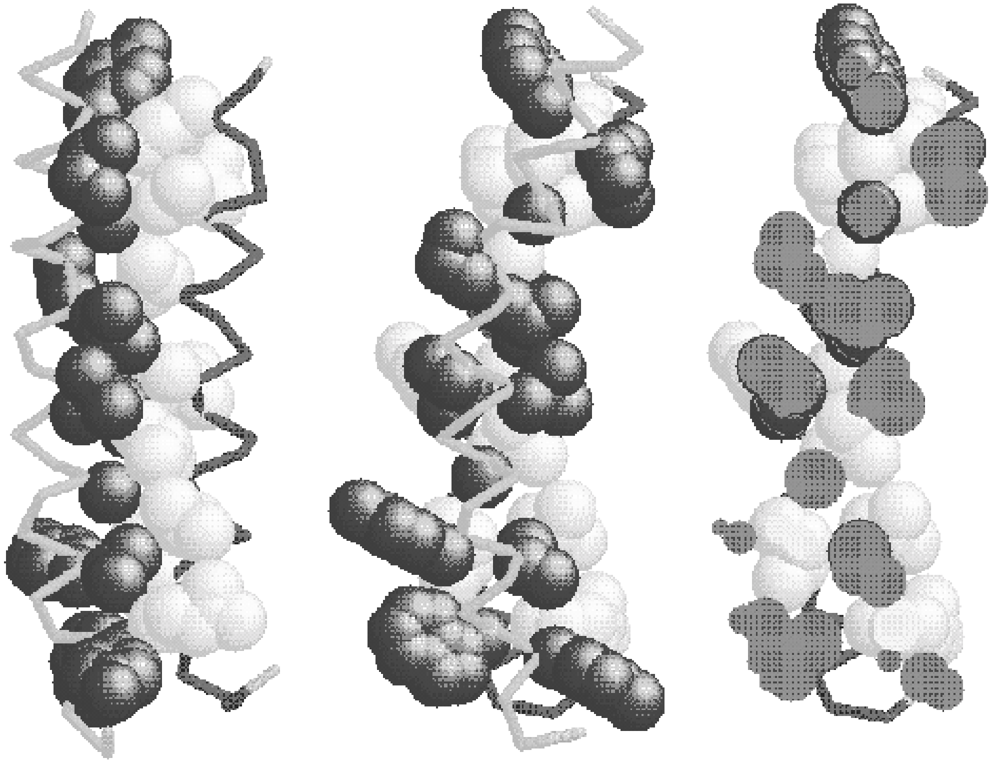


Other Applications of Volumes -- Motions, Docking

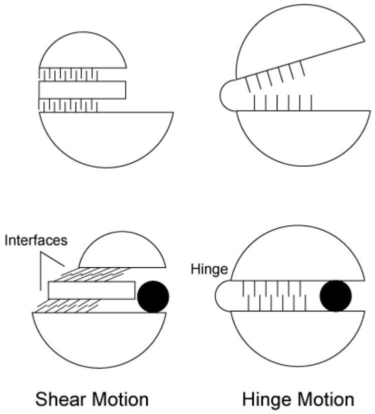
- What Structures Look Like?
- Structural Alignment by Iterated Dynamic Programming
  - ◊ RMS Superposition
- Scoring Structural Similarity
- Other Aspects of Structural Alignment
  - ◊ Distance Matrix based methods
  - ◊ Fold Library
- Relation of Sequence Similarity to Structural and Functional Similarity
- Protein Geometry
- Surfaces I (Calculation)
- Calculation of Volume
- Voronoi Volumes & Packing
- Standard Volumes & Radii
- Surfaces II (Relationship to Volumes)
- Other Applications of Volumes -- Motions, Docking

Interface Packing and Motions

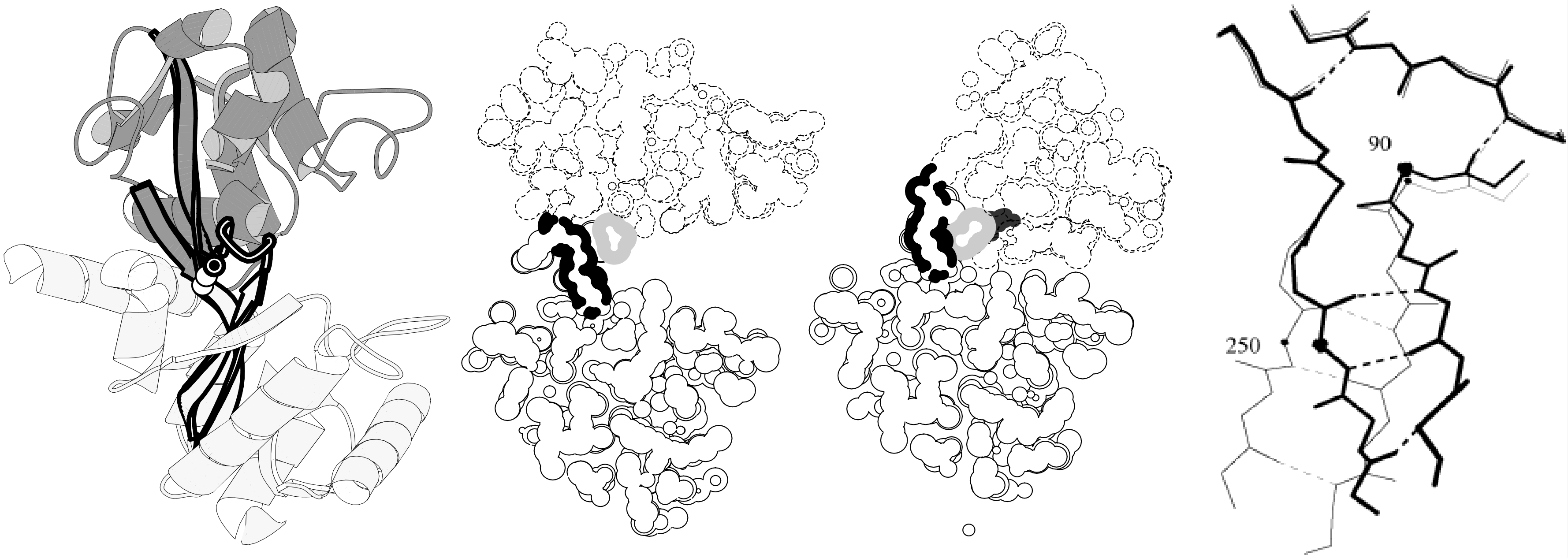
- Intercalcating Interface, Knobs into Holes
- Packing is a strong constraint on motions
  - ◊ Domain or loop motions have to be fast (~10 ps – 100 ns)
  - ◊ Can't cross big energy barriers involved in repacking an interface
- Not applicable to allosteric motions, which are much slower (~1 ms) and do involve repacking interfaces



Packing Based Classification: ***Hinge*** v Shear

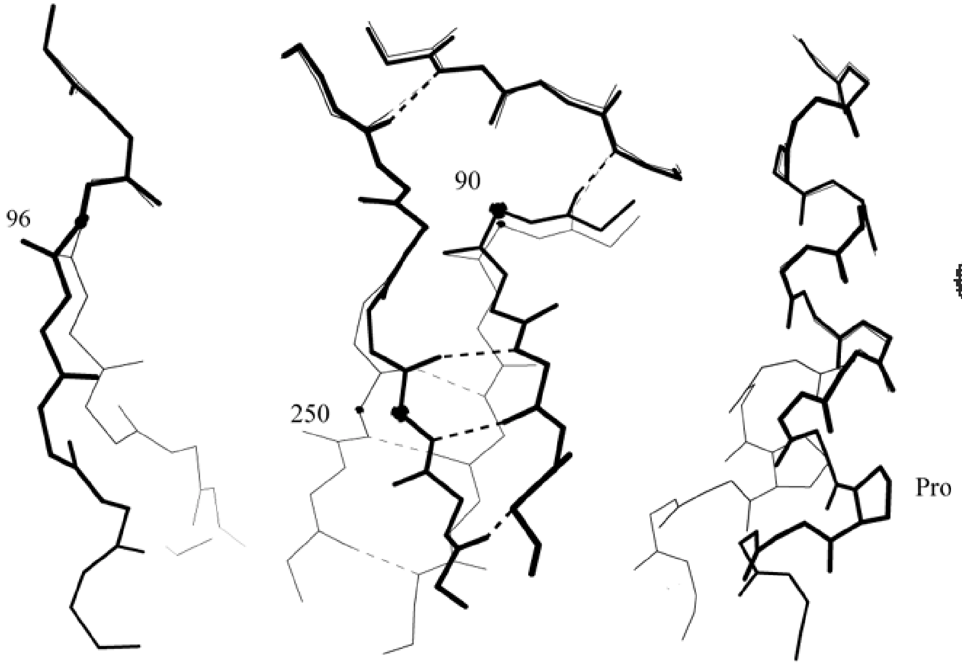


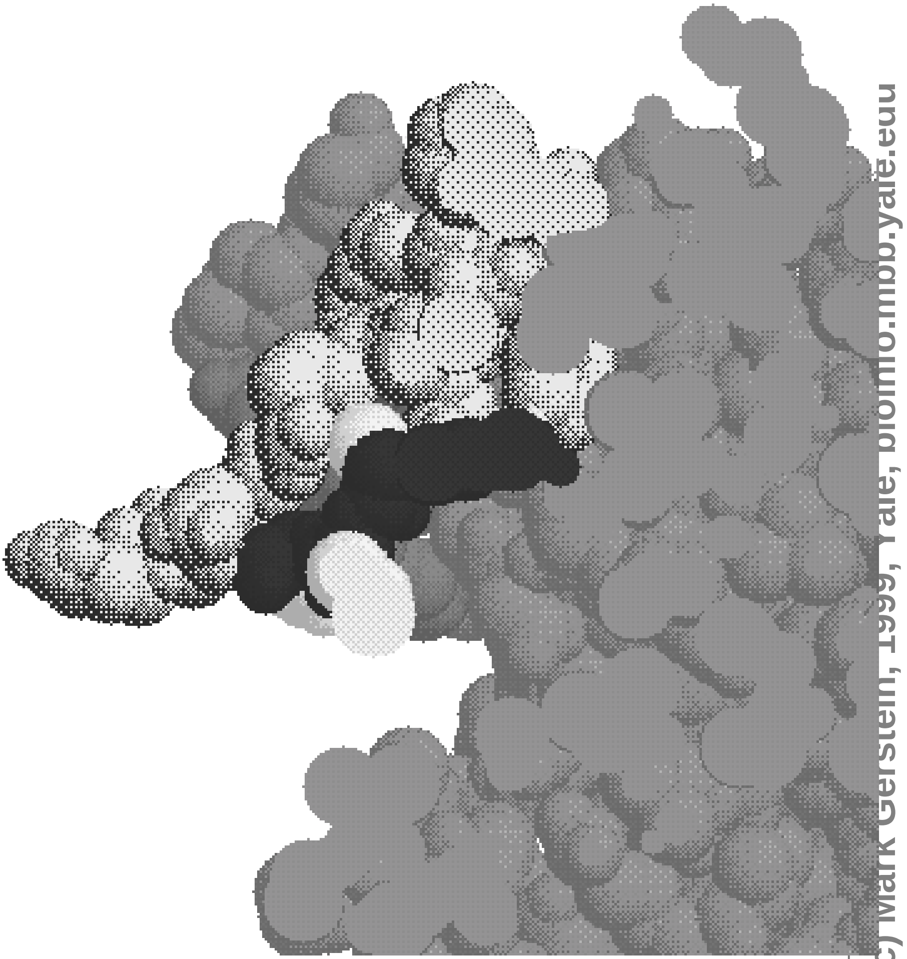
Hinge Mechanism involves absence of steric constraints (continuously maintained interface), esp. at hinge



Absence of Tight Packing at Hinge

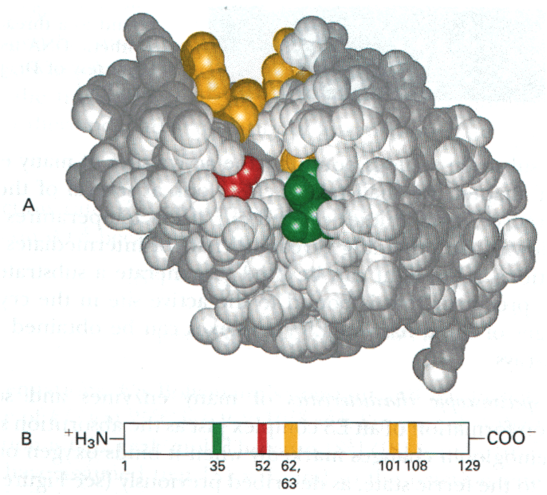
Chain Topology is not important

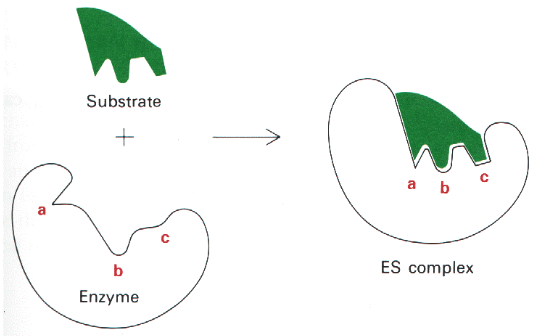




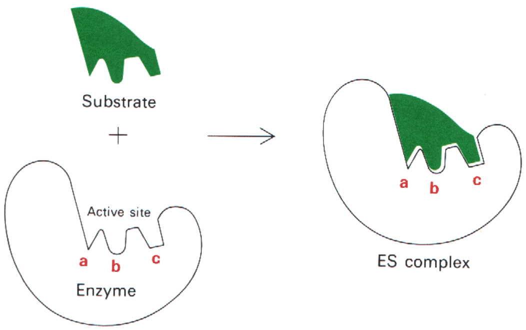
Enzyme Active Sites

Docking





Lock and Key



Induced Fit

•The active site of an enzyme is constituted from a relatively small part of the the total volume of an enzyme

•The active site is three-dimensional and formed from distant parts of the linear amino-acid or nucleic acid sequence

•Substrates are bound to enzymes by multiple weak interactions

•Active sites are usually clefts or crevices in the enzyme that maximize interaction with the substrate and exclude water

•The active site creates an unusual microenvironment that specifically stabilizes the chemical transition state

•The specificity of substrate binding depends upon the precise arrangements of atoms within the active site

•The active site can be prearranged (rigid lock and key mechanism) or have a dynamic interaction with the substrate (induced fit mechanism)