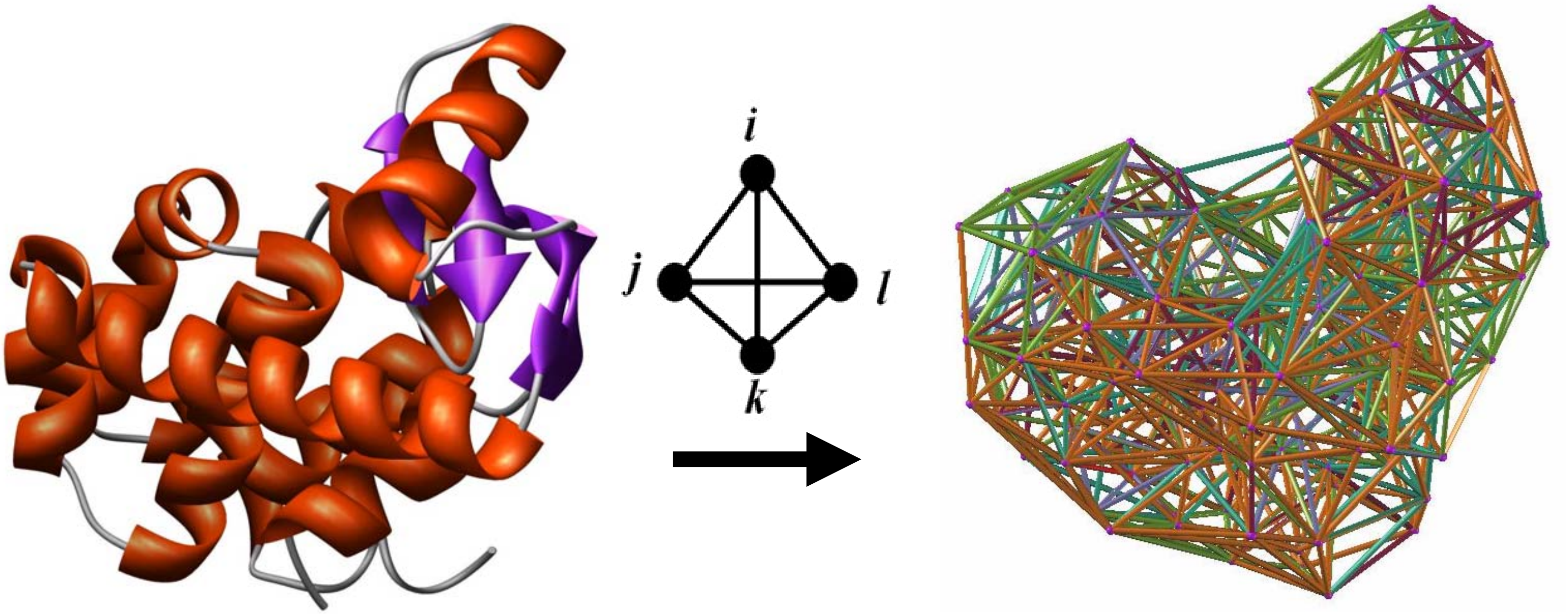
Using Biology to Teach Geometry: Protein Structure Tessellations in Matlab



Majid Masso, Ph.D.

Laboratory for Structural Bioinformatics

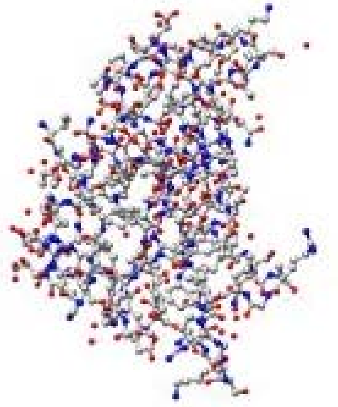
George Mason University

http://binf.gmu.edu/mmasso   **mmasso@gmu.edu**

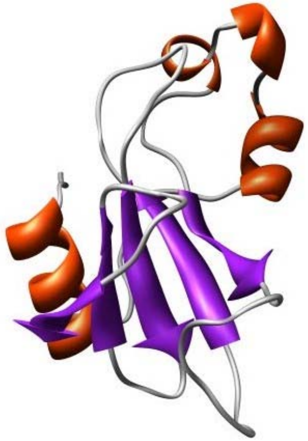
Proteins in Brief

- Intra- and inter-cellular workhorses of all organisms
- Building blocks: amino acids
  - 20 distinct types in nature (A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y)
  - ~200 ordered, successively linked amino acids/protein (varies widely across proteins, from tens to thousands)
- Protein structure representations:

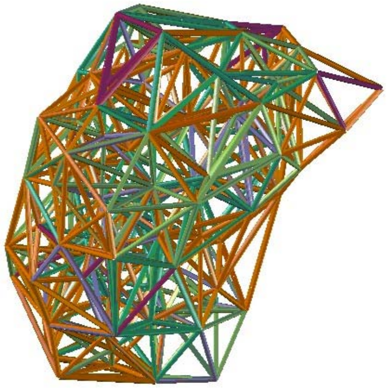
all-atom



backbone ribbon

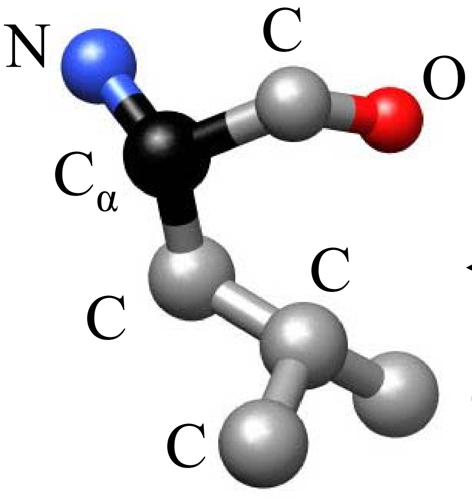


tessellation

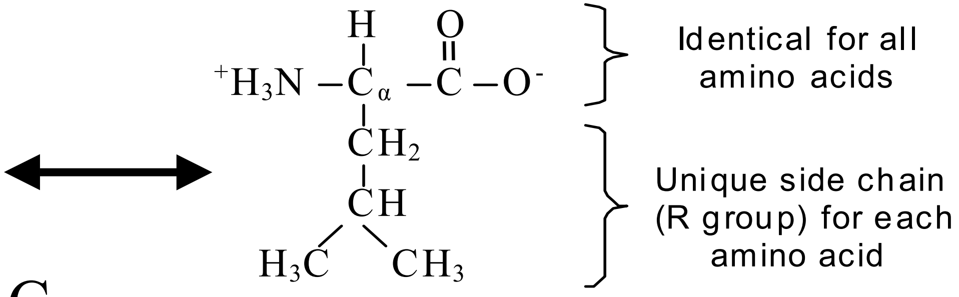


Amino Acids

- Atomic constituents: carbon (C), nitrogen (N), oxygen (O), and hydrogen (H)
- Amino Acids C (cysteine) and M (methionine) each also contain a sulfur (S) atom
- Coordinates of hydrogen atoms are only available for structures solved at very high resolution
- Example –



Leucine (Leu or L)



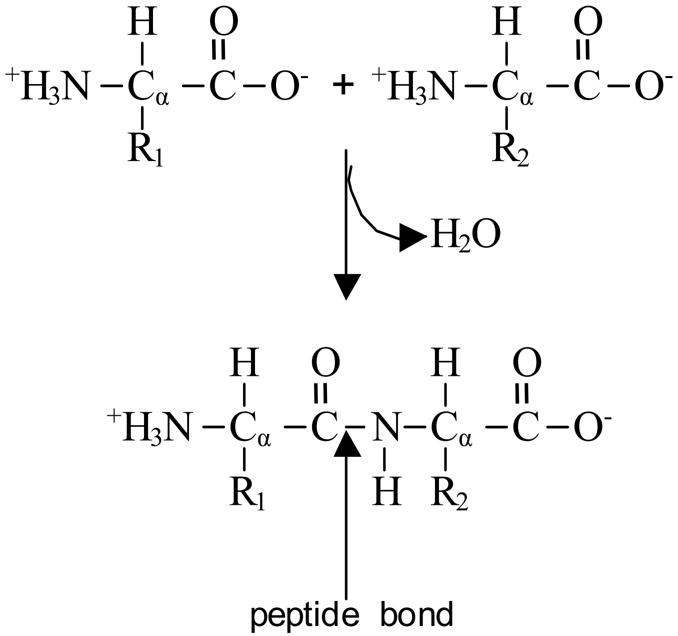
Identical for all amino acids

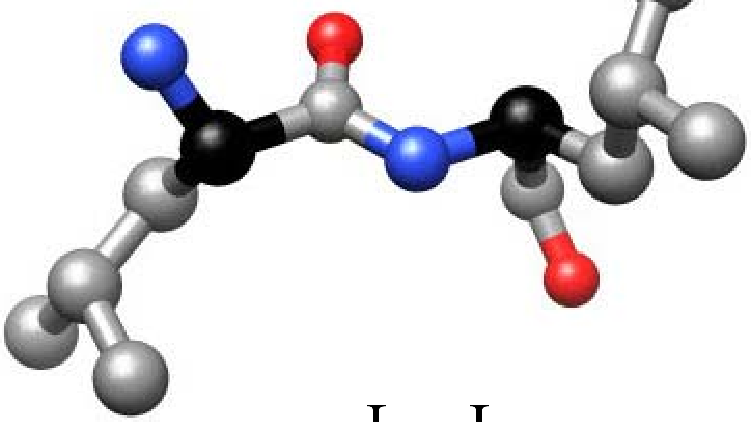
Unique side chain (R group) for each amino acid

Leucine (Leu or L)

Peptide Bond

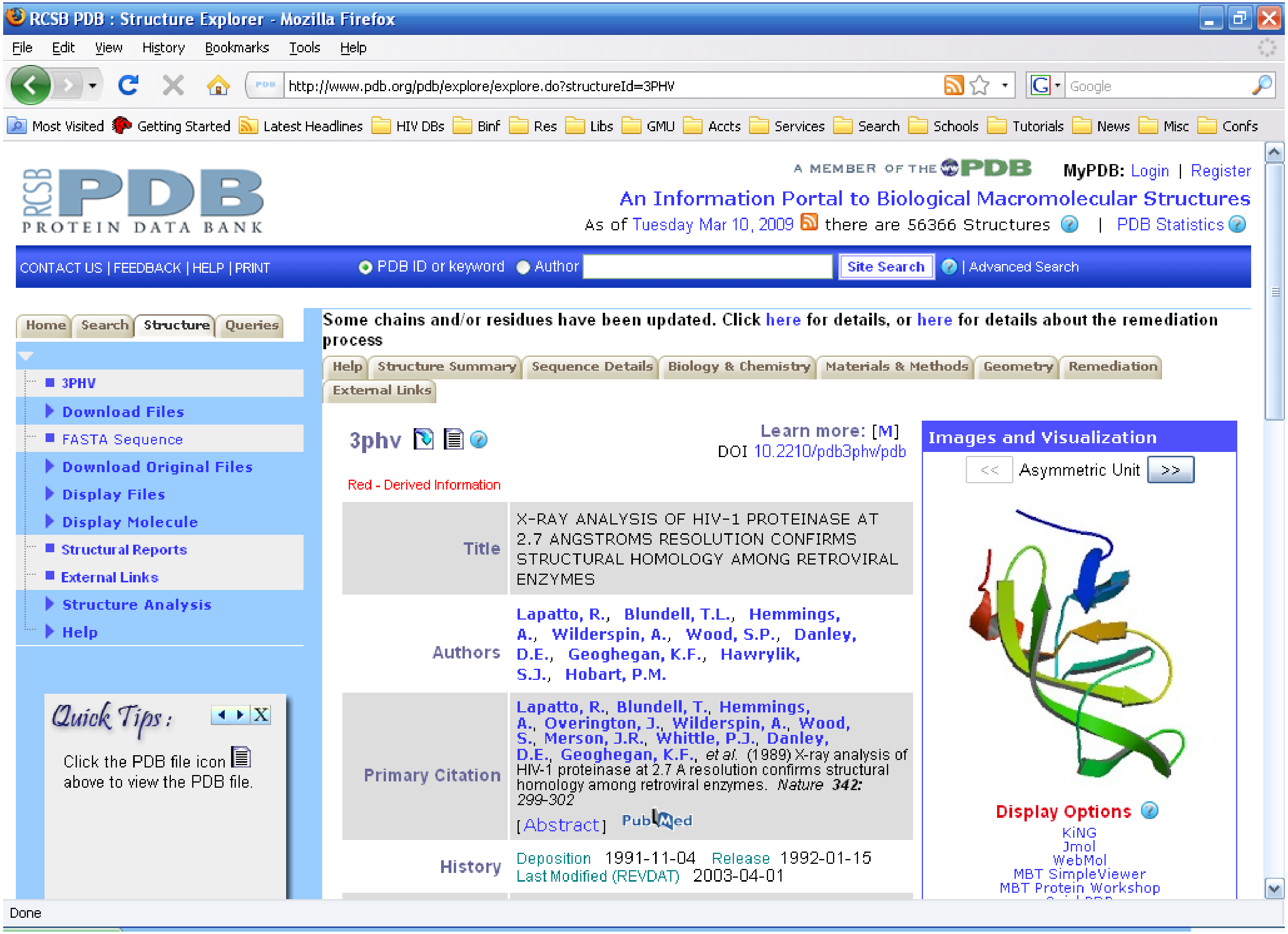
- Backbone linkage between consecutive amino acids in the growing, linear protein chain (the "primary sequence")
- Links the backbone C atom of amino acid *n*–1 to the backbone N atom of amino acid *n*, with release of H_2O





L – L

Protein Data Bank (PDB, http://www.pdb.org)



http://www.pdb.org/pdb/explore/explore.do?structureId=3PHV

MyPDB: Login | Register

An Information Portal to Biological Macromolecular Structures

As of Tuesday Mar 10, 2009 there are 56366 Structures | PDB Statistics

● PDB ID or keyword ○ Author

- 3PHV
- Download Files
- FASTA Sequence
- Download Original Files
- Display Files
- Display Molecule
- Structural Reports
- External Links
- Structure Analysis
- Help

Quick Tips:

Click the PDB file icon above to view the PDB file.

Some chains and/or residues have been updated. Click here for details, or here for details about the remediation process

3phv

Learn more: [M]
DOI 10.2210/pdb3phv/pdb

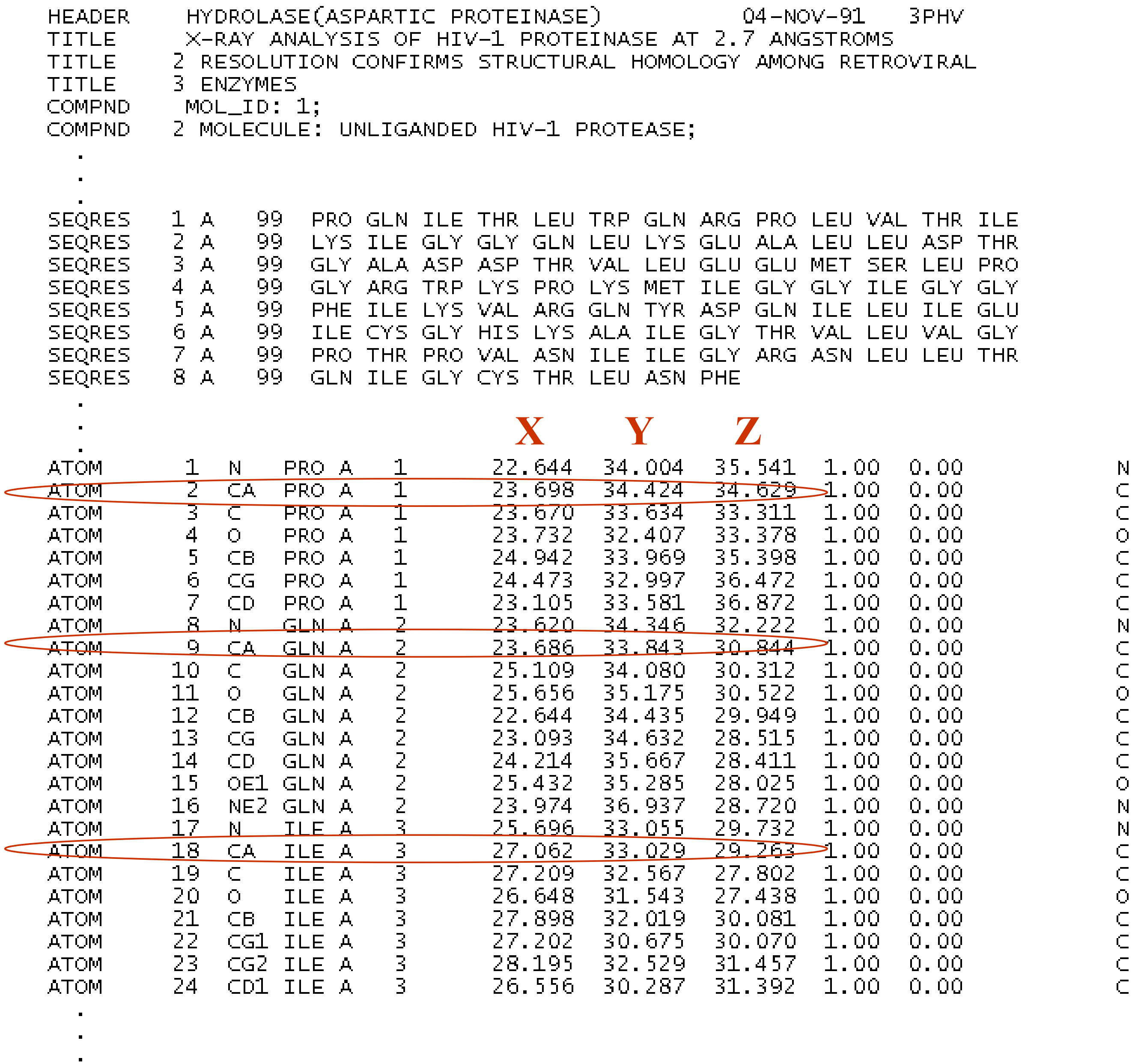
Red - Derived Information

Title	X-RAY ANALYSIS OF HIV-1 PROTEINASE AT 2.7 ANGSTROMS RESOLUTION CONFIRMS STRUCTURAL HOMOLOGY AMONG RETROVIRAL ENZYMES
Authors	Lapatto, R., Blundell, T.L., Hemmings, A., Wilderspin, A., Wood, S.P., Danley, D.E., Geoghegan, K.F., Hawrylik, S.J., Hobart, P.M.
Primary Citation	Lapatto, R., Blundell, T., Hemmings, A., Overington, J., Wilderspin, A., Wood, S., Merson, J.R., Whittle, P.J., Danley, D.E., Geoghegan, K.F., *et al.* (1989) X-ray analysis of HIV-1 proteinase at 2.7 A resolution confirms structural homology among retroviral enzymes. *Nature* ***342:*** *299-302* [Abstract] PubMed
History	Deposition 1991-11-04 Release 1992-01-15 Last Modified (REVDAT) 2003-04-01

Display Options

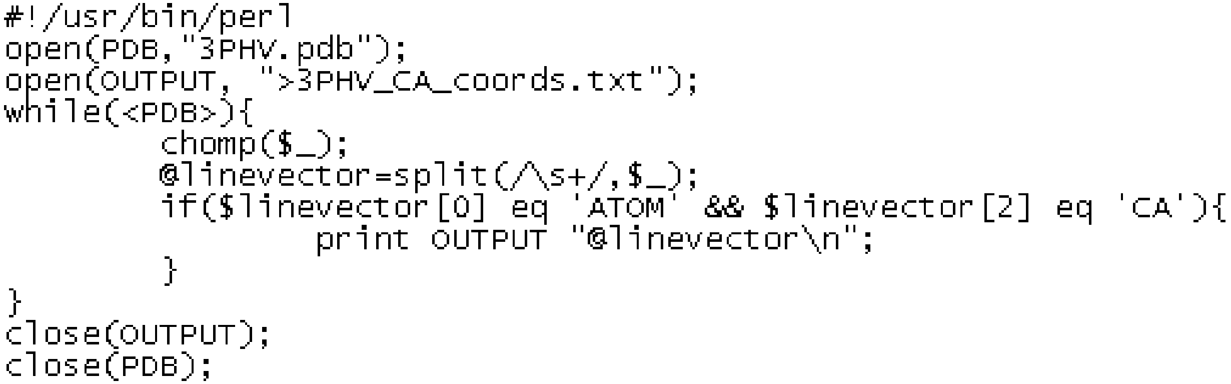
KiNG
Jmol
WebMol
MBT SimpleViewer
MBT Protein Workshop

PDB File Format

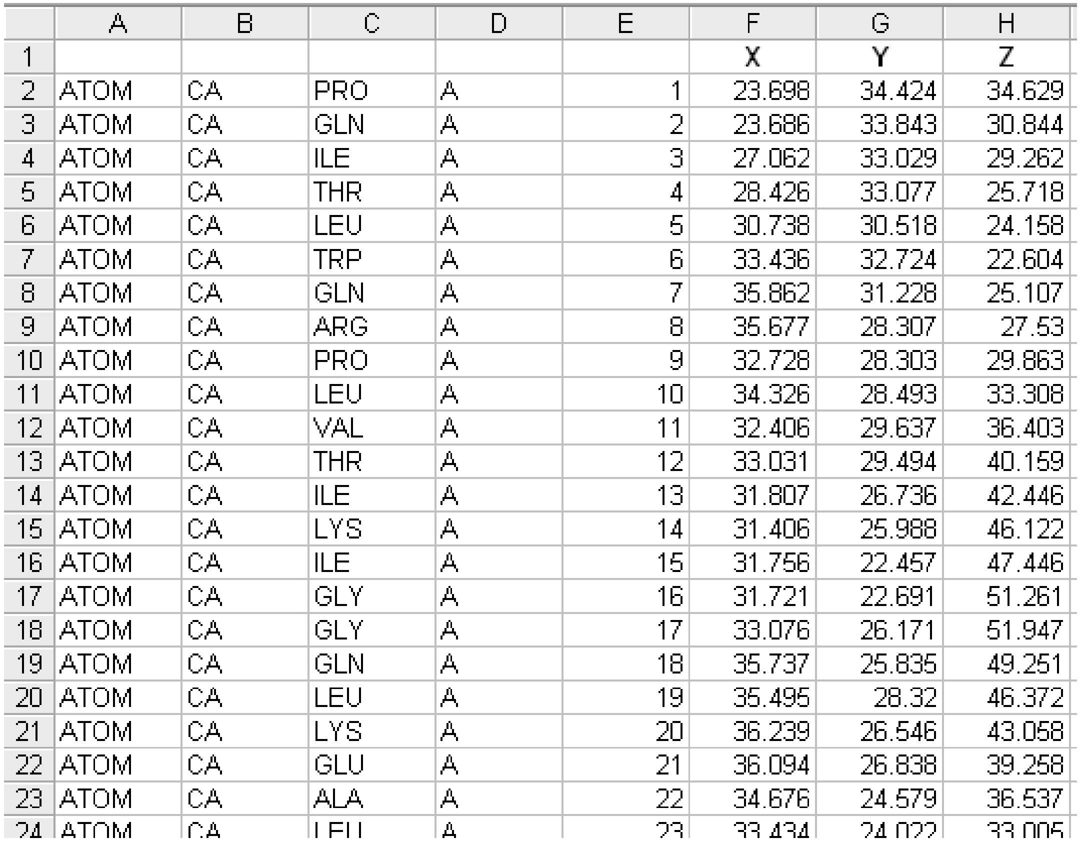


```
HEADER    HYDROLASE(ASPARTIC PROTEINASE)          04-NOV-91   3PHV
TITLE     X-RAY ANALYSIS OF HIV-1 PROTEINASE AT 2.7 ANGSTROMS
TITLE    2 RESOLUTION CONFIRMS STRUCTURAL HOMOLOGY AMONG RETROVIRAL
TITLE    3 ENZYMES
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: UNLIGANDED HIV-1 PROTEASE;
   .
   .
   .
SEQRES   1 A   99  PRO GLN ILE THR LEU TRP GLN ARG PRO LEU VAL THR ILE
SEQRES   2 A   99  LYS ILE GLY GLY GLN LEU LYS GLU ALA LEU LEU ASP THR
SEQRES   3 A   99  GLY ALA ASP ASP THR VAL LEU GLU GLU MET SER LEU PRO
SEQRES   4 A   99  GLY ARG TRP LYS PRO LYS MET ILE GLY GLY ILE GLY GLY
SEQRES   5 A   99  PHE ILE LYS VAL ARG GLN TYR ASP GLN ILE LEU ILE GLU
SEQRES   6 A   99  ILE CYS GLY HIS LYS ALA ILE GLY THR VAL LEU VAL GLY
SEQRES   7 A   99  PRO THR PRO VAL ASN ILE ILE GLY ARG ASN LEU LEU THR
SEQRES   8 A   99  GLN ILE GLY CYS THR LEU ASN PHE
   .
   .                              X       Y       Z
   .
ATOM      1  N   PRO A   1      22.644  34.004  35.541  1.00  0.00           N
ATOM      2  CA  PRO A   1      23.698  34.424  34.629  1.00  0.00           C
ATOM      3  C   PRO A   1      23.670  33.634  33.311  1.00  0.00           C
ATOM      4  O   PRO A   1      23.732  32.407  33.378  1.00  0.00           O
ATOM      5  CB  PRO A   1      24.942  33.969  35.398  1.00  0.00           C
ATOM      6  CG  PRO A   1      24.473  32.997  36.472  1.00  0.00           C
ATOM      7  CD  PRO A   1      23.105  33.581  36.872  1.00  0.00           C
ATOM      8  N   GLN A   2      23.620  34.346  32.222  1.00  0.00           N
ATOM      9  CA  GLN A   2      23.686  33.843  30.844  1.00  0.00           C
ATOM     10  C   GLN A   2      25.109  34.080  30.312  1.00  0.00           C
ATOM     11  O   GLN A   2      25.656  35.175  30.522  1.00  0.00           O
ATOM     12  CB  GLN A   2      22.644  34.435  29.949  1.00  0.00           C
ATOM     13  CG  GLN A   2      23.093  34.632  28.515  1.00  0.00           C
ATOM     14  CD  GLN A   2      24.214  35.667  28.411  1.00  0.00           C
ATOM     15  OE1 GLN A   2      25.432  35.285  28.025  1.00  0.00           O
ATOM     16  NE2 GLN A   2      23.974  36.937  28.720  1.00  0.00           N
ATOM     17  N   ILE A   3      25.696  33.055  29.732  1.00  0.00           N
ATOM     18  CA  ILE A   3      27.062  33.029  29.263  1.00  0.00           C
ATOM     19  C   ILE A   3      27.209  32.567  27.802  1.00  0.00           C
ATOM     20  O   ILE A   3      26.648  31.543  27.438  1.00  0.00           O
ATOM     21  CB  ILE A   3      27.898  32.019  30.081  1.00  0.00           C
ATOM     22  CG1 ILE A   3      27.202  30.675  30.070  1.00  0.00           C
ATOM     23  CG2 ILE A   3      28.195  32.529  31.457  1.00  0.00           C
ATOM     24  CD1 ILE A   3      26.556  30.287  31.392  1.00  0.00           C
   .
   .
   .
```

HIV-1 Protease CA Coordinate Data

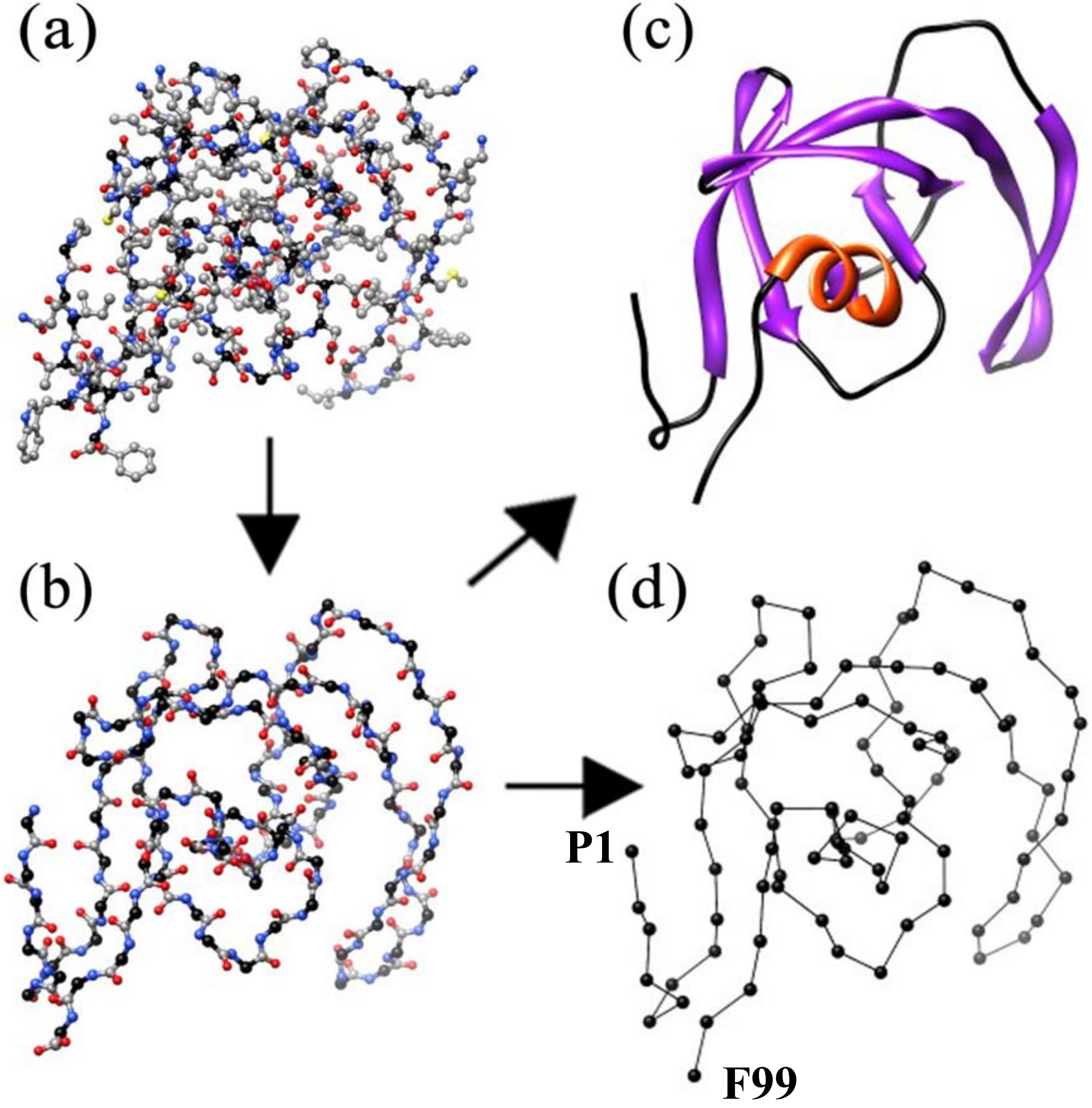


```
#!/usr/bin/perl
open(PDB,"3PHV.pdb");
open(OUTPUT, ">3PHV_CA_coords.txt");
while(<PDB>){
        chomp($_);
        @linevector=split(/\s+/,$_);
        if($linevector[0] eq 'ATOM' && $linevector[2] eq 'CA'){
                print OUTPUT "@linevector\n";
        }
}
close(OUTPUT);
close(PDB);
```



	A	B	C	D	E	F	G	H
1						X	Y	Z
2	ATOM	CA	PRO	A	1	23.698	34.424	34.629
3	ATOM	CA	GLN	A	2	23.686	33.843	30.844
4	ATOM	CA	ILE	A	3	27.062	33.029	29.262
5	ATOM	CA	THR	A	4	28.426	33.077	25.718
6	ATOM	CA	LEU	A	5	30.738	30.518	24.158
7	ATOM	CA	TRP	A	6	33.436	32.724	22.604
8	ATOM	CA	GLN	A	7	35.862	31.228	25.107
9	ATOM	CA	ARG	A	8	35.677	28.307	27.53
10	ATOM	CA	PRO	A	9	32.728	28.303	29.863
11	ATOM	CA	LEU	A	10	34.326	28.493	33.308
12	ATOM	CA	VAL	A	11	32.406	29.637	36.403
13	ATOM	CA	THR	A	12	33.031	29.494	40.159
14	ATOM	CA	ILE	A	13	31.807	26.736	42.446
15	ATOM	CA	LYS	A	14	31.406	25.988	46.122
16	ATOM	CA	ILE	A	15	31.756	22.457	47.446
17	ATOM	CA	GLY	A	16	31.721	22.691	51.261
18	ATOM	CA	GLY	A	17	33.076	26.171	51.947
19	ATOM	CA	GLN	A	18	35.737	25.835	49.251
20	ATOM	CA	LEU	A	19	35.495	28.32	46.372
21	ATOM	CA	LYS	A	20	36.239	26.546	43.058
22	ATOM	CA	GLU	A	21	36.094	26.838	39.258
23	ATOM	CA	ALA	A	22	34.676	24.579	36.537
24	ATOM	CA	LEU	A	23	33.434	24.022	33.005

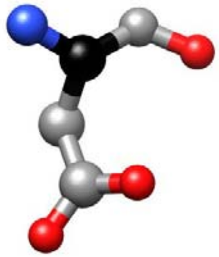
Example: HIV-1 Protease (PDB ID: 3PHV)



(a) all-atom (b) backbone only (c) ribbon diagram (d) C_α trace

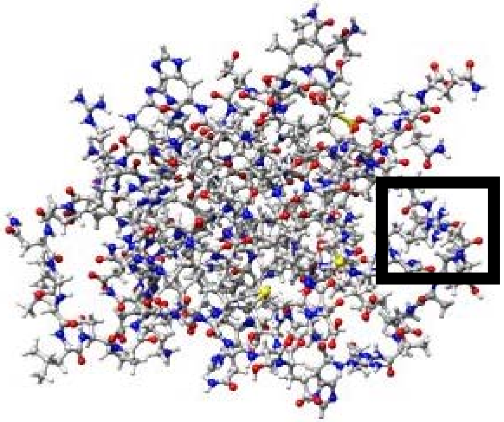
Delaunay Tessellation of Protein Structure

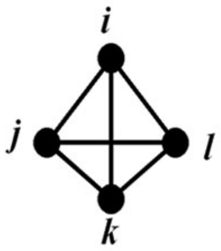
Aspartic Acid
(Asp or D)



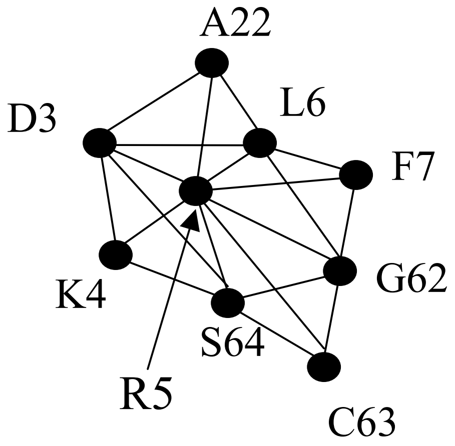
→ ● C_α coordinates in 3D

Abstract every amino acid to a point
Atomic coordinates – Protein Data Bank (PDB)



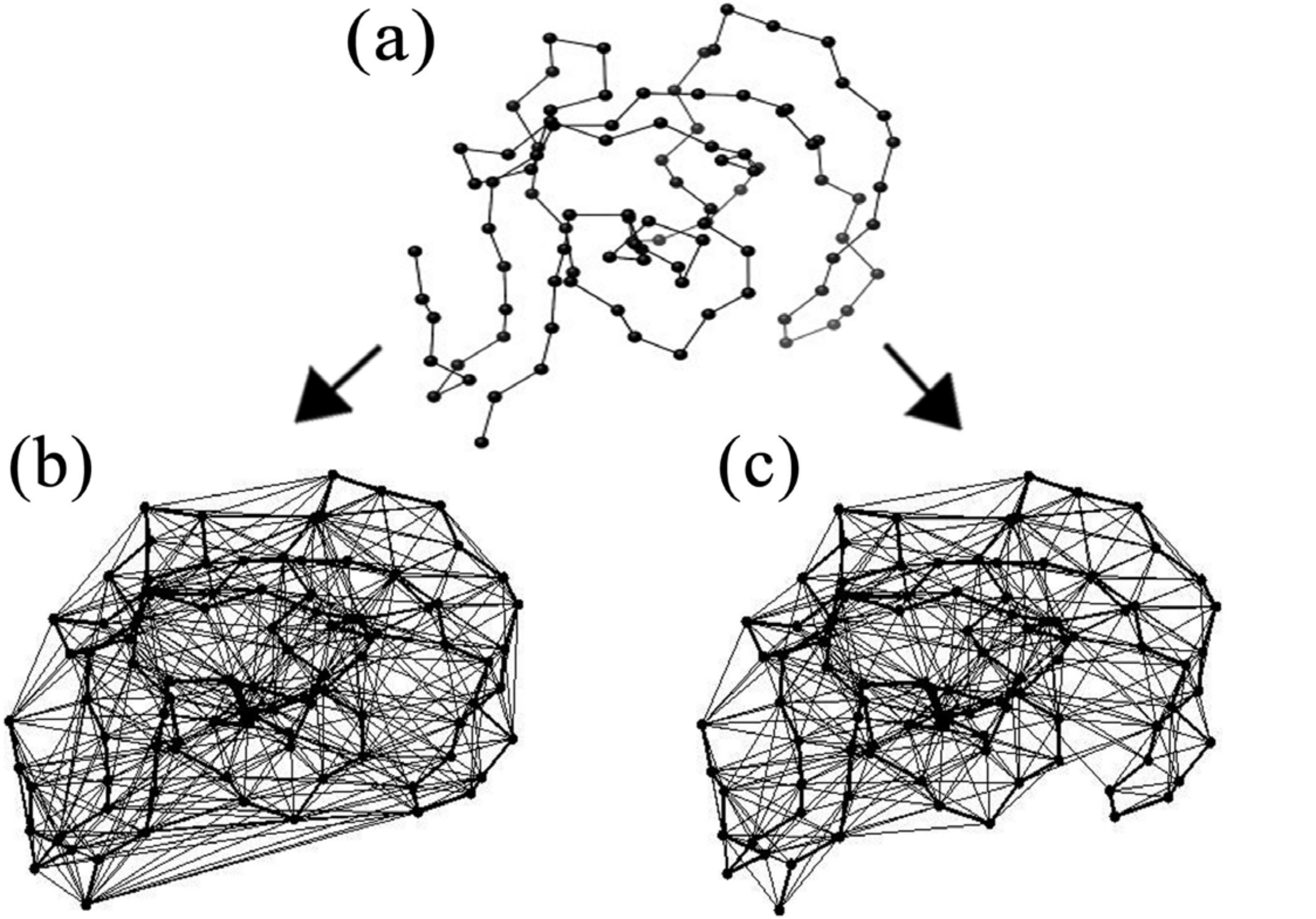


→



Delaunay tessellation: 3D “tiling” of space into non-overlapping, irregular tetrahedral simplices. Each simplex objectively identifies a quadruplet of nearest-neighbor amino acids at its vertices.

Tessellation Example: HIV-1 Protease (PR)

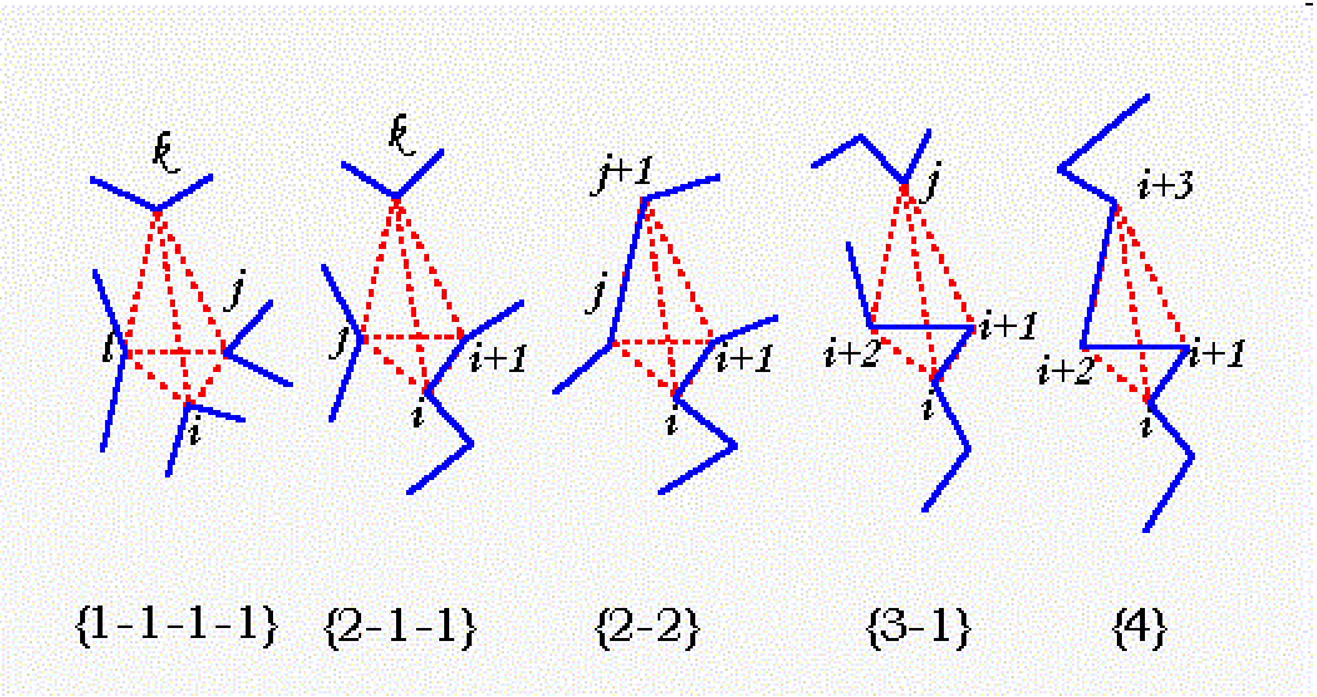


(a) C_α trace (b) complete tessellation (convex hull of simplices)
(c) tessellation subject to a 12 Angstrom edge length cutoff

Delaunay Tessellation in Matlab

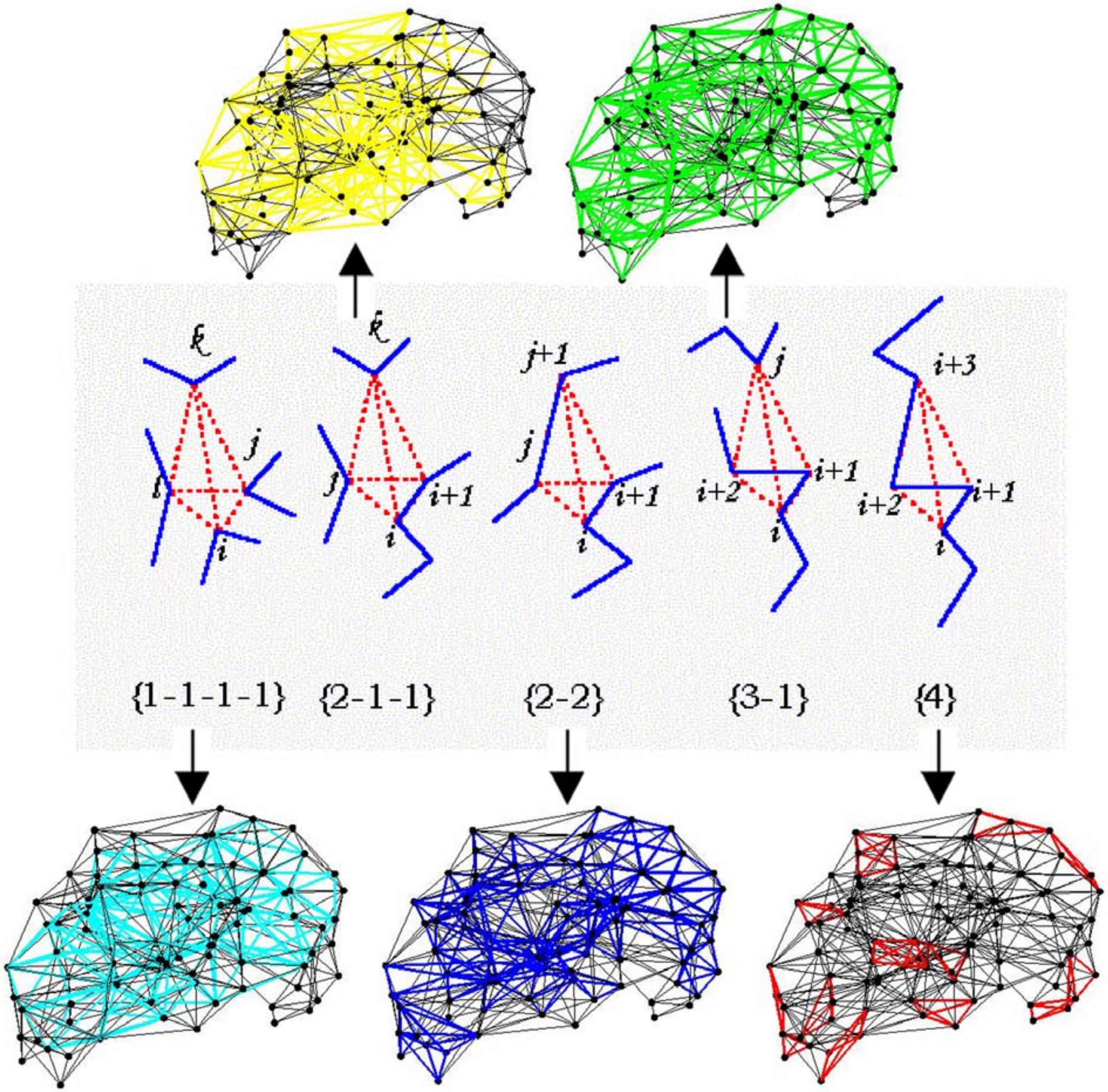
```
% Majid Masso, George Mason University, Manassas, VA
% Coordinates of the points (CA atoms) representing each of the N amino acids, each
column vector is N-dimensional
x=[];y=[];z=[];
% CA trace
plot3(x,y,z);
% Overlap graphs
hold on;
% Or just CA points
plot3(x,y,z,'.');
% No axes
axis off;
% Concatenate, w is an Nx3 matrix, each row gives the 3D coordinates of one CA
point, each CA point is indexed by its row number in w, from 1 to N
w=[x(:) y(:) z(:)];
% Tessellation, T is an rx4 matrix, r is the total number of tetrahedral simplices
in the tessellation, the 4 numbers in each row are the indices (row numbers in w) of
the CA points forming the vertices of a tetrahedral simplex
T=delaunay3(x,y,z);
% Full tessellation (convex hull of tetrahedral simplices)
% FaceAlpha is the the transparency of the triangular faces (set to 0)
tetramesh(T,w,'FaceAlpha',0);
% Alternative is tessellation subject to a 12.0 A edge-length cutoff
s=size(T);
r=s(1,1);
k=1;
for i=1:r
 e=T(i,1);f=T(i,2);g=T(i,3);h=T(i,4);
 if(sqrt((w(e,1)-w(f,1))^2+(w(e,2)-w(f,2))^2+(w(e,3)-w(f,3))^2) <=12.0 &&
sqrt((w(e,1)-w(g,1))^2+(w(e,2)-w(g,2))^2+(w(e,3)-w(g,3))^2)<=12.0 &&
sqrt((w(e,1)-w(h,1))^2+(w(e,2)-w(h,2))^2+(w(e,3)-w(h,3))^2)<=12.0 &&
sqrt((w(f,1)-w(g,1))^2+(w(f,2)-w(g,2))^2+(w(f,3)-w(g,3))^2)<=12.0 &&
sqrt((w(f,1)-w(h,1))^2+(w(f,2)-w(h,2))^2+(w(f,3)-w(h,3))^2)<=12.0 &&
sqrt((w(g,1)-w(h,1))^2+(w(g,2)-w(h,2))^2+(w(g,3)-w(h,3))^2)<=12.0)
  t(k,1)=e;t(k,2)=f;t(k,3)=g;t(k,4)=h;
  k=k+1;
 end
end
tetramesh(t,w,'FaceAlpha',0);
```

Five Simplex Categories

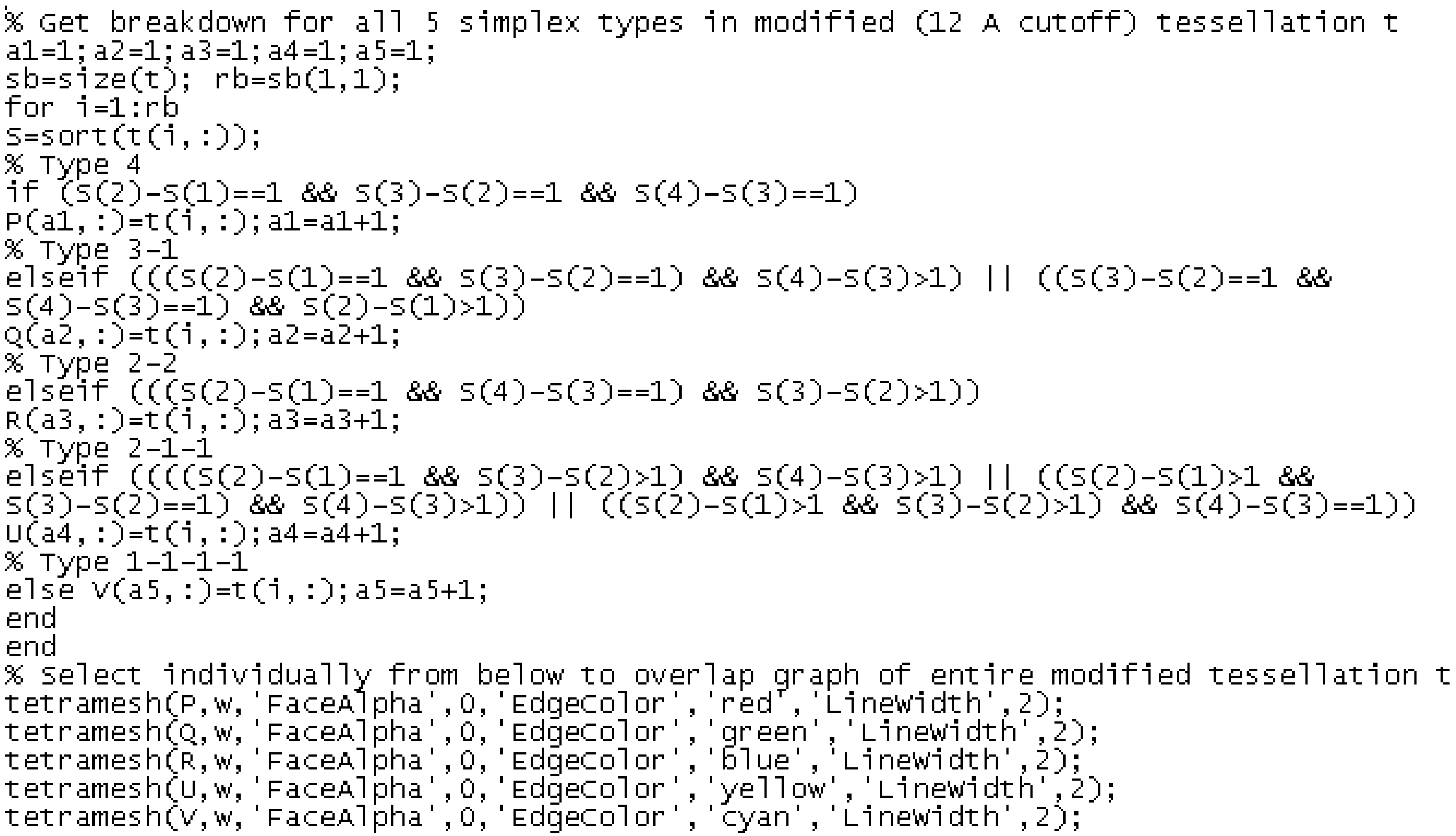


Singh *et al.* (1996) J. Comput. Biol., **3**, 213-222

Simplex Categories Example: HIV-1 PR

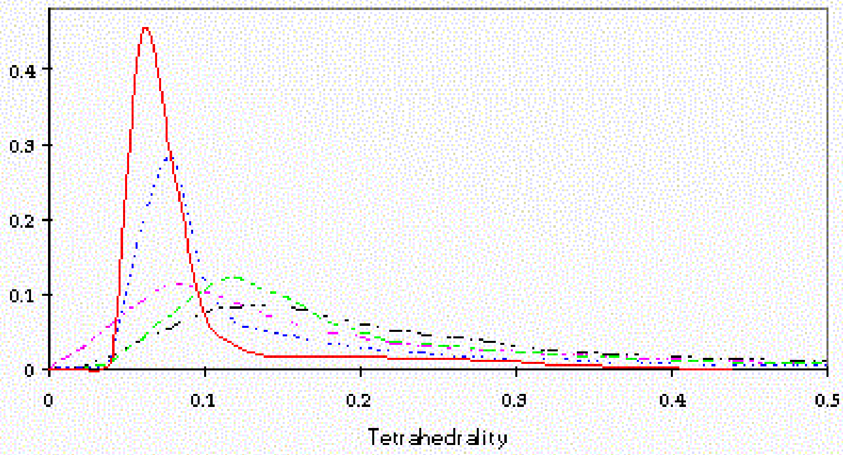


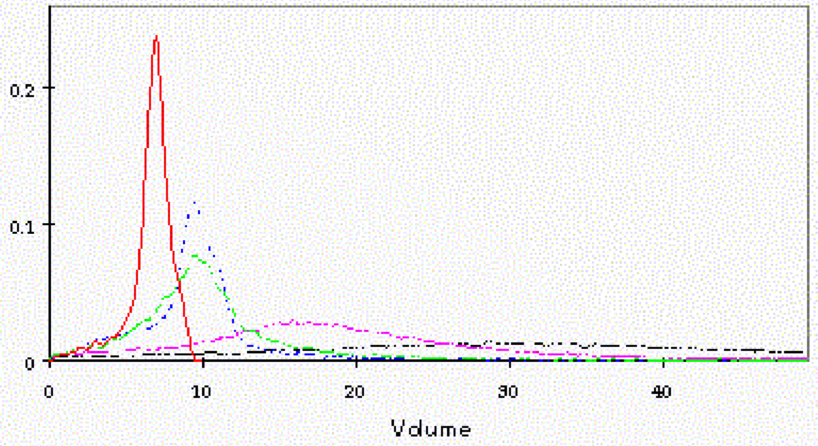
Simplex Categories in Matlab



```
% Get breakdown for all 5 simplex types in modified (12 A cutoff) tessellation t
a1=1;a2=1;a3=1;a4=1;a5=1;
sb=size(t); rb=sb(1,1);
for i=1:rb
S=sort(t(i,:));
% Type 4
if (S(2)-S(1)==1 && S(3)-S(2)==1 && S(4)-S(3)==1)
P(a1,:)=t(i,:);a1=a1+1;
% Type 3-1
elseif (((S(2)-S(1)==1 && S(3)-S(2)==1) && S(4)-S(3)>1) || ((S(3)-S(2)==1 &&
S(4)-S(3)==1) && S(2)-S(1)>1))
Q(a2,:)=t(i,:);a2=a2+1;
% Type 2-2
elseif (((S(2)-S(1)==1 && S(4)-S(3)==1) && S(3)-S(2)>1))
R(a3,:)=t(i,:);a3=a3+1;
% Type 2-1-1
elseif ((((S(2)-S(1)==1 && S(3)-S(2)>1) && S(4)-S(3)>1) || ((S(2)-S(1)>1 &&
S(3)-S(2)==1) && S(4)-S(3)>1)) || ((S(2)-S(1)>1 && S(3)-S(2)>1) && S(4)-S(3)==1))
U(a4,:)=t(i,:);a4=a4+1;
% Type 1-1-1-1
else V(a5,:)=t(i,:);a5=a5+1;
end
end
% Select individually from below to overlap graph of entire modified tessellation t
tetramesh(P,w,'FaceAlpha',0,'EdgeColor','red','Linewidth',2);
tetramesh(Q,w,'FaceAlpha',0,'EdgeColor','green','Linewidth',2);
tetramesh(R,w,'FaceAlpha',0,'EdgeColor','blue','Linewidth',2);
tetramesh(U,w,'FaceAlpha',0,'EdgeColor','yellow','Linewidth',2);
tetramesh(V,w,'FaceAlpha',0,'EdgeColor','cyan','Linewidth',2);
```

Tetrahedrality and Volume





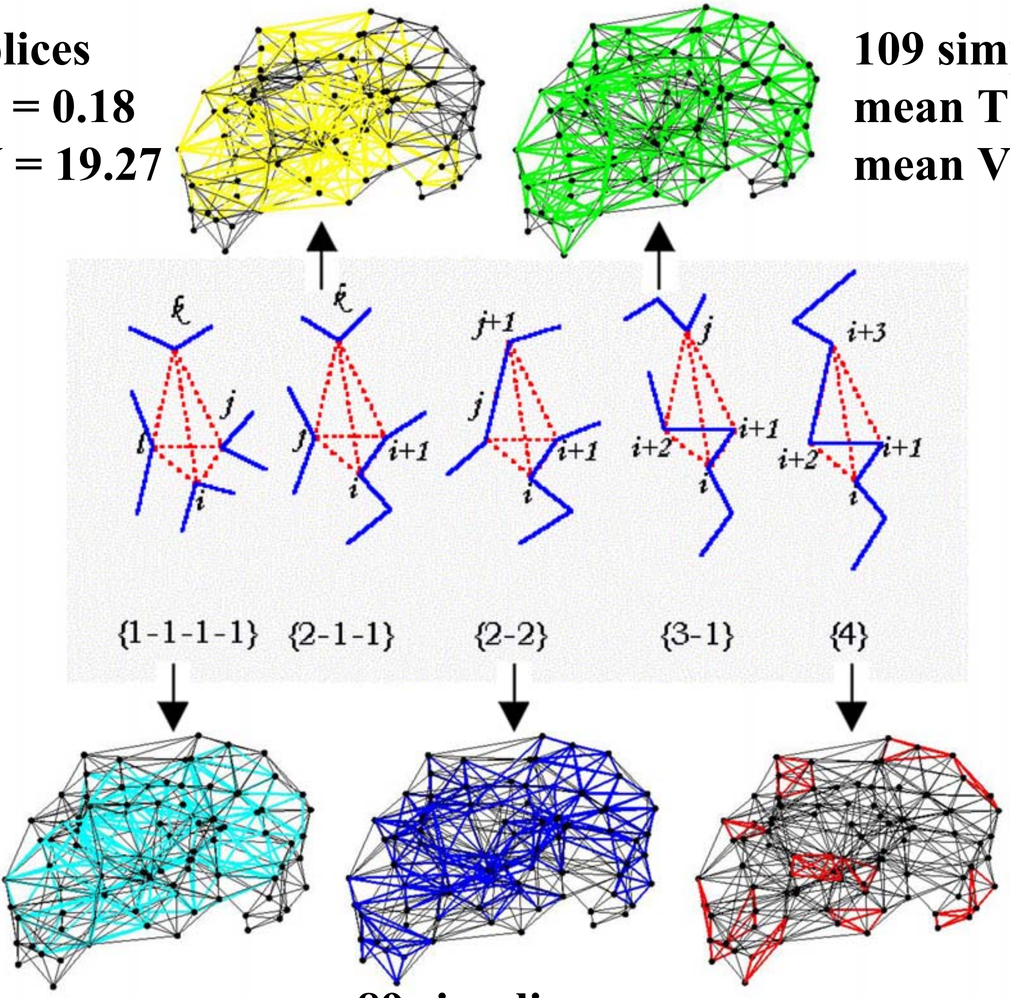
$$T = \sum_{i>j} (l_i - l_j)^2 \,/\, 15\bar{l}^2$$

$$V = \frac{|(\mathbf{a}-\mathbf{d}) \cdot ((\mathbf{b}-\mathbf{d}) \times (\mathbf{c}-\mathbf{d}))|}{6}$$

Vectors **a**, **b**, **c**, and **d** represent the 3D coordinates of the tetrahedral vertices.

Singh *et al.* (1996) J. Comput. Biol., **3**, 213-222

Tetrahedrality and Volume Example: HIV-1 PR



95 simplices
mean T = 0.18
mean V = 19.27

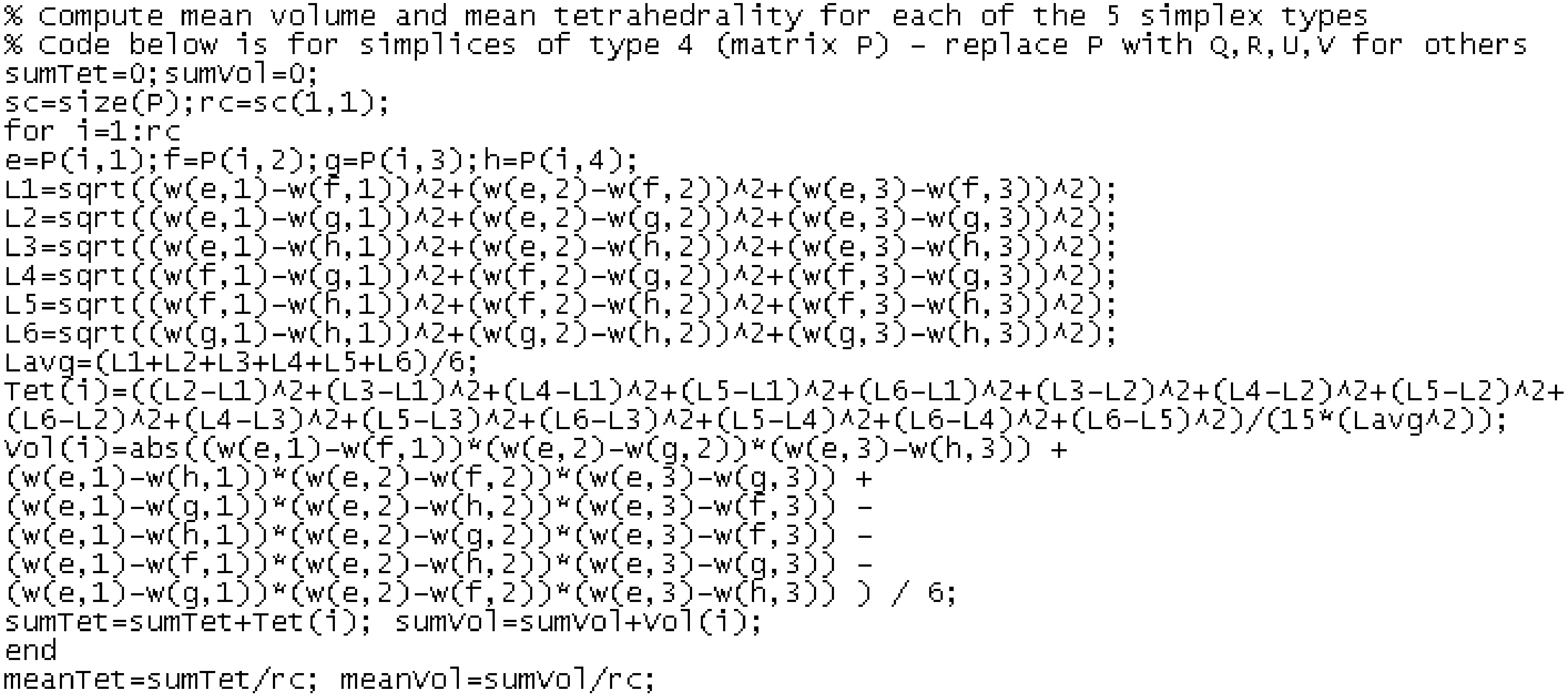
109 simplices
mean T = 0.20
mean V = 10.09

73 simplices
mean T = 0.11
mean V = 41.51

89 simplices
mean T = 0.15
mean V = 9.45

16 simplices
mean T = 0.18
mean V = 5.61

Tetrahedrality and Volume in Matlab



```
% Compute mean volume and mean tetrahedrality for each of the 5 simplex types
% Code below is for simplices of type 4 (matrix P) - replace P with Q,R,U,V for others
sumTet=0;sumVol=0;
sc=size(P);rc=sc(1,1);
for i=1:rc
e=P(i,1);f=P(i,2);g=P(i,3);h=P(i,4);
L1=sqrt((w(e,1)-w(f,1))^2+(w(e,2)-w(f,2))^2+(w(e,3)-w(f,3))^2);
L2=sqrt((w(e,1)-w(g,1))^2+(w(e,2)-w(g,2))^2+(w(e,3)-w(g,3))^2);
L3=sqrt((w(e,1)-w(h,1))^2+(w(e,2)-w(h,2))^2+(w(e,3)-w(h,3))^2);
L4=sqrt((w(f,1)-w(g,1))^2+(w(f,2)-w(g,2))^2+(w(f,3)-w(g,3))^2);
L5=sqrt((w(f,1)-w(h,1))^2+(w(f,2)-w(h,2))^2+(w(f,3)-w(h,3))^2);
L6=sqrt((w(g,1)-w(h,1))^2+(w(g,2)-w(h,2))^2+(w(g,3)-w(h,3))^2);
Lavg=(L1+L2+L3+L4+L5+L6)/6;
Tet(i)=((L2-L1)^2+(L3-L1)^2+(L4-L1)^2+(L5-L1)^2+(L6-L1)^2+(L3-L2)^2+(L4-L2)^2+(L5-L2)^2+
(L6-L2)^2+(L4-L3)^2+(L5-L3)^2+(L6-L3)^2+(L5-L4)^2+(L6-L4)^2+(L6-L5)^2)/(15*(Lavg^2));
Vol(i)=abs((w(e,1)-w(f,1))*(w(e,2)-w(g,2))*(w(e,3)-w(h,3)) +
(w(e,1)-w(h,1))*(w(e,2)-w(f,2))*(w(e,3)-w(g,3)) +
(w(e,1)-w(g,1))*(w(e,2)-w(h,2))*(w(e,3)-w(f,3)) -
(w(e,1)-w(h,1))*(w(e,2)-w(g,2))*(w(e,3)-w(f,3)) -
(w(e,1)-w(f,1))*(w(e,2)-w(h,2))*(w(e,3)-w(g,3)) -
(w(e,1)-w(g,1))*(w(e,2)-w(f,2))*(w(e,3)-w(h,3)) ) / 6;
sumTet=sumTet+Tet(i); sumVol=sumVol+Vol(i);
end
meanTet=sumTet/rc; meanVol=sumVol/rc;
```

Counting Amino Acid Quadruplets

***n* = size of amino acid alphabet = 20; *r* = size of the subsets = 4**

only realistic choice for proteins

Repetitions Allowed?	Permutations Allowed?	Number of Quadruplets
[✓ yes]	yes	$n^r = 20^4 = 160{,}000$
[✓ yes]	[✓ no]	$\begin{pmatrix} n+r-1 \\ r \end{pmatrix} = \begin{pmatrix} 23 \\ 4 \end{pmatrix} = 8855$
no	yes	$\dfrac{n!}{(n-r)!} = \dfrac{20!}{16!} = 116{,}280$
no	no	$\dfrac{n!}{r!(n-r)!} = \begin{pmatrix} n \\ r \end{pmatrix} = \begin{pmatrix} 20 \\ 4 \end{pmatrix} = 4845$

only realistic choice to get enough quadruplets for each of the 8855 types (by tessellating a large, diverse set of protein structures) and obtain a frequency distribution

Counting Amino Acid Quadruplets

Repetitions – yes, permutations – no:
a more “hands-on” counting approach

$\underbrace{\text{C}}\ \underbrace{\text{D}}\ \underbrace{\text{E}}\ \underbrace{\text{F}}$	$\binom{20}{4}$
$\text{C}\ \text{C}\ \underbrace{\text{D}}\ \underbrace{\text{E}}$	$20\cdot\binom{19}{2}$
$\underbrace{\text{C}\ \text{C}}\ \underbrace{\text{D}\ \text{D}}$	$\binom{20}{2}$
C C C D	$20\cdot 19$
C C C C	20

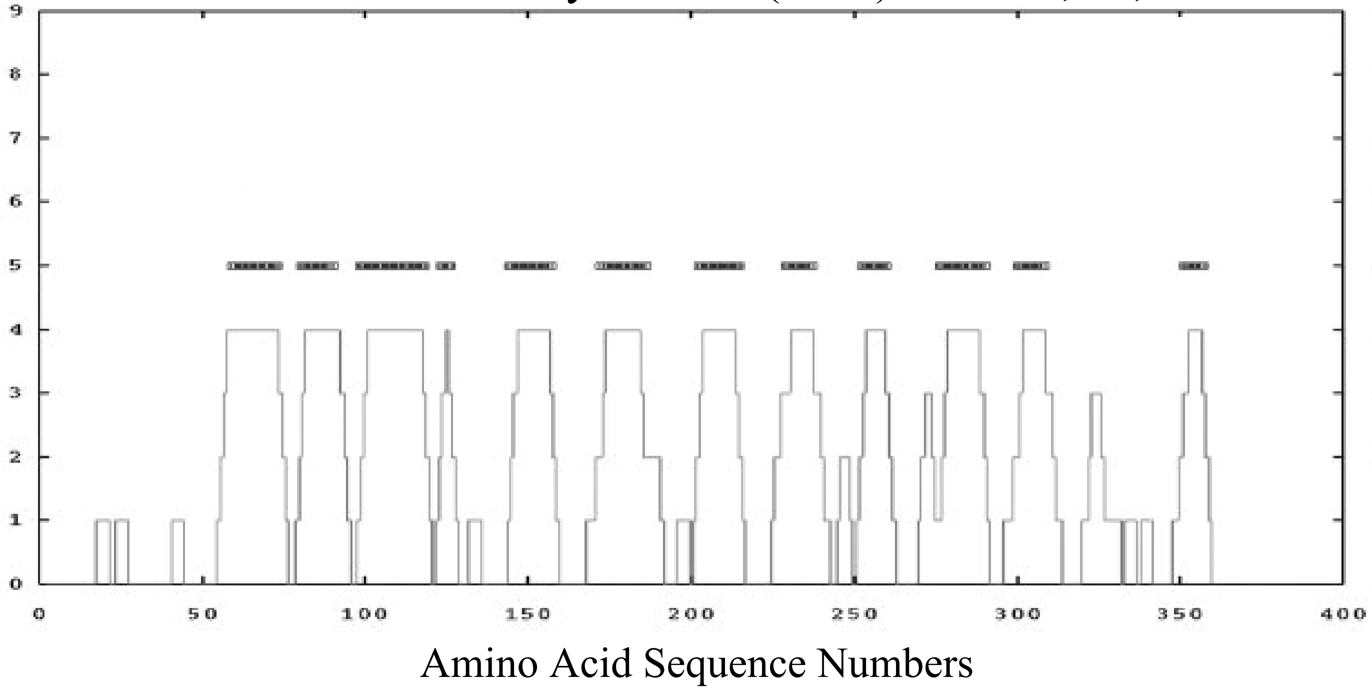
Total: 8,855 distinct quadruplets

Predicting Alpha Helix Locations in Proteins

Taylor *et al.* (2005) Proteins, **60**, 513-524

actual locations

predicted locations



Amino acid *i* can participate in up to 4 distinct simplices consisting of consecutive amino acids at the vertices: (*i*, *i*+1, *i*+2, *i*+3), (*i*-1, *i*, *i*+1, *i*+2), (*i*-2, *i*-1, *i*, *i*+1), and (*i*-3, *i*-2, *i*-1, *i*). The step-graph shows the number of such simplices ("t-number" values of 0, 1, 2, 3, or 4) for each amino acid in protein structure 2mnr. Amino acids with a t-number of 4 strongly correlate with those occurring in alpha helices.

References

- To obtain a copy of these slides: (**http://binf.gmu.edu/mmasso/MAA2010.pdf**)
- Protein structure repository: Protein Data Bank (**http://www.pdb.org**)
- Structure visualization: Chimera (**http://www.cgl.ucsf.edu/chimera/**)
- Delaunay tessellation:
  - Matlab (**http://www.mathworks.com/**)
  - Qhull (**http://www.qhull.org/**)
- Programming and data formatting: Perl (**http://www.activestate.com/activeperl/**)