Molecular Biophysics & Biochemistry 400a/700a (Advanced Biochemistry)

Computational Aspects of:
1, Secondary Structure Prediction;
2, Protein Packing

Mark Gerstein

Classes on 9/24/98 & 9/29/98

Yale University

The Handouts I

- Notes
  - ◊ this class (sec. str.) and next (packing)
- Presentation Paper
  - ◊ Frishman D, and Argos P. (1997) The Future of Protein Secondary Structure Prediction Accuracy. *Folding & Design* 2:159-62.
  - ◊ Controversial idea: secondary structure prediction to 80%?
  - ◊ http://bioinfo.mbb.yale.edu/~mbg/clippings/frishman-fad-acc-secstr.pdf (guest:guest)
- Secondary Structure Review
  - ◊ Handout from D Frishman
  - ◊ Garnier, J., Gibrat, J. F. & Robson, B. (1996b). GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol* **266**, 540-53.
- Problem Set Paper (Sec. Struc.)
  - ◊ King, R. D. & Sternberg, M. J. E. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Prot. Sci.* **5**, 2298-2310.

The Handouts II

- Problem Set Paper (Packing)
  - ◊ Joan Pontius, Jean Richelle, Shoshana J. Wodak (1996). Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures. Journal of Molecular Biology **264**: 121-136.
  - ◊ http://bioinfo.mbb.yale.edu/~mbg/clippings/wodak-jmb-volume.pdf (guest:guest)
- Packing Review
  - ◊ M Gerstein & F M Richards. "Protein Geometry: Distances, Areas, and Volumes," (eventually) to appear in *International Tables for Crystallography*. (International Union of Crystallography, Chester, UK).
  - ◊ http://bioinfo.mbb.yale.edu/e-print/geom-inttab/
- Optional Fun Reading (Only via Web)
  - ◊ Barry Cipra (1998). "Packing Challenge Mastered At Last," Science 281: 1267
  - ◊ http://www.sciencemag.org/cgi/content/full/281/5381/1267
  - ◊ Simon Singh (1998). "Mathematics 'Proves' What the Grocer Always Knew," *New York Times* (August 25).
  - ◊ http://bioinfo.mbb.yale.edu/~mbg/clippings-u/nyt-sci-packproof.txt

The Raw Material
Molecular Biology Information - DNA

- Raw DNA Sequence
  - ◊ Coding or Not?
  - ◊ Parse into genes?
  - ◊ 4 bases: AGCT
  - ◊ ~1 K in a gene, ~2 M in genome

```
atggcaattaaaattggtatcaatggttttggtcgtatcggccgtatcgtattccgtgca
gcacaacaccgtgatgacattgaagttgtaggtattaacgacttaatcgacgttgaatac
atggcttatatgttgaaatatgattcaactcacggtcgtttcgacggcactgttgaagtg
aaagatggtaacttagtggttaatggtaaaactatccgtgtaactgcagaacgtgatcca
gcaaacttaaactggggtgcaatcggtgttgatatcgctgttgaagcgactggtttattc
ttaactgatgaaactgctcgtaaacatatcactgcaggcgcaaaaaaagttgtattaact
ggcccatctaaagatgcaacccctatgttcgttcgtggtgtaaacttcaacgcatacgca
ggtcaagatatcgtttctaacgcatcttgtacaacaaactgtttagctccttagcacgt
gttgttcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
gcaactcaaaaaactgtggatggtccatcagctaaagactggcgcggcggccgcggtgca
tcacaaaacatcattccatcttcaacaggtgcagcgaaagcagtaggtaaagtattacct
gcattaaacggtaaattaactggtatggctttccgtgttccaacgccaaacgtatctgtt
gttgatttaacagttaatcttgaaaaaccagcttcttatgatgcaatcaaacaagcaatc
aaagatgcagcggaaggtaaaacgttcaatggcgaattaaaaggcgtattaggttacact
gaagatgctgttgtttctactgacttcaacggttgtgctttaacttctgtatttgatgca
gacgctggtatcgcattaactgattctttcgttaaattggtatc . . .

. . .  caaaaatagggttaatatgaatctcgatctccattttgttcatcgtattcaa
caacaagccaaaactcgtacaaatatgaccgcacttcgctataaagaacacggcttgtgg
cgagatatctcttggaaaaactttcaagagcaactcaatcaactttctcgagcattgctt
gctcacaatattgacgtacaagataaaatcgccatttttgcccataatatggaacgttgg
gttgttcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
acaatcgttgacattgcgaccttacaaattcgagcaatcacagtgcctatttacgcaacc
aatacagcccagcaagcagaatttatcctaaatcacgccgatgtaaaaattctcttcgtc
ggcgatcaagagcaatacgatcaaacattggaaattgctcatcattgtccaaaattacaa
aaaattgtagcaatgaaatccaccattcaattacaacaagatcctctttcttgcacttgg
```

Molecular Biology Information: Protein Sequence

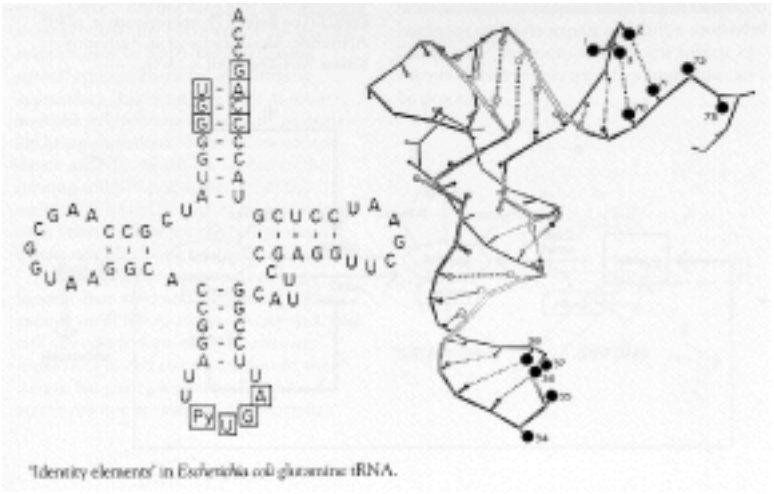
- 20 letter alphabet
  - ◊ ACDEFGHIKLMNPQRSTVWY but not BJOUXZ
- Strings of ~300 aa in an average protein (in bacteria), ~200 aa in a domain
- ~200 K known protein sequences

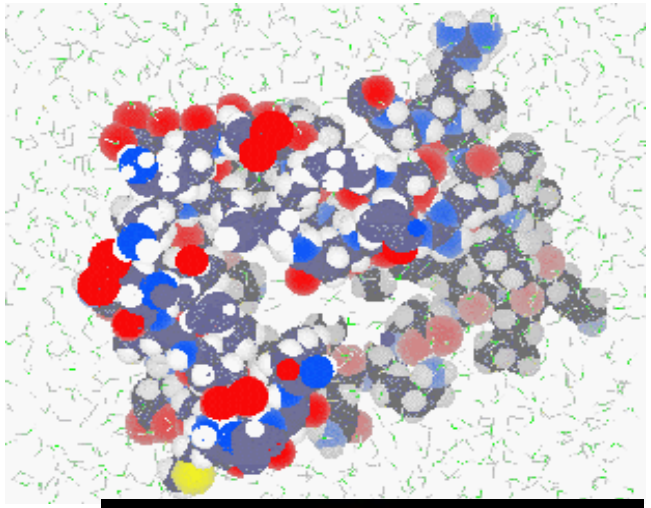
```
d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr__ LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPWN-LPADLAWFKRNTL--------NKPVIMGRHTWESI
d3dfr__ TAFLWAQDRDGLIGKDGHLPWH-LPDDLHYFRAQTV--------GKIMVVGRRTYESF

d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr__ LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPW-NLPADLAWFKRNTLD--------KPVIMGRHTWESI
d3dfr__ TAFLWAQDRNGLIGKDGHLPW-HLPDDLHYFRAQTVG--------KIMVVGRRTYESF

d1dhfa_ VPEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
d8dfr__ VPEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKP
d4dfra_ ---G-RPLPGRKNIILS-SQPGTDDRV-TWVKSVDEAIAACGDVP------EIMVIGGGRVYEQFLPKA
d3dfr__ ---PKRPLPERTNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQHLDQ----ELVIAGGAQIFTAFKDDV

d1dhfa_ -PEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
d8dfr__ -PEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKP
d4dfra_ -G---RPLPGRKNIILSSSQPGTDDRV-TWVKSVDEAIAACGDVPE-----  IMVIGGGRVYEQFLPKA
d3dfr__ -P--KRPLPERTNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQHLD----QELVIAGGAQIFTAFKDDV
```

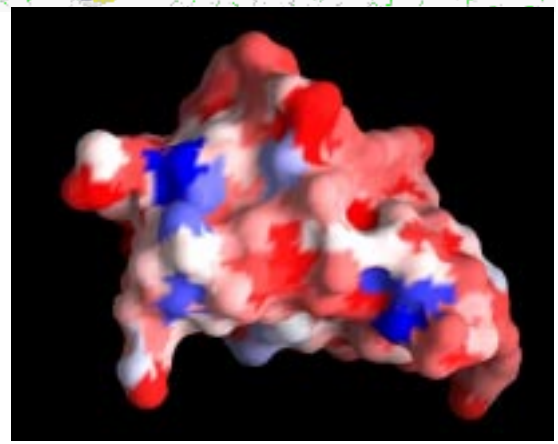
Molecular Biology Information: Macromolecular Structure

- DNA/RNA/Protein
  - ◊ Almost all protein

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



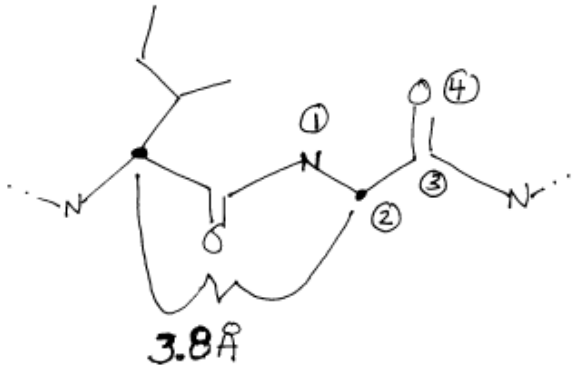




Molecular Biology Information: Protein Structure Details

- 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
    - => ~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
```

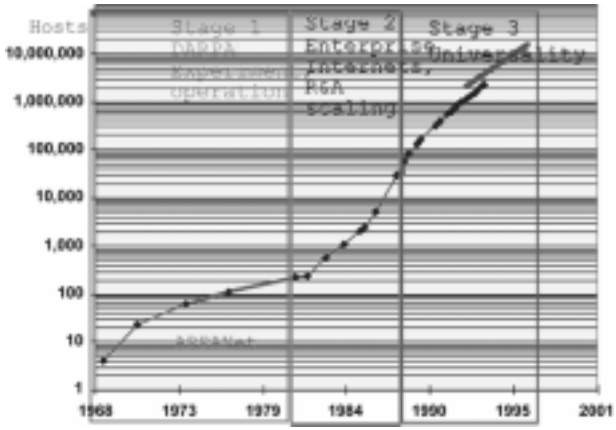


Explonential Growth of Data Matched by Development of Computer Technology

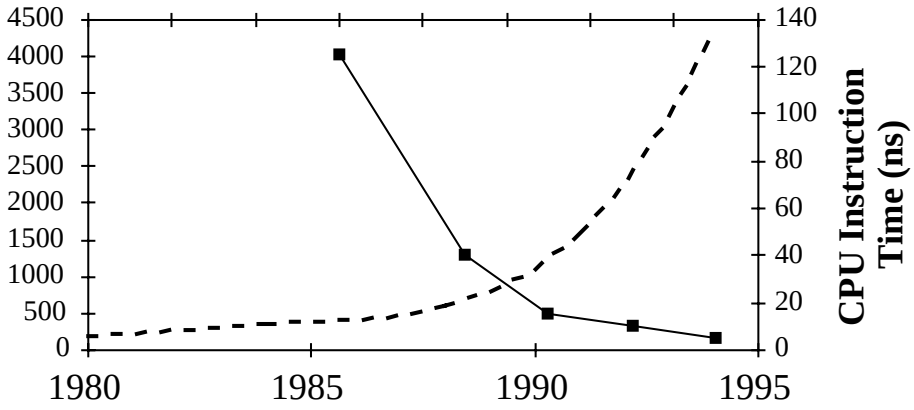
- CPU vs Disk & Net
  - ◊ As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Driving Force in Bioinformatics

(Internet picture adapted from D Brutlag, Stanford)

Internet Hosts



Num. Protein Domain Structures



Overview

- Why interesting?
  - ◊ Not tremendous success, but many methods brought to bear.
  - ◊ What does difficulty tell about protein structure?

- Start with TM Prediction (Simpler)
- Basic GOR Sec. Struc. Prediction
- Better GOR
  - ◊ GOR III, IV, semi-parametric improvements, DSC
- Other Methods
  - ◊ NN, nearest nbr.

What secondary structure prediction tries to accomplish?

Credits: Rost et al. 1993;
Fasman & Gilbert, 1990

- Not Same as Tertiary Structure Prediction -- no coordinates
- Need torsion angles of terms + slight diff. in torsions of sec. str.

(a) Residue-by-residue comparison of experimentally observed (OBS) and predicted [COM[22], ETH[26], PHD (Ref. 35 and B. Rost and C. Sander, submitted)] structures of the catalytic subunit of the cAMP-dependent protein kinase (1cpk). 'AA' is the amino acid sequence taken from Protein Data Bank entry 1cpk residues 27-287). Secondary structure: H = α-helix, E = β-sheet (extended), blank = loop. Predicted α-helices and β-strands that have insufficient overlap with an observed segment of the same type are underlined. Note the relatively good prediction of the location of segments for the ETH and PHD methods and overprediction of α-helices for the COM method.

(b) Ribbon view of the domain used in this blind test. The X-ray structure of catalytic subunit of the cAMP-dependent protein kinase. Drawn using Molscript[44].

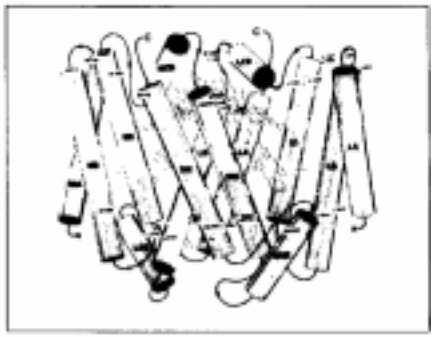


Figure 1
Column model for the core of the reaction center from *Rsp. viridis*. Reproduced, with permission, from Ref. 18.

```
Sequence   RPDFCLEPPYTGPCKARIIRYFYNAKAGLVQTFVYGGCRAKRNNFKSAEDAMRTCGGA
Structure  CCGGGGCCCCCCCCCCCEEEEEEETTTTEEEEEEECCCCCTTTTBTTHHHHHHHHHCC
```

Some TM scales:

GES

F	-3.7
M	-3.4
I	-3.1
L	-2.8
V	-2.6
C	-2.0
W	-1.9
A	-1.6
T	-1.2
G	-1.0
S	-0.6
P	+0.2
Y	+0.7
H	+3.0
Q	+4.1
N	+4.8
E	+8.2
K	+8.8
D	+9.2
R	+12.3

KD

I	4.5
V	4.2
L	3.8
F	2.8
C	2.5
M	1.9
A	1.8
G	-0.4
T	-0.7
W	-0.9
S	-0.8
Y	-1.3
P	-1.6
H	-3.2
E	-3.5
Q	-3.5
D	-3.5
N	-3.5
K	-3.9
R	-4.5

How to use GES to predict proteins

- Transmembrane segments can be identified by using the GES hydrophobicity scale (Engelman et al., 1986). The values from the scale for amino acids in a window of size 20 (the typical size of a transmembrane helix) were averaged and then compared against a cutoff of -1 kcal/mole. A value under this cutoff was taken to indicate the existence of a transmembrane helix.
- H-19(i) = [H(i-9)+H(i-8)+...+H(i) + H(i+1) + H(i+2) + . . . + H(i+9)] / 19

Graph showing Peaks in scales

Illustrations Adapted From: von Heijne, 1992; Smith notes, 1997

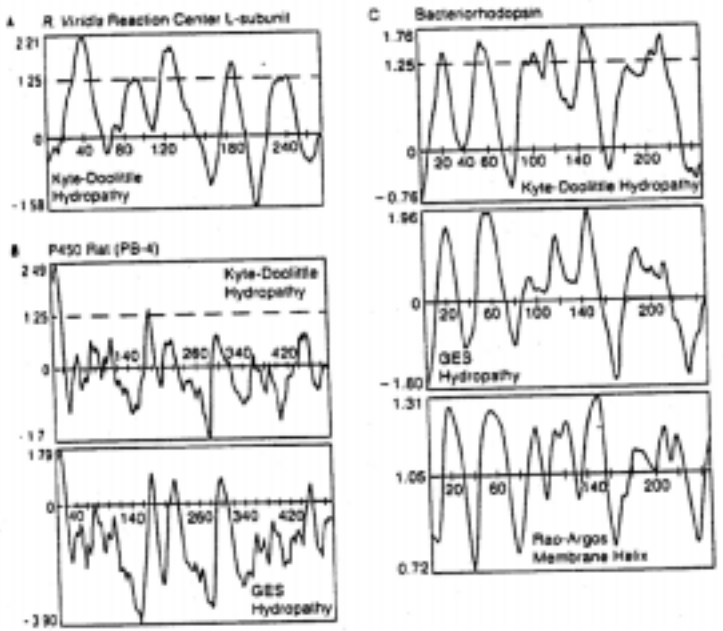
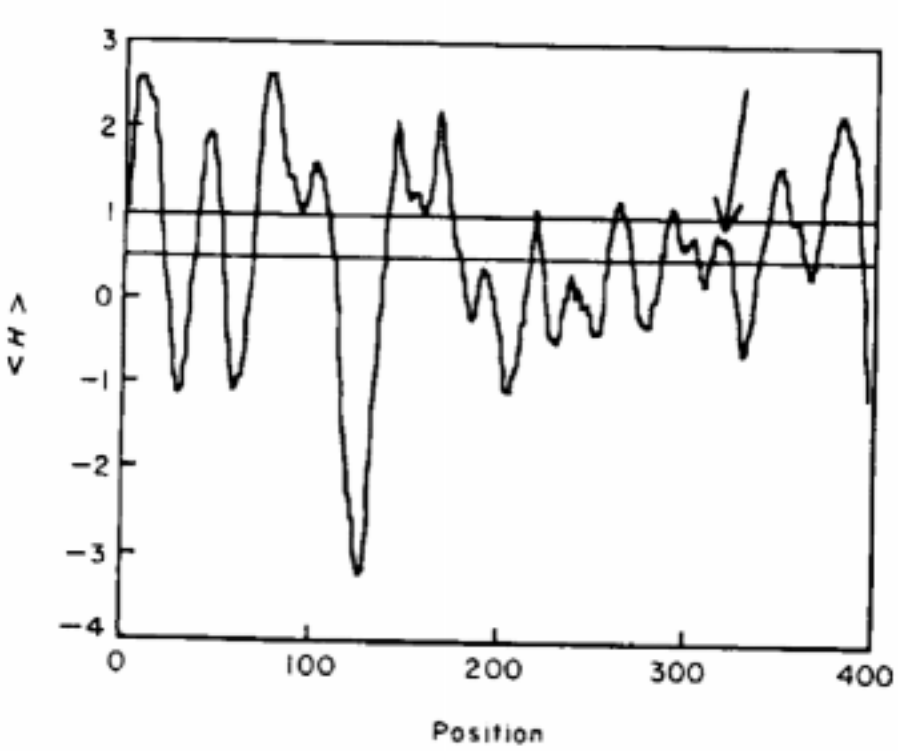
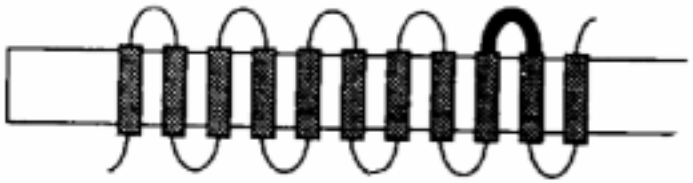


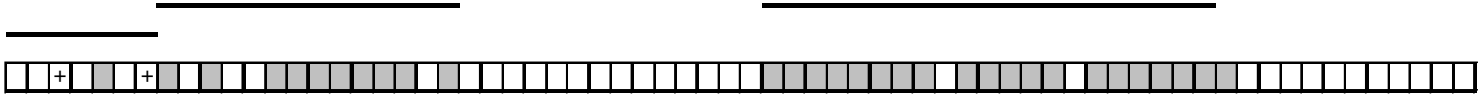
Figure 3.12. Representative profiles of three membrane proteins used to predict membrane-spanning helices. The amino acid scales of Kyte-Doolittle (804), Goldman-Engelman-Steitz (GES) (389), and Rao-Argos (1194) were used. A computer software package (SEQANAL) provided by Dr. A. Crofts (Univ. of Illinois) was used to generate these profiles. For comparative purposes, the Kyte-Doolittle and GES plots were obtained using a window of 19 residues and then smoothed using a second pass with a window of 7. The average value at each residue position is plotted as a function of residue number starting with the amino terminus on the left in each case. The values plotted for the Kyte-Doolittle and GES scales represent average hydropathy and transfer free energy per residue (kcal/mol). The Rao-Argos plot used a span of 7 residues and was smoothed using two additional passes with the same span of 7, as recommended by the authors. The scale values reflect the relative preference for being in a membrane-spanning helix. Note that the version of the GES algorithm which was used does not take into account possible ion pair formation. See text for details.





Removing Signal sequences

- Initial hydrophobic stretches corresponding to signal sequences for membrane insertion were excluded. (These have the pattern of a charged residue within the first 7, followed by a stretch of 14 with an average hydrophobicity under the cutoff).



Relation between Energy Scales and Statistics

- Statistical Mechanics: Energies give Probabilities of Observing an Object
- Inverse Statistical Mechanics: Derive Energies from observing probabilities in the database
- $E = = - kT \ln P$
- $dE = E1 - E2 = -kT \ln P1/P2$
- dE = E(state-1) - E(~state-1)
- $dE = -kT \ln P(state\text{-}1)/P(\sim state\text{-}1)$
- P(state-1)/P(~state-1) = odds ratio => dE = lod score

Ex. Pr(S) probability that residue j has secondary structure i

- Problem of DB Bias
- f(A) = frequency of residue A to have a helical conf. in db
- f(A,i) = f(A) at position i in a particular sequence
- E(α)=statistical energy of helix over a window
- p(i, α) = probability that residue i is in a helix

$$E_\alpha = \sum_i^N \ln f_\alpha^i$$

$$p_\alpha^i = \frac{e^{-Ea/RT}}{\sum_j^N e^{-E_j/RT}}$$

Statistics Based Methods: Persson & Argos

- Propensity P(A) for amino acid A to be in the middle of a TM helix or near the edge of a TM helix

$$P(A) = \frac{\dfrac{n(A,\mathrm{TM})}{\sum_A n(A,\mathrm{TM})}}{\dfrac{n(A,\mathrm{everywhere})}{\sum_A n(A,\mathrm{everywhere})}}$$

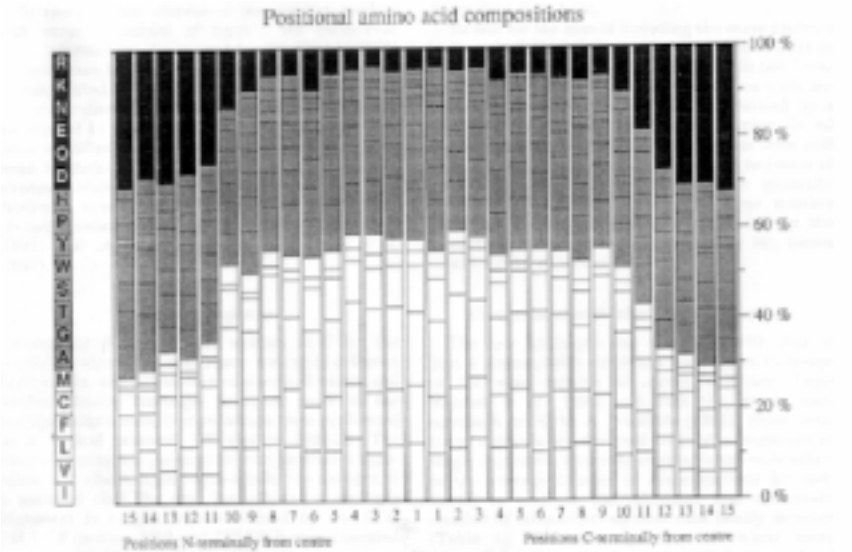
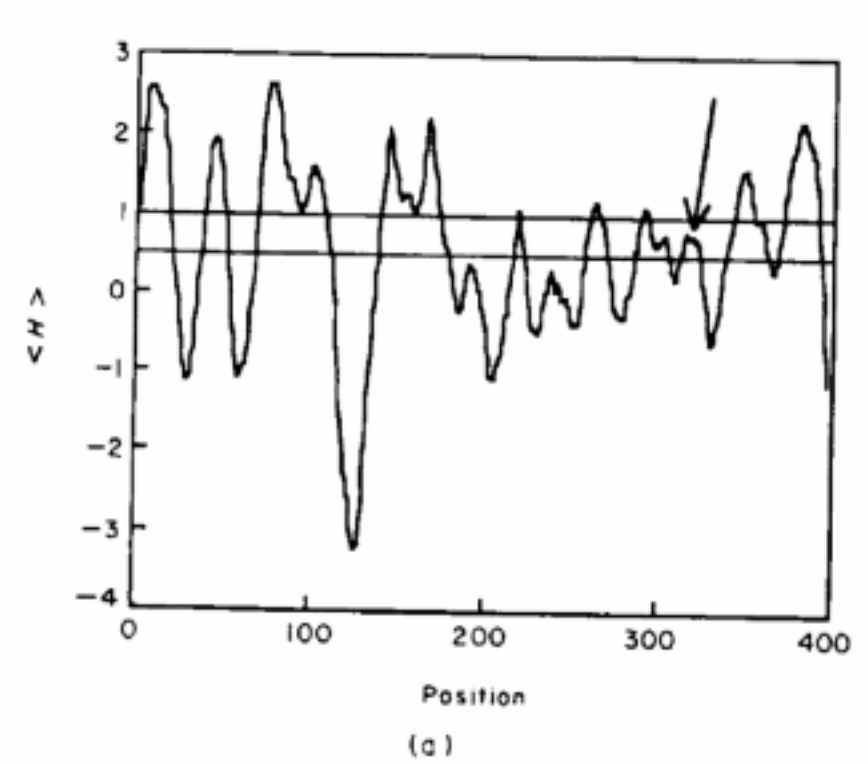


Figure 1. Positional amino acid compositions of transmembrane segments. Bar chart showing the amino acid compositions for 15 N and C-terminal positions relative to the centre of putative transmembrane segments listed in feature tables of the Swiss-Prot database. For each position, the percentage contribution of each amino acid type is shown according to the hydrophilic (top) to hydrophobic (bottom) order given in the ruler bar at the left. The hydrophobic residue contributions are illustrated in white, the hydrophilic in dark-gray, and intermediate in light-gray. The compositions of positions 11 to 15 at the N-terminal side and 12 to 15 at the C-terminal side differ significantly from the others, especially for the most hydrophobic and charged/hydrophilic residues. These results suggest that in general transmembrane spans consist of a hydrophobic portion 21 residues in length.

Illustration Credits: Persson & Argos, 1994

Refinements: Charge on the Outside, Positive Inside Rule

- for marginal helices, decide on basis of R+K inside (cytoplasmic)



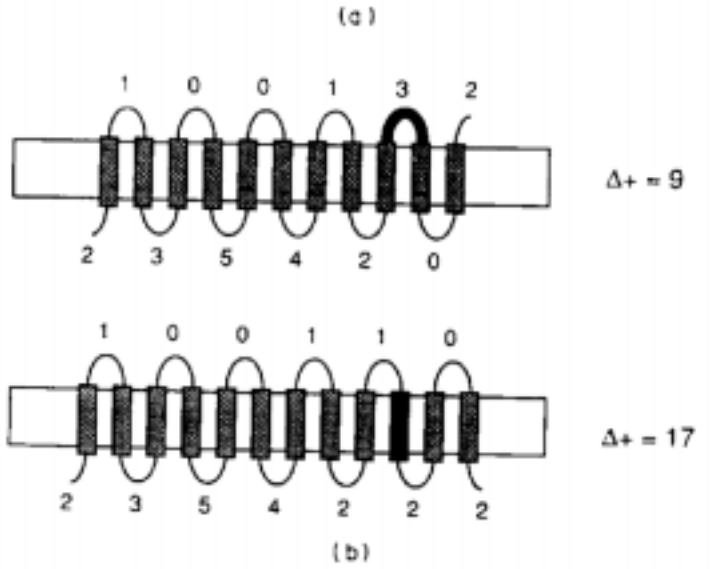
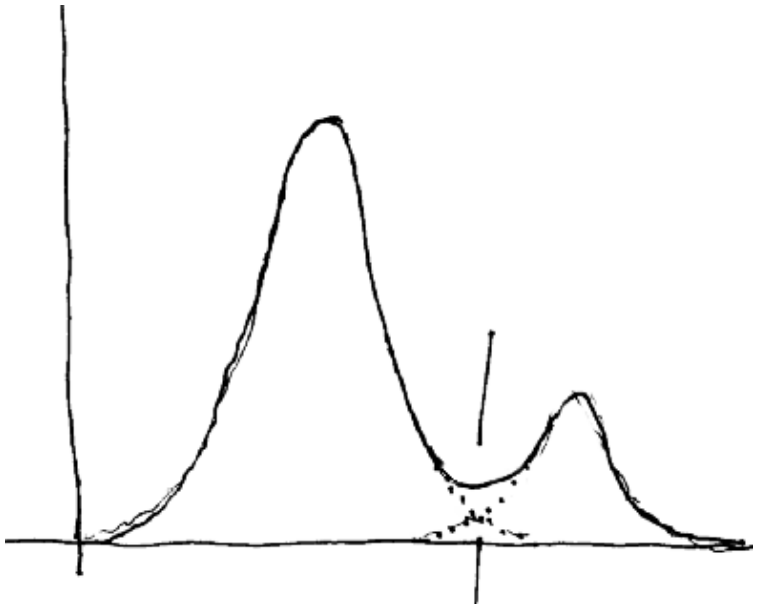


Figure 4. (a) Hydrophobicity plot for the SecY protein. The upper and lower cutoffs are marked. A tentative transmembrane segment with a mean hydrophobicity falling between the 2 cutoffs is marked by an arrow. (b) Two possible topologies for the SecY protein based on the hydrophobicity plot. The putative transmembrane segment is shown in black. The number of Arg + Lys residues is shown next to each polar segment. Note that the correct alternative (bottom, including the putative transmembrane segment) has a much higher charge-bias than the incorrect one.

Credits: von Heijne, 1992

Refinements: MaxH



- How to train to find right threshold? Not that many TM helices
- Marginal TM helices are not that hydrophobic but 1/3 of TM's are very hydrophobic, so focus on these.
- Sosui, Klein & Delisi, Boyd
- Discriminant analysis: set threshold to be best partition of dataset

Secondary Structure Prediction: GOR is based on Info. Theory

- Analogous to P&A,
  but more data and more involved stats
- Based on information theory, want to maximize the "Information"
  - ◊ I(x;y) = information that event y carries about the occurrence of x
  - ◊ I(x;y) = log (P(x|y)/P(x))
  - ◊ I = 0 for no information
  - ◊ I > 0 if A favors helix
  - ◊ I < 0 if disfavors
- Information sort of like entropy
  I increases, entropy increases

GOR: Simplifications

- For independent events just add up the information
- $I(S_j ; R_1, R_2, R_3,...R_{last})$ = Information that first through last residue of protein has on the conformation of residue j (S_j)
  - ◊ Could get this just from sequence sim. or if same struc. in DB (homology best way to predict sec. struc.!)
- Simplify using a 17 residue window: $I(S_j=H ; R[j-8], R[j-7], ..., R[j], ... R[j+8])$
- Difference of information for residue to be in helix relative to not: I(dSj;y) = I(Sj=H;y)-I(Sj=~H;y)
  - ◊ odds ratio: I(dSj;y)= ln P(Sj;y)/P(~Sj;y)
  - ◊ I determined by observing counts in the DB, essentially a lod value

Basic GOR

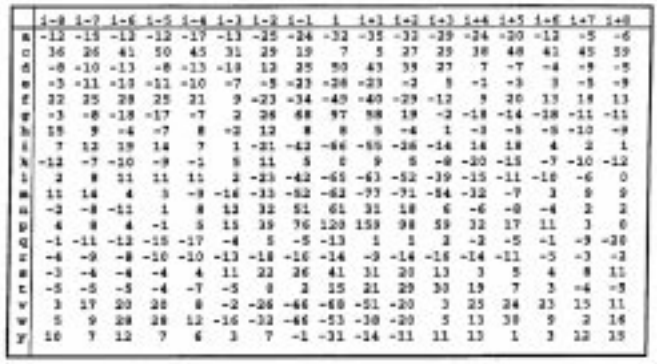
- Pain & Robson, 1971;
  Garnier, Osguthorpe, Robson, 1978
- I ~ sum of I(Sj,R[j+m]) over 17 residue window centered on j and indexed by m
  - ◊ I(Sj,R[j+m]) = information that residue at position m in window has about conformation of protein at position j
  - ◊ 1020 bins=17*20*3
- In Words
  - ◊ Secondary structure prediction can be done using the GOR program (Garnier et al., 1996; Garnier et al., 1978; Gibrat et al., 1987). This is a well-established and commonly used method. It is statistically based so that the prediction for a particular residue (say Ala) to be in a given state (i.e. helix) is directly based on the frequency that this residue (and taking into account neighbors at +/- 1, +/- 2, and so forth) occurs in this state in a database of solved structures. Specifically, for version II of the GOR program (Garnier et al., 1978), the prediction for residue i is based on a window from i-8 to i+8 around i, and within this window, the 17 individual residue frequencies (singlets),

Directional Information

helix

strand

coil



	i-8	i-7	i-6	i-5	i-4	i-3	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	i+6	i+7	i+8
a	-12	-15	-12	-12	-17	-13	-25	-24	-32	-35	-32	-29	-24	-20	-12	-5	-6
c	36	26	41	50	45	31	29	19	7	5	27	29	38	48	41	45	59
d	-8	-10	-13	-8	-13	-18	13	25	50	43	39	27	7	-7	-4	-9	-5
e	-3	-11	-13	-11	-10	-7	-5	-23	-28	-23	-3	5	-1	-3	3	-5	-9
f	32	25	28	25	21	9	-23	-34	-45	-40	-29	-12	9	20	13	14	13
g	-3	-8	-18	-17	-7	2	26	68	97	98	19	-2	-18	-14	-18	-11	-11
h	19	9	-4	-7	8	-2	12	8	8	5	-4	1	-3	-5	-5	-10	-9
i	7	13	19	14	7	1	-21	-43	-66	-59	-36	-14	14	18	4	2	1
k	-12	-7	-10	-9	-1	5	11	5	0	9	5	-8	-20	-15	-7	-10	-12
l	2	8	11	11	11	2	-23	-42	-65	-63	-32	-39	-15	-11	-16	-6	0
m	11	14	4	3	-8	-16	-33	-52	-62	-77	-71	-54	-32	-7	3	8	9
n	-2	-8	-11	1	8	12	32	51	61	31	18	6	-6	-8	-4	2	2
p	4	8	4	-1	5	15	39	76	120	159	98	59	32	17	11	3	0
q	-1	-11	-12	-15	-17	-4	5	-5	-13	1	1	2	-2	-5	-1	-9	-20
r	-4	-9	-8	-10	-10	-13	-18	-16	-14	-9	-14	-16	-14	-11	-5	-3	-2
s	-3	-4	-4	-4	4	11	22	36	41	31	20	13	3	5	4	8	11
t	-5	-5	-5	-4	-7	-5	8	3	15	21	29	30	19	7	3	-4	-5
v	3	17	20	28	8	-2	-26	-46	-68	-51	-20	3	25	24	23	15	11
w	5	9	28	28	12	-16	-33	-46	-53	-30	-20	5	13	30	9	2	16
y	10	7	12	7	6	3	7	-1	-31	-14	-11	11	13	1	3	12	15

Credits: King & Sternberg, 1996

Table 3. *Directional informational parameters: I(S,j) = x:x'; R(j = m) for residue position versus residue type for α-helices*[a]

	i-8	i-7	i-6	i-5	i-4	i-3	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	i+6	i+7	i+8
a	19	21	22	24	34	36	44	47	60	60	53	50	44	40	31	23	24
c	-47	-45	-44	-47	-44	-36	-44	-55	-56	-58	-54	-55	-58	-58	-59	-53	-66
d	14	15	14	15	17	21	15	17	-7	-11	-31	-42	-28	-12	-8	1	-5
e	14	16	15	20	26	27	34	52	62	57	32	15	19	12	6	7	9
f	-19	-14	-10	-4	-2	-1	6	-1	10	10	12	12	-4	-5	2	0	2
g	5	2	1	-5	-22	-30	-50	-70	-92	-52	-28	-21	-13	-17	-8	-6	-6
h	-22	-20	-9	-10	-19	-10	-14	-7	-11	-4	0	-3	-2	2	6	11	12
i	7	7	0	0	1	1	2	-5	1	2	1	7	-6	-3	10	8	6
k	-2	-1	-1	-1	-6	-9	-6	5	17	17	21	27	35	33	21	22	23
l	0	-1	0	6	9	16	30	33	45	47	51	53	37	32	30	25	18
m	4	3	15	23	30	30	39	36	45	54	57	53	44	29	30	14	1
n	2	3	2	-5	-9	-10	-16	-17	-31	-16	-17	-16	-9	-8	-9	-10	-5
p	-12	-15	-14	-19	-23	-25	-30	-48	-82	-195	-145	-104	-67	-49	-43	-33	-17
q	-4	3	7	4	13	8	10	24	35	32	31	21	18	18	9	8	6
r	5	3	6	13	7	13	19	27	34	32	36	41	33	29	23	21	18
s	-10	-7	-10	-10	-16	-17	-25	-21	-39	-35	-39	-41	-32	-35	-34	-35	-33
t	1	-1	-6	-8	-6	-11	-16	-25	-48	-47	-48	-46	-34	-31	-34	-26	-24
v	-5	-12	-13	-14	-13	-19	-17	-20	-15	-22	-22	-20	-26	-19	-15	-10	-5
w	0	-4	-12	-19	-7	14	16	12	18	17	12	8	1	-6	1	3	-13
y	-22	-19	-17	-20	-16	-21	-30	-32	-8	-10	-4	-12	-17	-9	-10	-14	-15

[a]Note that the convention used is the reverse of that adopted by (Garnier et al., 1978), for example the first entry for alanine at position j-8 is the amount of information that an alanine residue eight positions toward the N terminus has for predicting an α-helix

Table 4. *Directional informational parameters for residue position versus residue type for β-strands*

	i-8	i-7	i-6	i-5	i-4	i-3	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	i+6	i+7	i+8
a	-8	-7	-13	-17	-23	-33	-26	-32	-43	-37	-30	-30	-26	-27	-26	-25	-25
c	3	13	-9	-20	-15	-3	9	33	47	51	21	19	9	-5	7	-5	-14
d	-7	-5	0	-9	-4	-14	-42	-73	-83	-59	-21	10	22	24	16	11	13
e	-14	-5	-5	-11	-21	-27	-45	-44	-57	-54	-46	-29	-25	-12	-12	-2	0
f	-9	-20	-32	-34	-30	-12	24	44	49	39	24	2	-9	-23	-24	-29	-23
g	-3	9	24	29	34	30	18	-23	-48	-27	6	27	39	38	33	23	23
h	6	11	17	22	12	16	0	-2	3	-2	5	3	8	4	-1	1	-3
i	-21	-30	-31	-21	-12	-3	26	58	76	64	33	11	-14	-24	-20	-14	-11
k	20	12	15	14	8	4	-8	-14	-25	-40	-39	-27	-20	-24	-20	-15	-15
l	-2	-10	-18	-27	-30	-27	-6	15	27	21	2	-19	-31	-29	-28	-26	-25
m	-22	-26	-29	-40	-31	-17	-7	23	24	28	17	2	-15	-31	-53	-36	-16
n	1	8	14	5	0	-6	-30	-65	-62	-28	-6	11	18	21	16	10	3
p	9	7	12	24	20	8	-22	-65	-108	-64	-8	17	25	30	32	31	21
q	6	12	8	16	8	-5	-22	-27	-30	-52	-49	-34	-22	-17	-9	2	20
r	0	8	3	-3	5	2	1	-14	-26	-32	-30	-35	-27	-26	-25	-25	-21
s	16	14	17	19	14	5	-3	-13	-15	-4	15	27	32	32	31	28	21
t	6	8	14	15	16	21	19	25	31	22	13	9	12	25	34	34	34
v	1	-11	-15	-11	4	25	51	75	91	81	49	19	-6	-12	-16	-11	-11
w	-8	-8	-28	-19	-9	5	23	44	45	30	13	-18	-22	-40	-15	-7	-9
y	13	13	4	14	12	20	24	37	48	31	20	-1	2	11	7	0	-4

Types of Residues

Table 3. *Directional informational parameters: I(Sj = x:x': Rj + m) for residue position versus residue type for α-helices*[a]

	i-8	i-7	i-6	i-5	i-4	i-3	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	i+6	i+7	i+8
a	19	21	22	24	34	36	44	47	60	60	53	50	44	40	31	23	24
c	-47	-45	-44	-47	-44	-36	-44	-45	-56	-58	-54	-55	-58	-58	-59	-53	-66
d	14	15	14	15	17	21	15	17	-7	-11	-31	-42	-28	-12	-8	1	-5
e	14	16	15	20	26	27	34	52	62	57	32	15	19	12	6	7	9
f	-19	-14	-10	-4	-2	-1	6	-1	10	10	12	12	-4	-5	2	0	2
g	5	2	1	-5	-22	-30	-50	-70	-92	-52	-28	-21	-13	-17	-8	-6	-6
h	-22	-20	-9	-10	-19	-10	-14	-7	-11	-4	0	-3	-2	2	6	11	12
i	7	7	0	-3	1	1	2	-5	1	2	1	7	-6	-3	10	8	-1
k	-2	-1	-1	-1	-6	-9	-6	5	17	17	21	27	35	33	21	22	23
l	0	-1	0	-2	9	16	30	33	45	47	51	53	37	32	30	25	19
m	4	3	15	23	30	30	39	36	45	54	57	53	44	29	30	14	1
n	2	3	2	-5	-9	-10	-16	-15	-31	-16	-17	-16	-9	-8	-9	-10	-5
p	-12	-15	-14	-19	-23	-25	-30	-48	-82	-195	-141	-104	-67	-49	-43	-33	-17
q	-4	3	7	4	13	8	10	24	35	32	31	21	18	18	9	8	6
r	5	3	6	13	7	13	19	27	34	32	36	41	33	29	23	21	18
s	-10	-7	-10	-10	-16	-17	-25	-21	-39	-35	-39	-41	-32	-35	-34	-35	-33
t	1	-1	-6	-8	-6	-11	-16	-25	-48	-47	-48	-46	-34	-31	-34	-26	-24
v	-5	-12	-13	-14	-13	-19	-17	-20	-15	-22	-22	-20	-26	-19	-15	-10	-5
w	0	-4	-12	-19	-7	14	16	12	18	17	12	8	1	-6	1	3	-13
y	-22	-19	-17	-20	-16	-21	-30	-32	-8	-10	-4	-12	-17	-9	-10	-14	-15

[a]Note that the convention used is the reverse of that adopted by (Garnier et al., 1978), for example the first entry for alanine at position j-8 is the amount of information that an alanine residue eight positions toward the N terminus has for predicting an α-helix at position j.

Credits: King & Sternberg, 1996

- Group I favorable residues and Group II unfavorable one:
- A, E, L -> H; V, I, Y, W, C -> E; G, N, D, S -> C
- P complex; largest effect on proceeding residue
- Some residues favorable at only one terminus (K)

GOR III

- Improvements in GOR -- Full GOR decomposition
- $I(S_j ; R_1, R_2, R_3,...R_{last})$ = Information that first through last residue of protein has on the conformation of residue j (S_j)
  - ◊ looked at singlets, now pairs, eventually triplets....
- GOR III
- I(Sj; R[j+m], R[j]) = information that pair of residues at postions 0 & m in window has about conformation of protein at position j
  - ◊ 16 pairs =>16*20*20*3=19200 bins

GOR IV

- I(Sj; R[j+m], R[j+n]) = the frequencies of all 136 (=16*17/2) possible di-residue pairs (doublets) in the window.
  - ◊ 20*20*3*16*17/2=163200 pairs
- Parameter Explosion Problem: 1000 dom. struc. * 100 res./dom. = 100k counts, over how many bins
- Dummy counts for low values (Bayes)

Assessment

- Q3 + other assess, 3x3
- Q3 = total number of residues predicted correctly over total number of residues
- GOR gets 65%
  - ◊ sum of diagonal over total number of residue -- (14K+5K+21K)/ 64K
- Under predict strands & to a lesser degree, helices: 5.9 v 4.1, 10.9 v 10.6

TABLE II
GLOBAL RESULTS FOR DATABASE PREDICTION

	Observed			
	H	E	C	Total
Predicted				
H	14,460	3094	4790	22,344
E	1124	4965	2089	8178
C	6002	5546	21,496	33,044
Total	21,586	13,605	28,375	63,566
Q_{prd} [a]	64.7	60.7	65.1	
Q_{obs} [b]	67.0	36.5	75.8	
Q_3 [c] = 64.4%				

[a] Number of correctly predicted residues/number of predicted residues.
[b] Number of correctly predicted residues/number of observed residues
[c] Total number of correctly predicted residues/total number of residues.

Credits: Garnier et al., 1996

Training and Testing Set

- Cross Validation: Leave one out, seven-fold

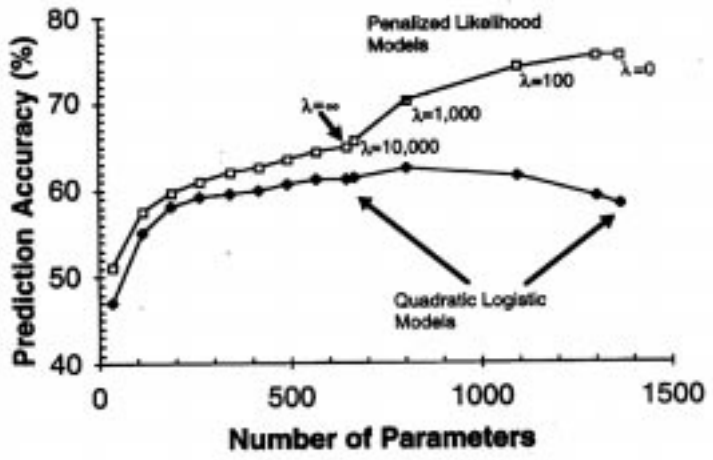


Figure 2. Comparison of prediction accuracy (correctly predicted residues as a proportion of total residues) versus effective number of parameters for linear-logistic models (number of parameters ≤640) and penalized likelihood models for crossvalidated (◆) and uncrossvalidated (❑) results. The values of the penalty parameter λ are shown.

Credits: Munson, 1995; Garnier et al., 1996

TABLE I
DATABASE PROTEINS[a]

1aaj.x	1aak.x	1aap.a	1aba.x	1abk.x	1abm.a	1add.x
1ads.x	1alk.a	1aoz.a	1apa.x	1apm.e	1arb.x	1atr.x
1avh.a	1ayh.x	1bab.a	1bbh.a	1bbp.a	1bet.x	1bge.a
1bll.e	1bmd.a	1bov.a	1bpb.x	1brs.d	1btc.x	1c2r.a
1caj.x	1cau.a	1cau.b	1cde.x	1cdt.a	1cew.i	1cgt.x
1chm.a	1cmb.a	1cob.a	1col.a	1cpc.a	1cpc.b	1cpt.x
1crl.x	1cse.i	1ctf.x	1ctm.x	1cus.x	1ddt.x	1dhr.x
1dog.x	1dsb.a	1eaf.x	1eco.x	1ede.x	1end.x	1epa.a
1fba.a	1fdd.x	1fha.x	1fia.a	1fkb.x	1fna.x	1fnr.x
1fxi.a	1gal.x	1gd1.o	1gdh.a	1gky.x	1glt.x	1gmf.a
1gof.x	1gox.x	1gp1.a	1gpb.x	1gpr.x	1gsr.a	1hbq.x
1hdx.a	1hiv.a	1hlb.x	1hle.a	1hmy.x	1hoe.x	1hpl.a
1hrh.a	1hsl.a	1huw.x	1ifc.x	1ipd.x	1isu.a	1ith.a
1l29.x	1le4.x	1len.a	1lga.a	1lis.x	1lla.x	1lmb.3
1lts.a	1lts.d	1mdc.x	1mgn.x	1min.a	1min.b	1mjc.x
1mpp.x	1mup.x	1nar.x	1nba.a	1ndk.x	1noa.x	1nsb.a
1nxb.x	1ofv.x	1olb.a	1omf.x	1omp.x	1onc.x	1osa.x
1pda.x	1pfk.a	1pgb.x	1pgd.x	1phh.x	1php.x	1pii.x
1plf.a	1poc.x	1poh.x	1pox.a	1ppa.x	1ppf.e	1ppf.i
1ppn.x	1prc.c	1prc.h	1prc.l	1prc.m	1pts.a	1pya.a
1pya.b	1pyd.a	1rcb.x	1rec.x	1rib.a	1rnd.x	1rop.a
1rve.a	1s01.x	1sac.a	1sbp.x	1ses.a	1sgt.x	1sha.a
1shf.a	1sim.x	1slt.b	1snc.x	1spa.x	1stf.i	1tbe.a
1tca.x	1tie.x	1tml.x	1tnd.a	1tpl.a	1trb.x	1trk.a
1tro.a	1ttb.a	1utg.x	1vaa.a	1vaa.b	1vmo.a	1wht.a
1wht.b	1wsy.a	1wsy.b	1yhb.x	1zaa.c	256b.a	2aai.b
2aza.a	2bop.a	2ccy.a	2cdv.x	2chs.a	2cmd.x	2cp4.x
2cpl.x	2cro.x	2ctc.x	2cts.x	2cyp.x	2dnj.a	2er7.e
2hbg.x	2hhm.a	2hip.a	2hpd.a	2ihl.x	2lh2.x	2liv.x
2mhr.x	2mnr.x	2msb.a	2mta.c	2mta.h	2mta.l	2pf1.x
2pia.x	2pol.a	2por.x	2reb.x	2rn2.x	2rsl.a	2sar.a
2sas.x	2scp.a	2sga.x	2sn3.x	2spc.a	2tgi.x	2tmd.a
2tpr.a	2tsc.a	3aah.a	3aah.b	3adk.x	3b5c.x	3cd4.x
3chy.x	3cla.x	3cox.x	3dfr.x	3eca.a	3gap.a	3gbp.x
3ink.c	3rub.l	3rub.s	3sdh.a	3tgl.x	451c.x	4blm.a
4enl.x	4fgf.x	4gcr.x	4ts1.a	4xis.x	5fbp.a	5p21.x
5tim.a	6fab.h	6fab.l	6taa.x	8abp.x	8acn.x	8atc.a
8atc.b	8cat.a	8i1b.x	8rxn.a	8tln.e	9ldt.a	9rnt.x
9wga.a						

[a] The database was prepared by J. M. Levin and checked for homologous sequences with the help of V. Di Francesco. This database has been modified to restore the total length of the sequences as defined in the SEQRES field of the Protein Data Bank (PDB) file (the DSSP program omits residues whose coordinates are missing in the PDB file, and thus if this occurs in the middle of the polypeptide chain it is split into two or more chains). Residues having no coordinates were assigned the conformation X and were not taken into account for the prediction accuracy although the prediction was done with the whole sequence length. The PDB code is followed by the chain name a, b, c, d, h (heavy), l (light), x (one chain only), e (enzyme), or i (inhibitor).

Is 100% Accuracy Possible?

Quoted from Barton (1995):

One problem that has arisen is how to evaluate secondary structure predictions. For prediction of a single protein sequence one might expect the best residue by residue accuracy to be 100%. It is not possible to define the secondary structure of a protein exactly, however. There is always room for alternative interpretations of where a helix or strand begins or ends so failure of a prediction to match exactly the secondary structure definition is not a disaster [24]. The problem of evaluation is more complicated for prediction from multiple sequences, as the prediction is a consensus for the family and so is not expected to be 100% in agreement with any single family member. The expected range in accuracy for a perfect consensus prediction is a function of the number, diversity and length of the sequences. Russell and I have calculated estimates of this range [11].

Simple residue by residue percentage accuracy has long been the standard method of assessment of secondary structure predictions. Although a useful guide, high percentage accuracies can be obtained for predictions of structures that are unlike proteins. For example, predicting myoglobin to be entirely helical (no strand or coil) will give over 80% accuracy but the prediction is of little practical use. Rost *et al.* [25] and Wang [26] explore these problems and suggest some alternative measures of predictive success based on secondary structure segment overlap. Although such measures help in an objective assessment of the prediction, there is no complete substitute for visual inspection. By eye, serious errors stand out and predictions of structures that are unlike proteins are usually recognizable. By eye, it is also straightforward to weight the importance of individual secondary structures. For example, prediction of what is in fact a core strand to be a helix would seriously hamper attempts to generate the correct tertiary structure of the protein from the predicted secondary structure, whereas prediction of a non-core helix as coil may have little impact on the integrity of the tertiary structure.

'Predictable' regions of secondary structure

Quoted from Barton (1995):

When recent predictions are examined in the light of the corresponding experimentally determined structures, the results look good. In general, the regions predicted with the highest confidence measure are also the most accurate. For example, Livingstone and I [13] assigned 41% of the tyrosine phosphatase structure with high confidence. Within these regions 88% of the residues were correctly predicted. Interestingly, these figures agree with Rost and Sander's observation that 40% of a sequence will be predicted with >88% accuracy by their method [1]. This agreement suggests that there is a core of 'predictable' regions in a protein. Examination of six blind predictions shows that the most accurately predicted regions are those that have clear periodicity in conservation, where conserved positions either alternate (beta-strand) or have a 1, 4, 5, 8 pattern characteristic of one face of an alpha-helix (CD Livingstone, personal communication). Problems remain with buried alpha-helices that comprise short runs of conserved hydrophobic amino acids. These often look like potential beta-strands and can mislead both automatic and manual predictive methods.

Types of Secondary Structure Prediction Methods

- Parametric Statistical
  - ◊ struc. = explicit numerical func. of the data (GOR)
- Non-parametric
  - ◊ struc. = NON- explicit numerical func. of the data
  - ◊ generalize Neural Net, seq patterns, nearest nbr, &c.
- Semi-parametric: combine both
- single sequence
- multi sequence
  - ◊ with or without multiple-alignment

GOR Semi-parametric Improvements

- Filtering GOR to regularize

$[\neg a, \neg a, c, \mathbf{b}, *, \neg b] \rightarrow \mathbf{c}$

$[\neg a, *, *, \mathbf{a}, b] \rightarrow \mathbf{b}$

$[\neg a, *, *, \mathbf{a}, c] \rightarrow \mathbf{c}$

$[a, *, *, \mathbf{a}, c, *, \neg c] \rightarrow \mathbf{c}$

$[\neg a, \neg a, \mathbf{a}, a, c, \neg a] \rightarrow \mathbf{c}$

$[\neg a, c, \neg c, \mathbf{a}, a, c, \neg a] \rightarrow \mathbf{c}$

$[\neg a, c, c, \mathbf{a}, a, \neg b, \neg a] \rightarrow \mathbf{c}$

$[a, c, *, \mathbf{a}, a, a, \neg a] \rightarrow \mathbf{c}$

$[*, c, *, \mathbf{a}, a, b, \neg a] \rightarrow \mathbf{c}$

$[c, b, b, \mathbf{a}, a, *, a] \rightarrow \mathbf{b}$

$[c, * \ \mathbf{a}, a, \neg a, a] \rightarrow \mathbf{c}$

a = α-helix, b β-strand, c = coil, * = wildcard (α-helix or β-strand or coil) $\neg$ = not.

If the pattern on the left is met in a prediction, then the secondary structure in bold on the left is rewritten as the secondary structure on the right of the rule. For example:

$[b, b, b, \mathbf{a}, c] \rightarrow [b, b, b, \mathbf{c}, c]$

$[b, b, c, \mathbf{a}, c] \rightarrow [b, b, c, \mathbf{c}, c]$

$[b, b, b, \mathbf{a}, b, b, b] \rightarrow [b, b, b, \mathbf{b}, b, b, b].$

Illustration Credits: King & Sternberg, 1996

Multiple Sequence Methods

- Average GOR over multiple seq. Alignment
- The GOR method only uses single sequence information and because of this achieves lower accuracy (65 versus >71 %) than the current "state-of-the-art" methods that incorporate multiple sequence information (e.g. King & Sternberg, 1996; Rost, 1996; Rost & Sander, 1993).

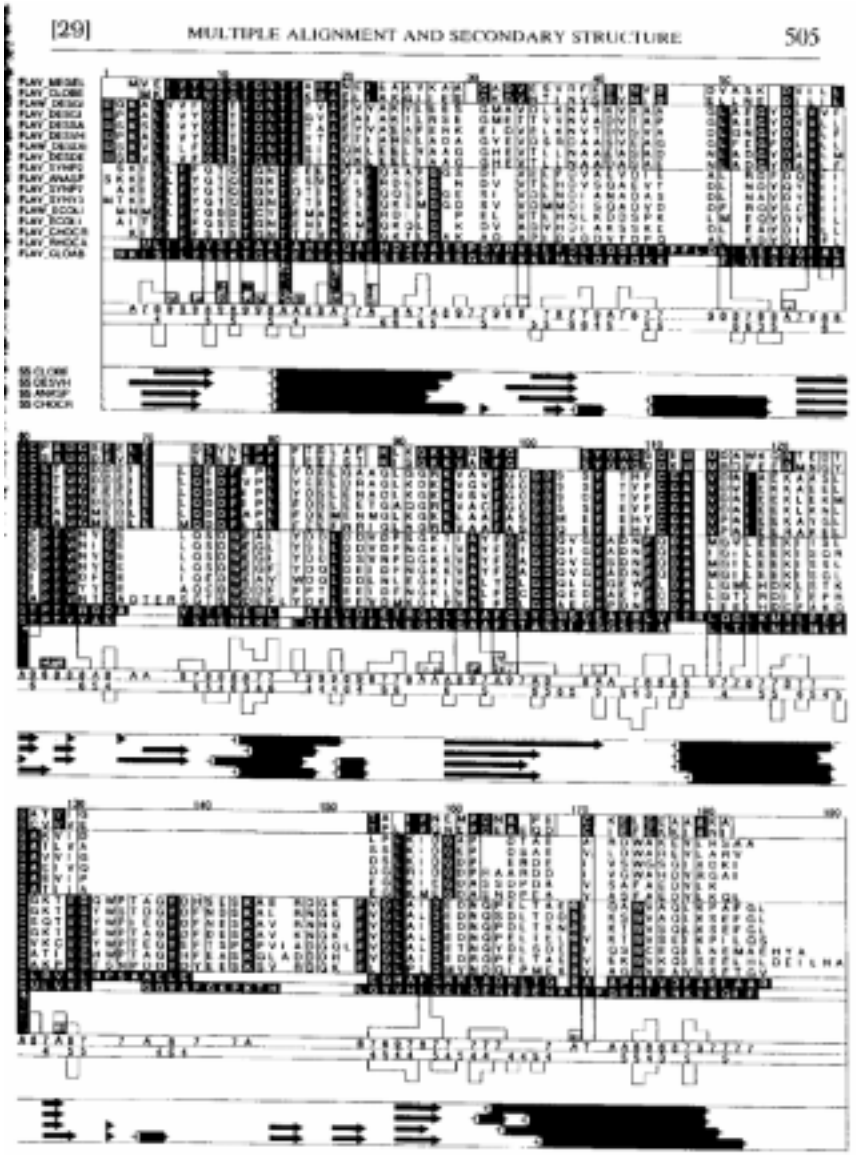


FIG. 5. Conservation analysis of the 17 flavodoxin sequences clustered in Fig. 3. The Taylor Venn diagram was used (Fig. 1) with a threshold of $T = 7$. See text for details.

Illustration Credits: Livingston & Barton, 1996

DSC -- an improvement on GOR

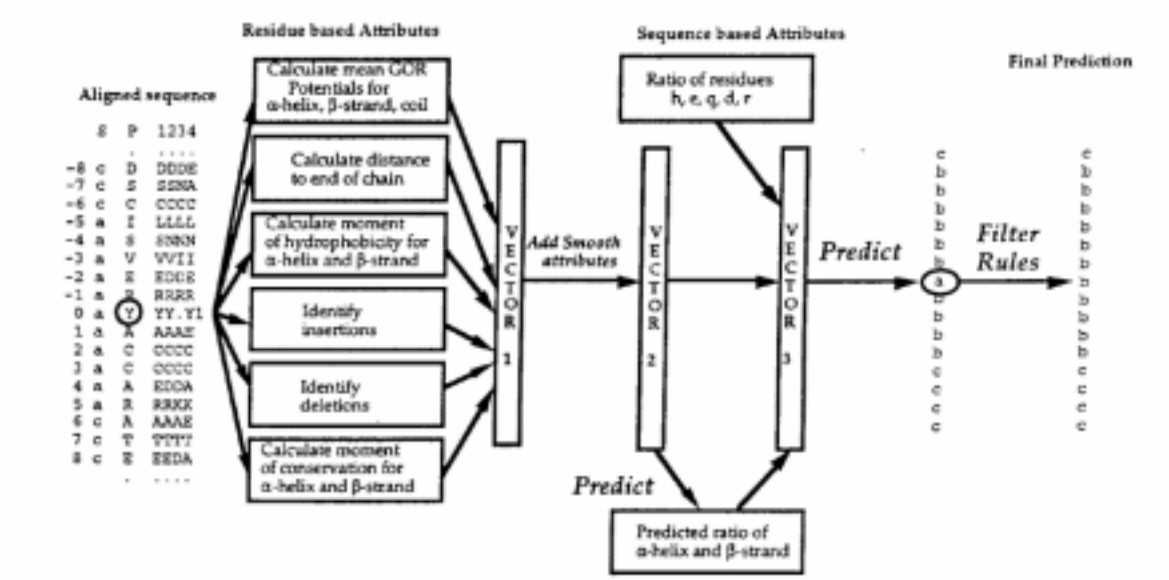


Fig. 1. DSC prediction method. For the aligned sequence: S is the observed secondary structure of the primary sequence, P. The residue at position 0 is predicted (circled).

- GOR parms
- + simple linear discriminant analysis on:
  - ◊ dist from C-term, N-term
  - ◊ insertions/deletes
  - ◊ overall composition
  - ◊ hydrophobic moments
  - ◊ autocorrelate: helices
  - ◊ conservation moment

Illustration Credits: King & Sternberg, 1996

Predator

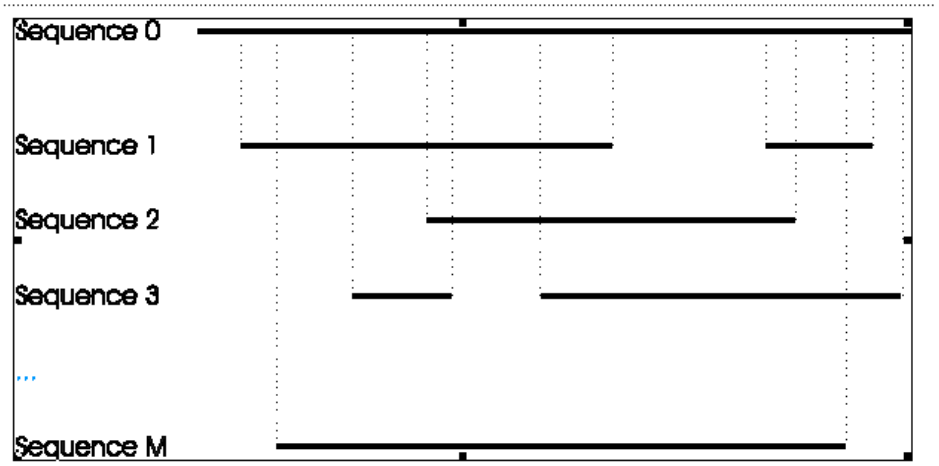


Figure 5. Pairwise local alignments of the query sequence 0 with the related sequences m=1,2...M. Every alignment is characterized by its length and residue percentage identity.

- Predator
  - ◊ **Frishman & Argos (1997)** proposed an alternate way to utilize the additional information contained in a set of related sequences. A careful pairwise alignment of the query sequence with all related sequences is performed.
- NNSSP, Salamov & Solovyev, 1995
  - ◊ small Segments, Dist metric betw. them (use eisenberg env.), k-closest neighbors, >71% accuracy
- Yi & Lander, 1994; Presnell et al., 1992

Illustration Credits: D Frishman handout

Neural Networks

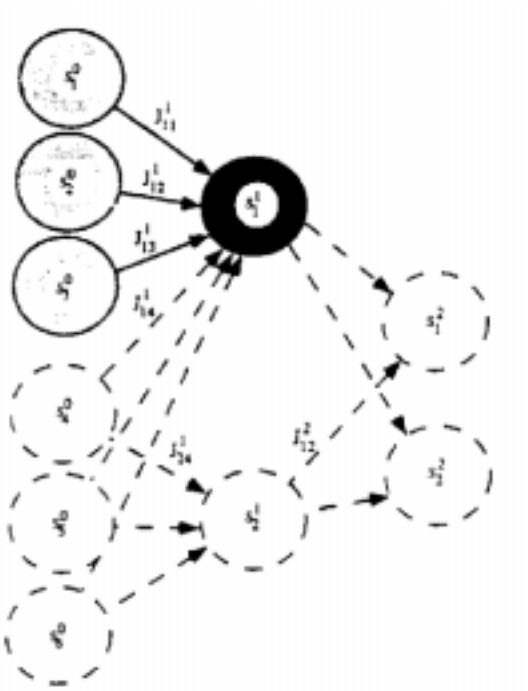
Figure 1. Function of a perceptron, the simplest neural network. A simple perceptron has only 1 output unit (black). Each of the left nodes receives a certain input signal (e.g. binary, i.e. =0 or 1). All units are connected to the output node by the junctions J^1, with e.g. J^1_{1j} connecting input unit j with output unit 1. The contribution of each left node (e.g. the jth) to the signal arriving at the right one is a product of the strength of the junction connecting the 2 units, and the input: e.g. $J^1_{1j}s^0_j$. All products (here 3) are summed by the right node (here s^1_1). This sum is then evaluated by a non-linear trigger function. The resulting map of the sum onto an interval between 0 and 1 is the actual output of the network. The broken-line nodes show a potential extension of the perceptron to a 2-layered feed-forward network. Stippled circles, input units, signal = 1 or 0. Black circle, output unit. Step 1, the input to this unit is summed according to:

$$h^1_i = \sum_{j=1}^{N^0-1} J_{ij}s^0_j \quad \text{(here, } i=1\text{)}.$$

Step 2, the output from this unit is computed by a sigmoid trigger function:

$$s^1_i = \frac{1}{1+\exp(-h^1_i)}.$$

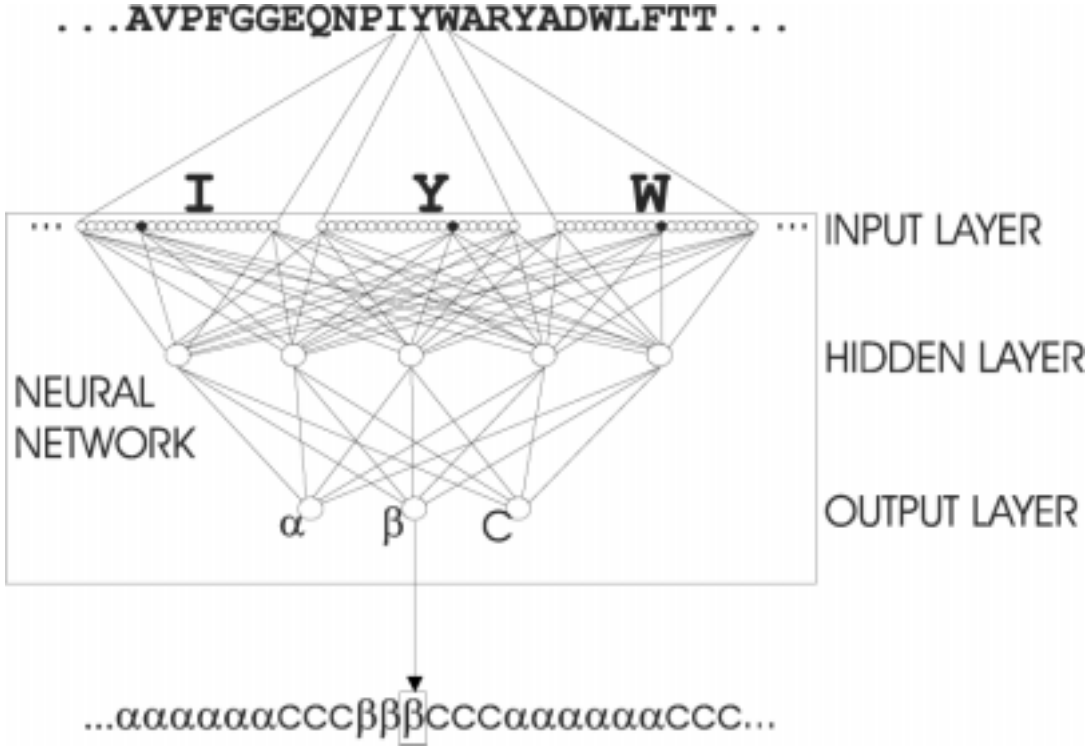
Broken-line circles, the potential extension to a 2-layered feed-forward network.



- Somehow generalize and learn patterns
- Black Box
- Rost, Kneller, Qian….
- Perceptron (above) is Simplest network
  - ◊ Multiply junction * input, sum, and threshold

Illustration Credits: Rost & Sander, 1993

More NN



- Hidden Layer
- Learning
  - ◊ Steepest descent to minimize an error function
- Jury Decision
  - ◊ Combine methods
  - ◊ Escape initial conditions

Illustration Credits: D Frishman handout

Yet more methods….

- struc class predict
  - ◊ Vect dist. between composition vectors
- threading via pair pot
- seq comparison
- ab initio from md
- ab initio from pair pot.

Mail Servers and Web Forms

Method	URL	Institution	Source code Availability
ANTHE-PROT	http://www.ibcp.fr/antheprot.html (currently unreachable)	Institute of Biology and Chemistry of Proteins (Lion)	YES
PSSP	http://dot.imgen.bcm.tmc.edu:9331/psspredict ion/pssp.html	Baylor College of Medicine (Houston)	NO
DSC	http://bonsai.lif.icnet.uk/bmm/dsc/dsc_form_align.html	Imperial Cancer Research Center (London)	YES
GOR	http://molbiol.soton.ac.uk/compute/GOR.html	University of Southampton	NO
nnPredict	http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html	University of California (San Francisco)	NO
Predict-Protein	http://www.embl-heidelberg.de/predictprotein/predictprotein.html	EMBL (Heidelberg)	NO
PRED-ATOR	http://www.embl-heidelberg.de/argos/predator/predator_form.html	EMBL (Heidelberg)	YES
PSA	http://bmerc-www.bu.edu/psa/	BioMolecular Engineering Research Center, Boston	NO
SSPRED	http://www.embl-heidelberg.de/sspred/sspred_info.html	EMBL (Heidelberg)	NO
GOR and DSC	http://genome.imb-jena.de/cgi-bin/GDEWWW/menu.cgi	IMB (Jena)	NO
GOR	http://absalpha.dcrt.nih.gov:8008/gor.html	DCRT/NIH (Washington)	NO
GOR	ftp://ftp.virginia.edu/pub/fasta	University of Virginia	YES
Mult-Predict	http://kestrel.ludwig.ucl.ac.uk/zpred.html	Ludwig Institute for Cancer Research (London)	NO

Illustration Credits: D Frishman handout

References

Argos P. (1976) Prediction of the secondary structure of mouse nerve growth factor and its comparison with insulin. *Biochemical and Biophysical Research Communications* 3:805-811.

Bairoch A and Apweiler R. (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res* 24:21-25.

Barton GJ. (1995) Protein secondary structure prediction. *Curr Opinion Struct Biol* 5:372-376.

Benner SA, Gerloff DL, and Jenny TF. (1994) Predicting protein crystal structures. *Science* 265:1642-1644.

Benner SA. (1995) Predicting the conformation of proteins from sequences. Progress and future progress. *J Mol Recogn* 8:9-28.

Boyd, D., Schierle, C. & Beckwith, J. (1998). How many membrane proteins are there? *Prot. Sci.* **7**, 201-205.

Crawford IP, Niermann T, and Kirschner K. (1987) Prediction of secondary structure by evolutionary comparison: application to the alpha subunit of thryptophan synthase. *Proteins: Struct Func Genet* 2:118-129.

Deleage G and Roux B. (1987) An algorithm for protein secondary structure prediction based on class prediction. *Protein Engineering* 4:289-294.

Eigenbrot C, Randal M, and Kossiakoff AA. (1992) Structural Effects Induced by Mutagenesis Affected by Crystal Packing Factors: the Structure of a 30-51 Disulfide Mutant of Basic Pancreatic Trypsin Inhibitor. *Proteins* 14:75.

Fasman, G. D. & Gilbert, W. A. (1990). The prediction of transmembrane protein sequences and their conformation: an evaluation. *Trends Biochem Sci* **15**, 89-92.

Frishman D, and Argos P. (1995) Knowledge-Base Protein Secondary Structure Assignment. *Proteins: Structure, Function, and Genetics* 23:566-79.

Frishman D, and Argos P. (1996) Incorporation of Non-Local Interactions in Protein Secondary Structure Prediction From the Amino Acid Sequence. *Protein Engineering* 2:in the press.

Frishman D, and Argos P. (1997) The Future of Protein Secondary Structure Prediction Accuracy. *Folding & Design* 2:159-62.

Frishman, D, and P Argos. (1996) 75% Accuracy in Protein Secondary Structure Prediction. Proteins 1997 Mar;27(3):329-335.

Garnier J and Levin JM. (1991) The protein structure code: what is its present status. *Comput Appl Biosci* 7:133-142.

Garnier, J. (1990). Protein structure prediction. *Biochimie* **72**, 513-24.

Garnier, J., Gibrat, J. F. & Robson, B. (1996a). GOR method for predicting protein secondary structure from amino acid sequence. *Meth. Enz.* **266**, 540-553.

References

Garnier, J., Gibrat, J. F. & Robson, B. (1996b). GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol* **266**, 540-53.

Garnier, J., Osguthorpe, D. & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97-120.

Geourjon C, and Deléage G. (1995) SOPMA: Significant Improvements in Protein Secondary Structure Prediction by Consensus Prediction From Multiple Sequences. *Comput Appl Biosci* 11:681-84.

Gibrat, J., Garnier, J. & Robson, B. (1987). Further developments of protein secondary structure prediction using information theory. *J. Mol. Biol.* **198**, 425-443.

Gilbert RJ. (1992) Protein structure prediction from predicted residue prperties utilizing a digital encoding algorithm. *J Mol Graph* 10:112-119.

Holley LH, and Karplus M. (1989) Protein Secondary Structure Prediction With a Neural Network. *Proc Natl Acad Sci USA* 86:152-56.

Hunt NG, Gregoret LM, and Cohen FE. (1994) The origins of protein secondary structure. Effects of packing density and hydrogen bonding studied by a fast conformational search. *J Mol Biol* 241:214-225.

Kabsch W and Sander C. (1984) On the use of sequence homologies to predict protein structure. Identical pentapeptides can have completely different conformation. *Proc Natl Acad Sci USA* 81:1075-1078.

Kabsch W, and Sander C. (1983) Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* 22:2577-637.

Kendrew JC, Klyne W, Lifson S, Miyazawa T, Nemethy G, Phillips DC, Ramachandran GN, and Sheraga HA. (1970). *Biochemistry* 9:3471-79.

King, R. D. & Sternberg, M. J. E. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Prot. Sci.* **5**, 2298-2310.

King, R. D., Saqi, M., Sayle, R. & Sternberg, M. J. (1997). DSC: public domain protein secondary structure predication. *Comput Appl Biosci* **13**, 473-4.

Levin J, Pascarella S, Argos P, and Garnier J. (1993) Quantification of secondary structure prediction improvement using multiple alignments. *Protein Engineering* 6:849-854.

Levin JM, Robson B, and Garnier J. (1986) An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett* 205:303-308.

Levitt M, and Greer J. (1977) Automatic Identification of Secondary Structure in Globular Proteins. *J Mol Biol* 114:181-239.

Lim VI. (1974) Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins. *J Mol Biol* 88:873-894.

Livingstone CD, Barton GJ (1996). Identification of functional residues and secondary structure from protein multiple sequence alignment. Methods Enzymol 266:497-512

Lupas A, Koster AJ, Walz J, and Baumeister W. (1994) Predicted secondary structure of the 20 S proteasome and model structure of the putative peptide channel. *FEBS Lett* 354:45-49.

Matthews B. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochem Biophys Acta* 405:442-451.

Mehta PK, Heringa J, and Argos P. (1995) A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Science* 4:2517-2525.

References

Muggleton S, King RD, and Sternberg MJE. (1992) Protein Secondary Structure Prediction Using Logic-Based Machine Learning. *Protein Engineering* 5:647-57.

Nishikawa K, and Ooi T. (1986) Amino Acid Sequence Homology Applied to the Prediction of Protein Secondary Structures, and Joint Prediction With Existing Methods. *Biochimica Et Biophysica Acta* 871:45-54.

Pauling L, Corey RB, and Branson HR. (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci USA* 37:205-211.

Persson, B. & Argos, P. (1997). Prediction of membrane protein topology utilizing multiple sequence alignments. *J Protein Chem* **16**, 453-7.

Presnell SR, Cohen BI, and Cohen FE. (1992) A segment-based approach to protein secondary structure prediction. *Biochemistry* 31:983-993.

Ptitsyn OB and Finkelstein AV. (1983) Theory of protein secondary structure and algorithm of its prediction. *Biopolymers* 22:15-25.

Qian N, and Sejnowski TJ. (1988) Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. *J Mol Biol* 202:865-84.

Rackovsky S. (1993) On the nature of the protein folding code. *Proc Natl Acad Sci U S A* 90:644-648.

Ramakrishnan C, and Soman KV. (1982) Identification of Secondary Structures in Globular Proteins - a New Algorithm. *Int J Pept Protein Res* 20:218-37.

Rao S, Zhu Q-L, Vaida S, and Smith T. (1993) The local information content of the protein structural database. *FEBS Lett* 2:143-146.

Rice CM, Fuchs R, Higgins DG, Stoehr PJ, and Cameron G N. (1993) The EMBL data library. *Nucleic Acids Res* 21:2967-2971.

Richards FM, and Kundrot CE. (1988) Identification of Structural Motifs From Protein Coordinate Data: Secondary Structure and First-Level Supersecondary Structure. *Proteins: Struct Func Genet* 3:71-84.

Robson B, and Garnier J. (1993) Protein Structure Prediction. *Nature* 361:506.

Rost B, and Sander C. (1993) Prediction of Protein Secondary Structure at Better Than 70% Accuracy. *J Mol Biol* 232:584-99.

Rost B, Sander C, and Schneider R. (1994) Redefining the goals of protein secondary structure prediction. *J Mol Biol* 235:13-26.

Rost B, Sander C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci USA* 90:7558-7562

Rost, B., Schneider, R. & Sander, C. (1993). Progress in protein structure prediction? *Trends Biochem Sci* **18**, 120-3.

Rumelhart DE, Hinton GE, and Williams R. (1986) Learning representations by back-propagating errors. *Nature* 323:533-536.

Salamov AA, and Solovyev VV. (1995) Prediction of Protein Secondary Structure by Combining Nearest-Meighbour Akgorithms Amd Multiple Sequence Alignments. *J Mol Biol* 247:11-15.

Salamov AA, and Solovyev VV. (1997) Protein Secondary Structure Prediction Using Local Alignments. *Journal of Molecular Biology* 268:31-36.

Sayle RA, and Milner-White EJ. (1995) RASMOL: Biomolecular Graphics for All. *Trends in Biochemical Sciences* 20:374-76.

References

Sayle RA, and Milner-White EJ. (1995) RASMOL: Biomolecular Graphics for All. *Trends in Biochemical Sciences* 20:374-76.

Sklenar H, Etchebest C, and Lavery R. (1989) Describing Protein Structure: a General Algorithm Yielding Complete Helicoidal Parameters and a Unique Overall Axis. *Proteins: Struct Func Genet* 6:46-60.

Solovyev VV and Salamov AA. (1994) Predicting alpha-helix and beta-strand segments of globular proteins. *Comput Appl Biosci* 10:661-669.

Stolorz P, Lapedes A, and Xia Y. (1992) Predicting Protein Secondary Structure Using Neural Net and Statistical Methods. *J Mol Biol* 225:363-77.

Sumpter BG, Getino C, and Noid DW. (1994) Theory and applications of neural computing in chemical science. *Ann Rev phys Chem* 45:439-481.

Sumpter BG, Getino C, and Noid DW. (19949 Theory and applications of neural computing in chemical science. *Ann Rev phys Chem* 45:439-481.

Taylor WR and Thornton JM. (1984) Recognition of super-secondary structure in proteins. *J Mol Biol* 173:487-5141984.

Thompson JD, Higgins DG, and Gibson TJ. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673-4680.

Thornton JM, Flores TP, Jones DT, and Swindells MB. (1991) Prediction of progress at last. *Nature* 354:105-106.

von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol* **225**, 487-94.

Wasserman PD. (1989) *Neural Computing. Theory and Practice.* New York.

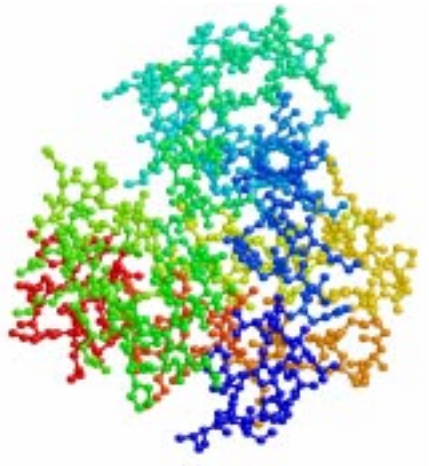
Zhang X, Mesirov JP, and Waltz DL. (1992) Hybrid System for Protein Secondary Structure Prediction. *J Mol Biol* 225:1049-63.

Zvelebil MJ, Barton GJ, Taylor WR, and Sternberg MJ. (1987) Prediction of Protein Secondary Structure and Active Sites Using the Alignment of Homologous Sequences. *J Mol Biol* 195:957-61.

Subject II:

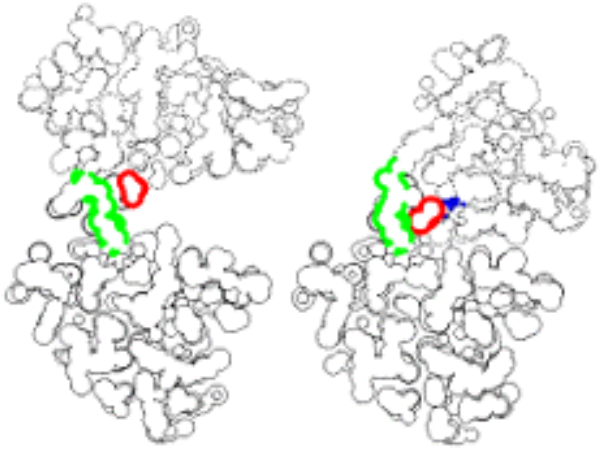
Packing

Other Aspects of Structure, Besides just Comparing Atom Positions



Atom Position, XYZ triplets

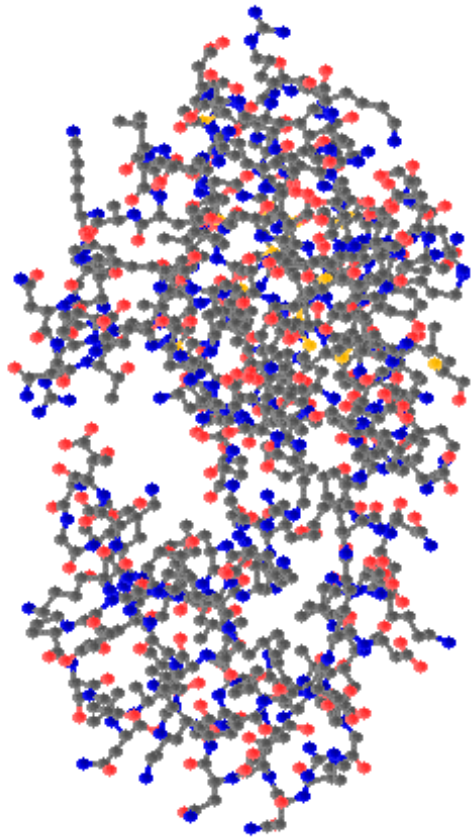
Lines, Axes, Angles



Surfaces, Volumes

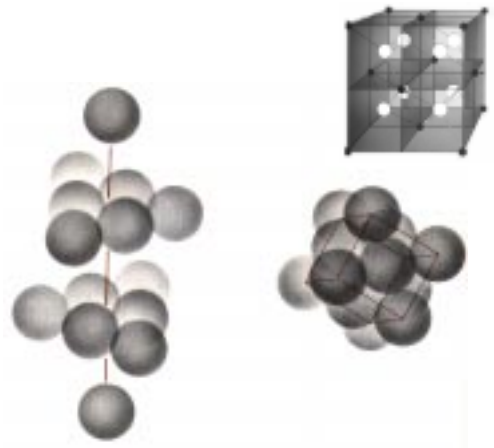
What is Protein Geometry?

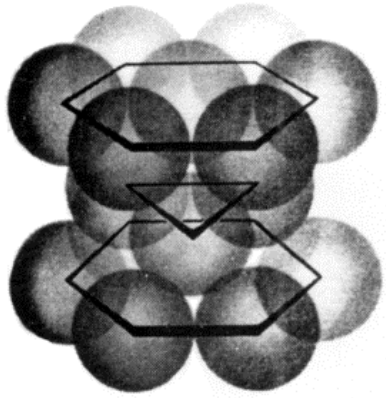
- Coordinates (X, Y, Z's)
- Derivative Concepts
  - ◊ Distance, Surface Area, Volume, Cavity, Groove, Axes, Angle, &c
- Relation to
  - ◊ Function, Energies (E(x)), Dynamics (dx/dt)

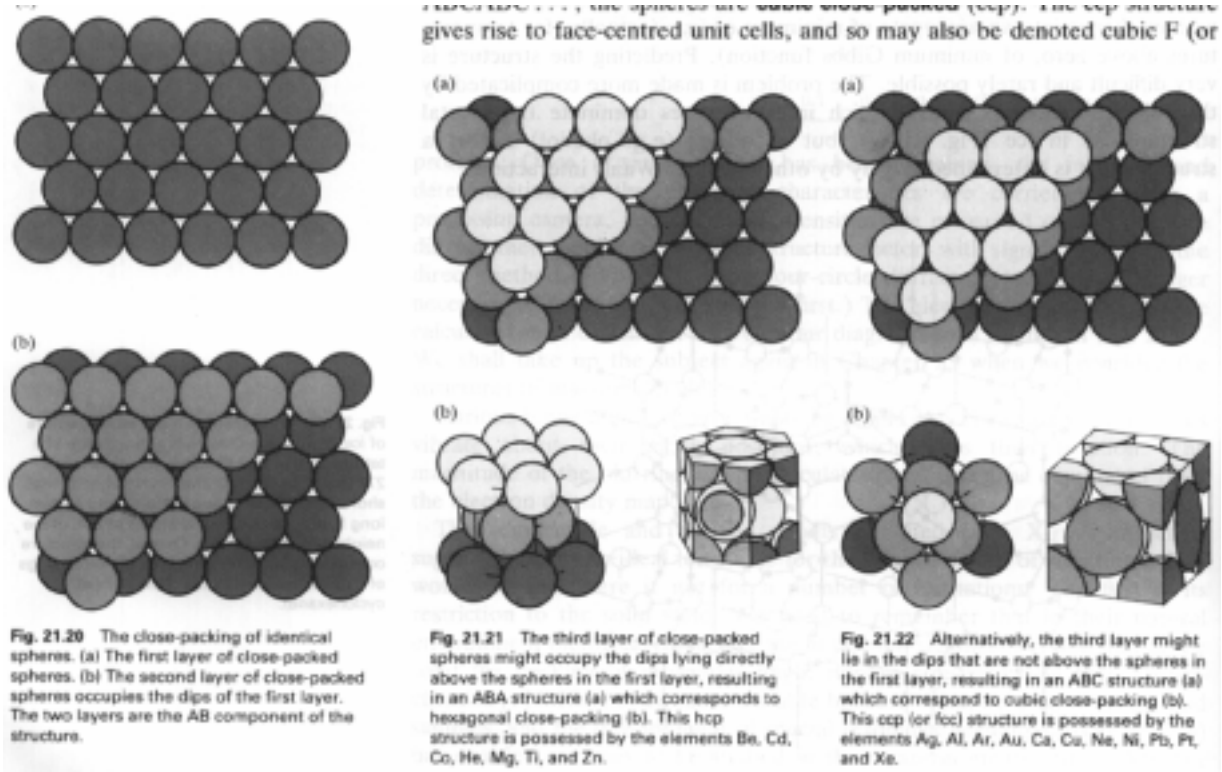


Close-Packing of Spheres

- 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







gives rise to face-centred unit cells, and so may also be denoted cubic F (or

(a)

(b)

(a)

(b)

(b)

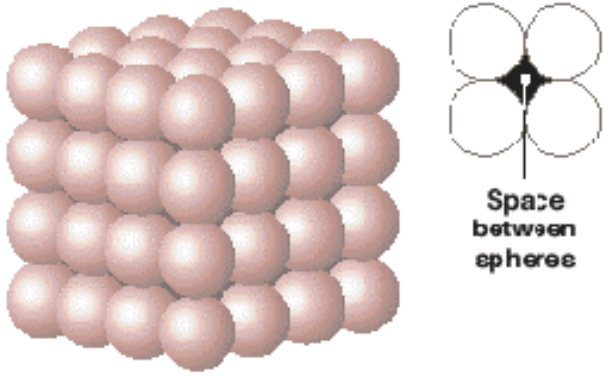
Fig. 21.20 The close-packing of identical spheres. (a) The first layer of close-packed spheres. (b) The second layer of close-packed spheres occupies the dips of the first layer. The two layers are the AB component of the structure.

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.

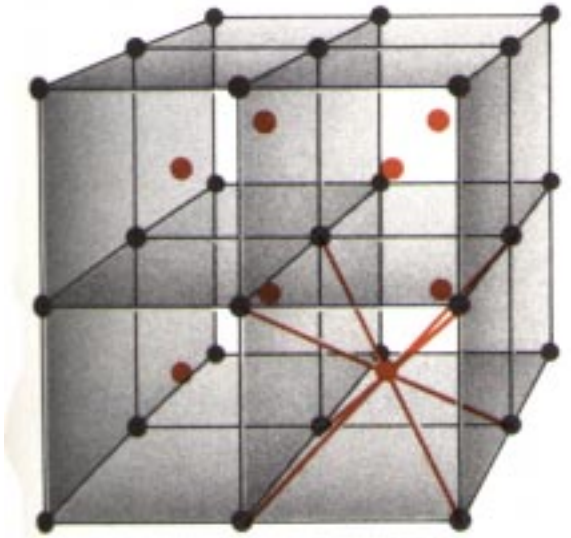
Fig. 21.22 Alternatively, the third layer might lie in the dips that are not above the spheres in the first layer, resulting in an ABC structure (a) which correspond to cubic close-packing (b). This ccp (or fcc) structure is possessed by the elements Ag, Al, Ar, Au, Ca, Cu, Ne, Ni, Pb, Pt, and Xe.

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

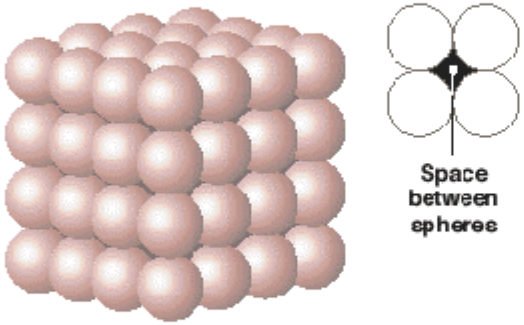


Optimal Packing Finally Proved

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.

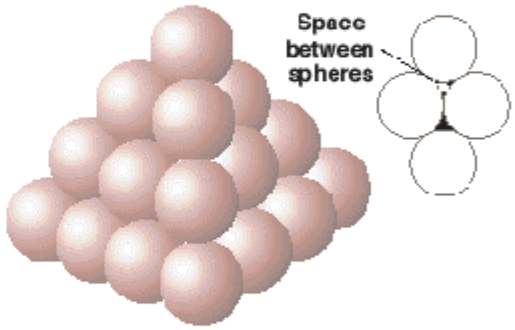
SIMPLE CUBIC LATTICE



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



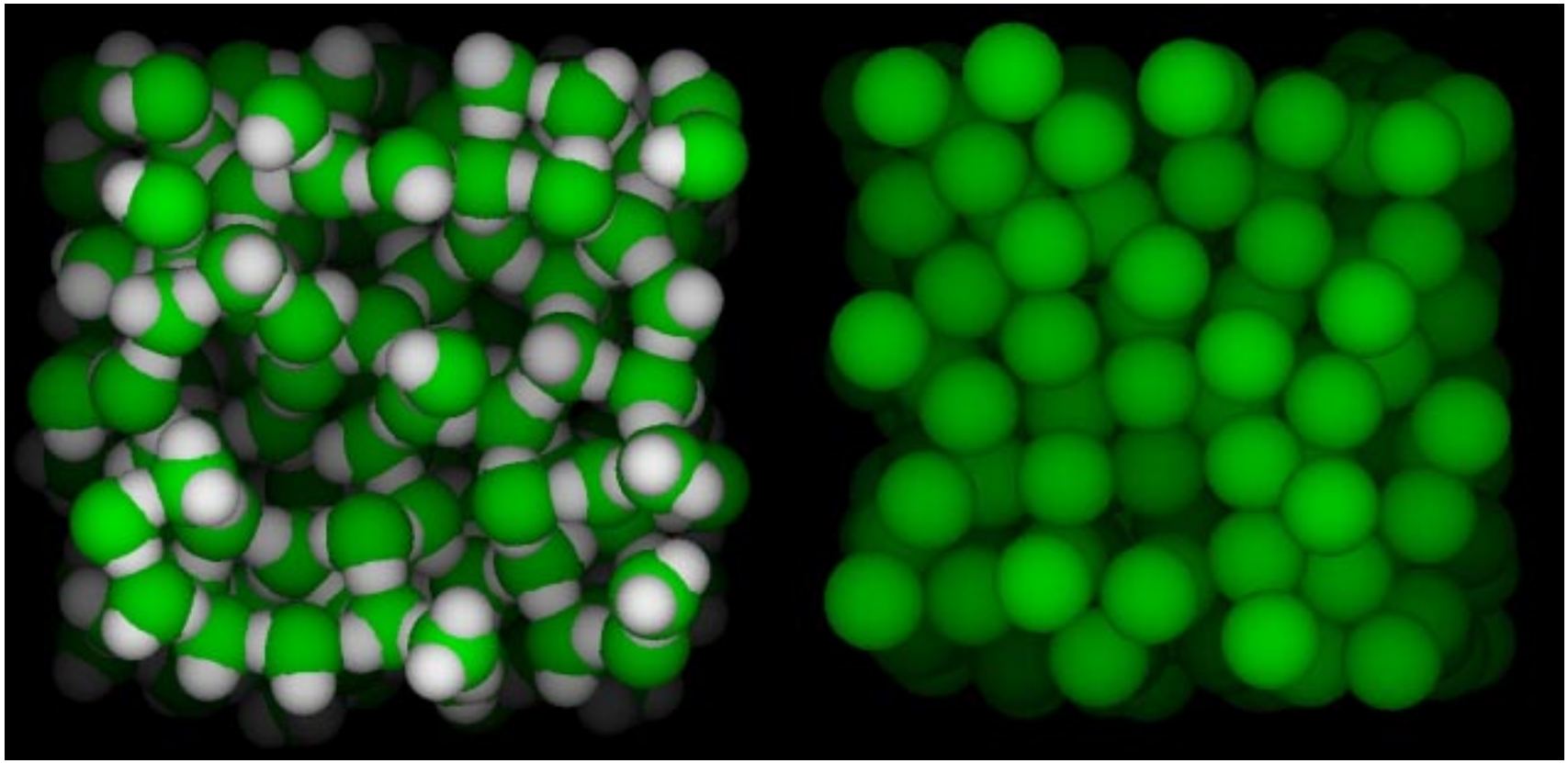
STACKING EFFICIENCY **74%**

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

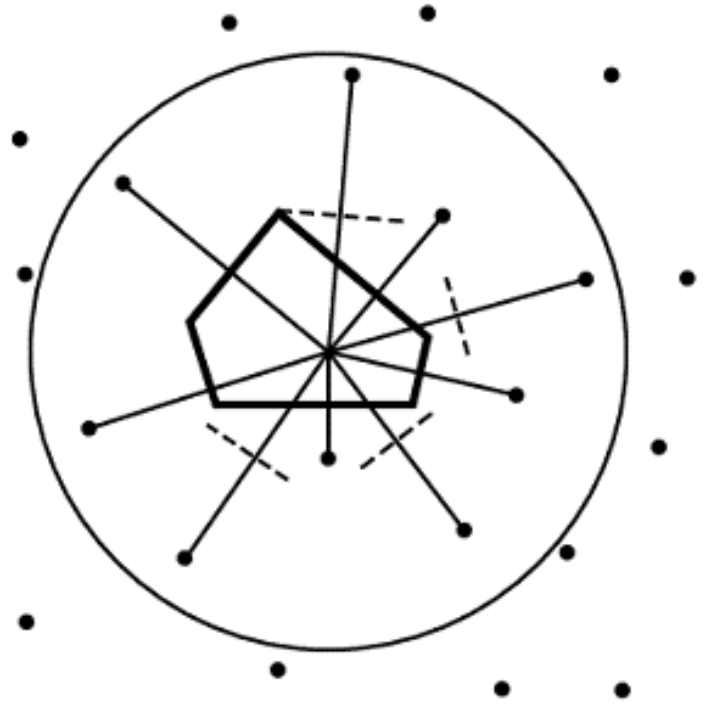
Water v. Argon



More Complex Systems -- what to do?

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.

Voronoi Volumes, the Natural Way to Measure Packing

Packing Efficiency

= Volume-of-Object

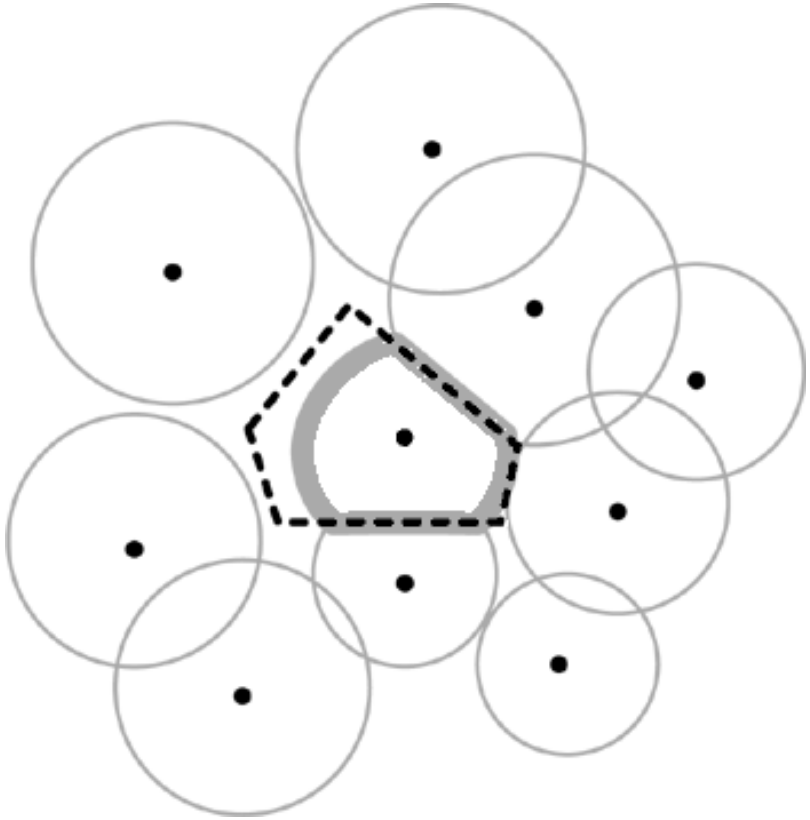
Space-it-occupies

= V(VDW) / V(Voronoi)

- Absolute v relative eff.

  V1 / V2

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

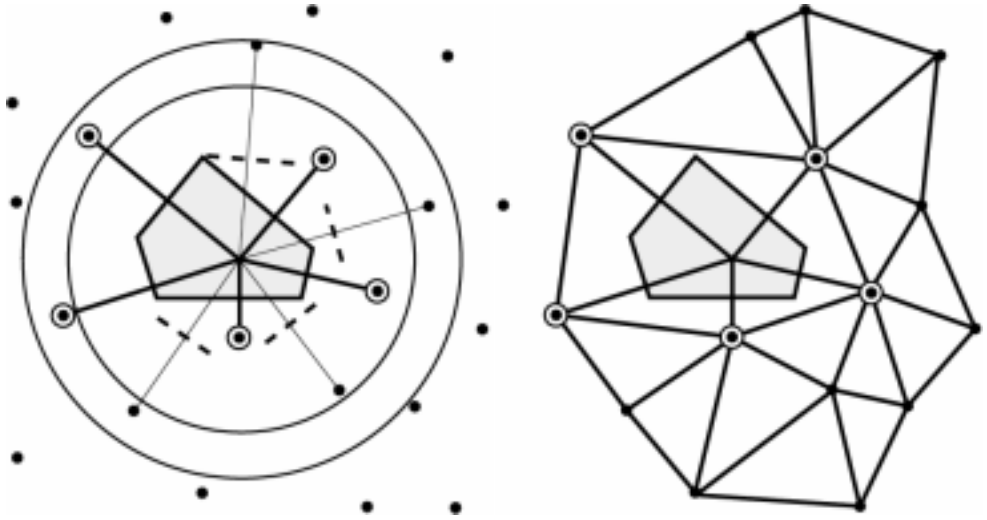


Calclating Volumes with Voronoi polyhedra

- 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.

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

Surfaces
(slides 1-10, 20-40 from website)

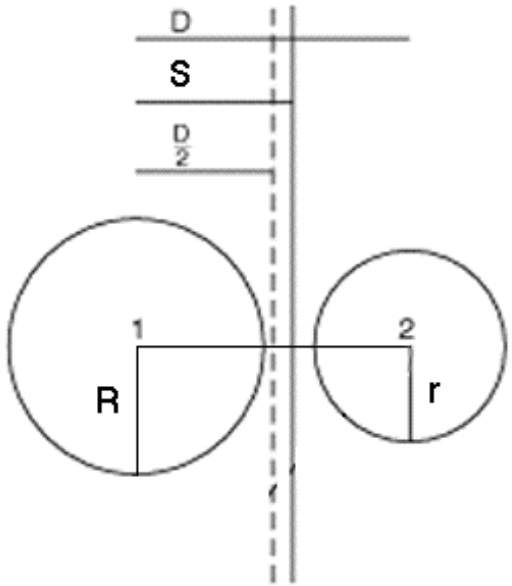
These are detailed slides on how to do Voronoi construction.

Go to http://bioinfo.mbb.yale.edu/geometry

and follow links to “HyperTalk” tutorial on surfaces 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



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

Different Sets of Radii

Atom Type & Symbol		Bondi 1968	Lee & Richards 1971	Shrake & Rupley 1973	Richards 1974	Chothia 1975	Rich-mond & 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

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...

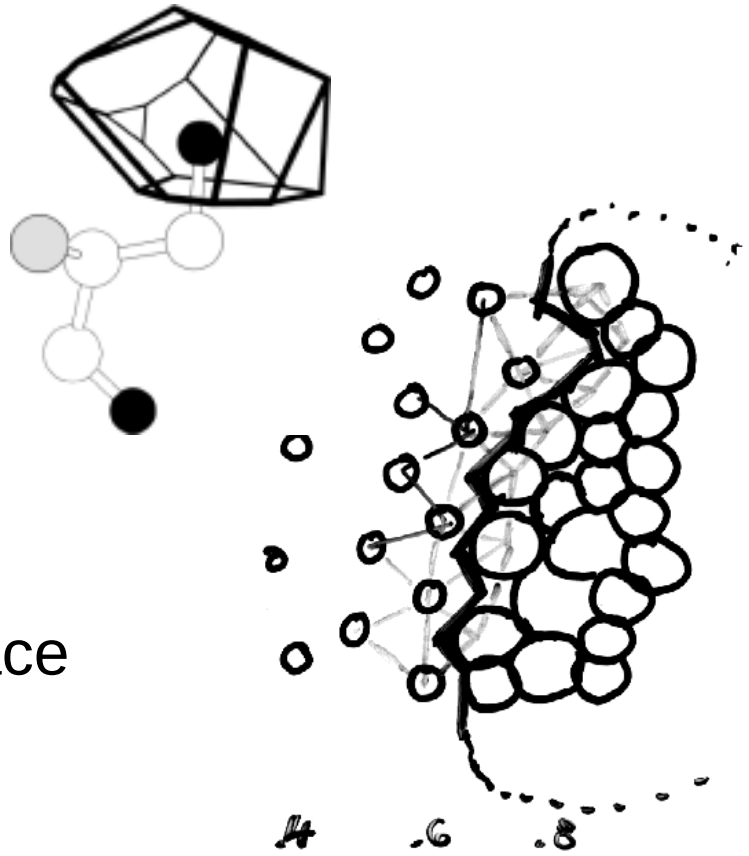
G 64	c 105	T 120	V 139	H 159	M 168	R 194
A 90	C 113	P 124	E 140	L 165	K 170	Y 198
S 94	D 117	N 128	N 150	I 165	F 193	W 233

Standard Core Volumes (Prelim.)

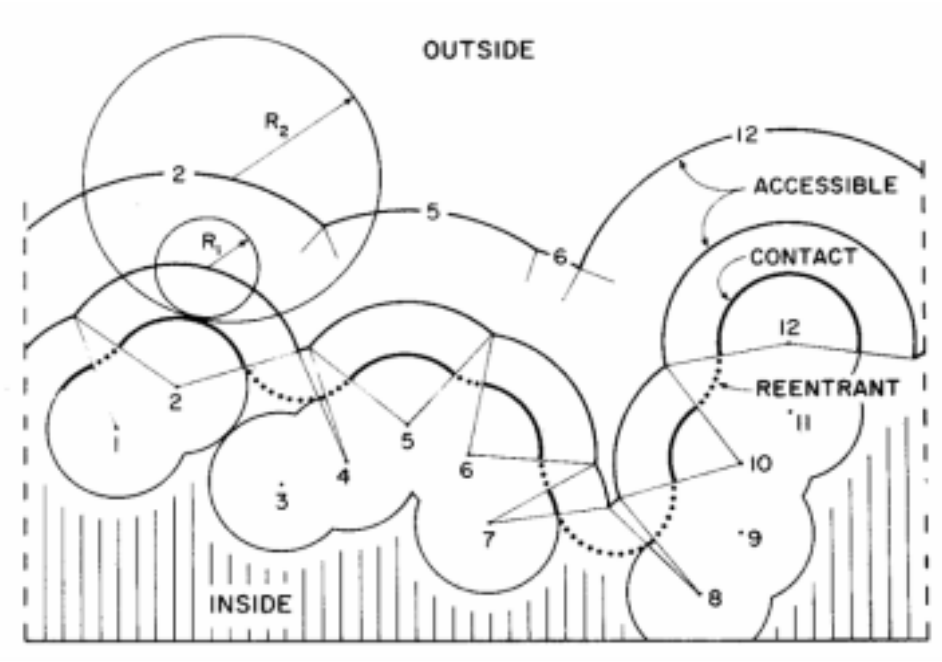
Atom Types		Num.	Volume (Å^3)	Error (%)
Mainchain Atoms				
carbonyl carbon (except G)	C	8361	9.2	.08
alpha carbon (except G)	CA	7686	13.4	.09
nitrogen (except P)	N	9042	13.9	.09
carbonyl oxygen	O	7831	15.8	.10
Gly C		811	10.2	.27
Gly CA		522	23.5	.39
Pro N		334	8.6	.39
Sidechain atoms				
trigonal or 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

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



Richards' Molecular and Accessible Surfaces

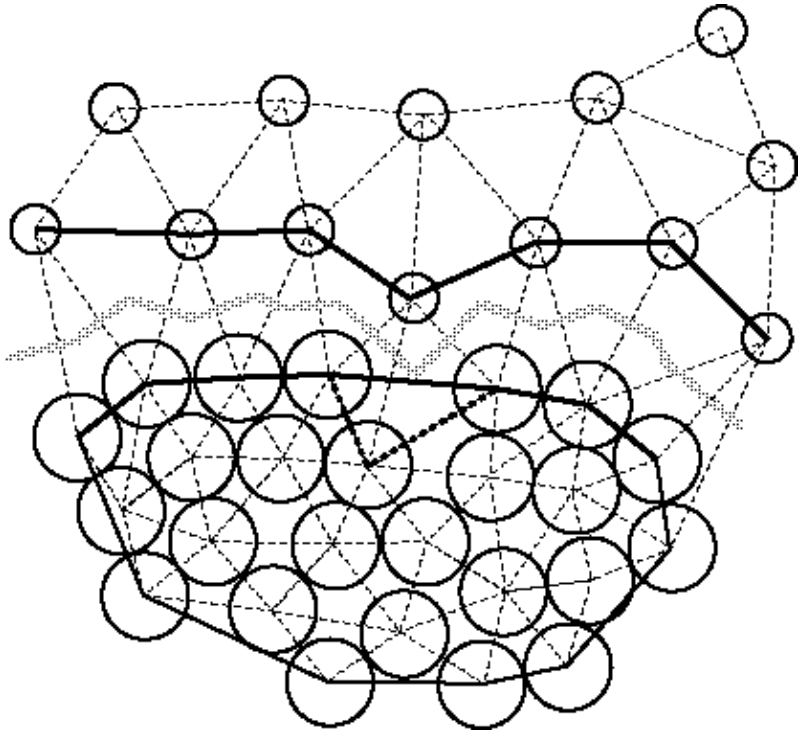


Probe Radius	Part of Probe Sphere	Type of Surface
0	Center (or Tangent)	Van der Waals Surface (vdWS)
1.4 Å	Center	Solvent Accessible Surface (SAS)
""	Tangent (1 atom)	Contact Surface (CS, from parts of atoms)
""	Tangent (2 or 3 atoms)	Reentrant Surface (RS, from parts of Probe)
""	Tangent (1,2, or 3 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

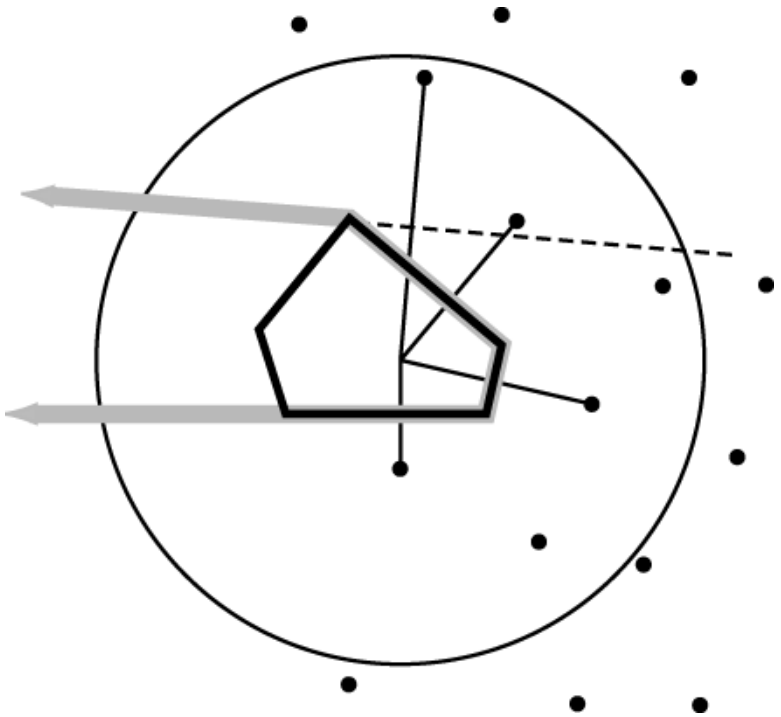
Packing defines the "Correct Definition" of the Protein Surface

- 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

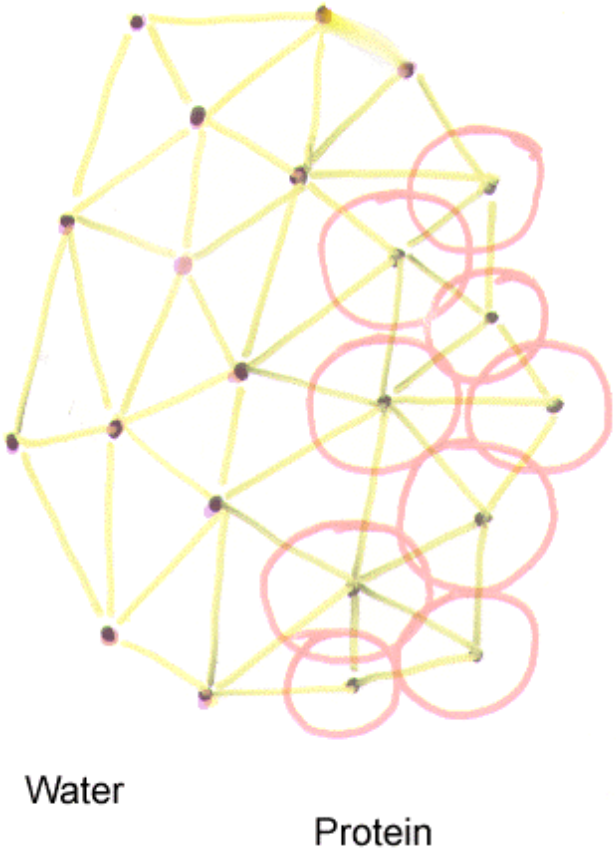
Surface and Volume Definitions Linked

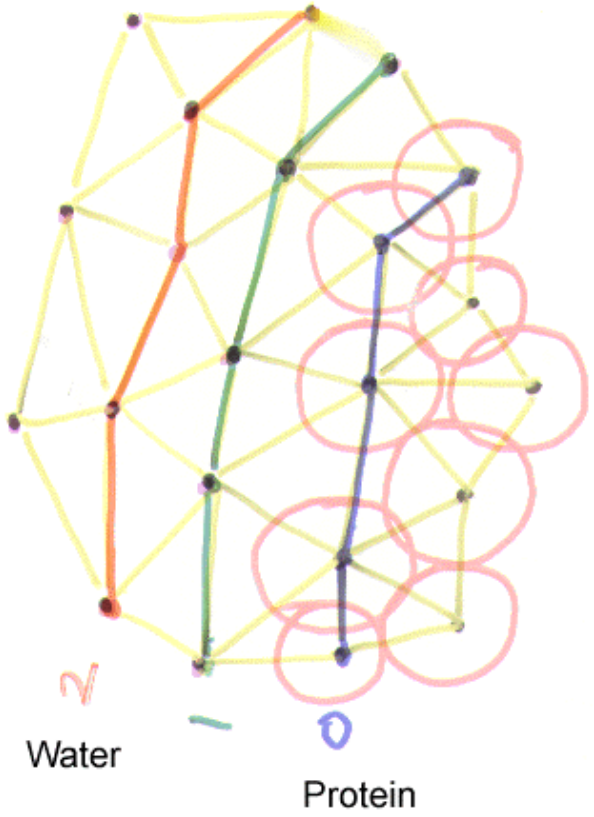


Problem of Protein Surface for Voronoi Construction

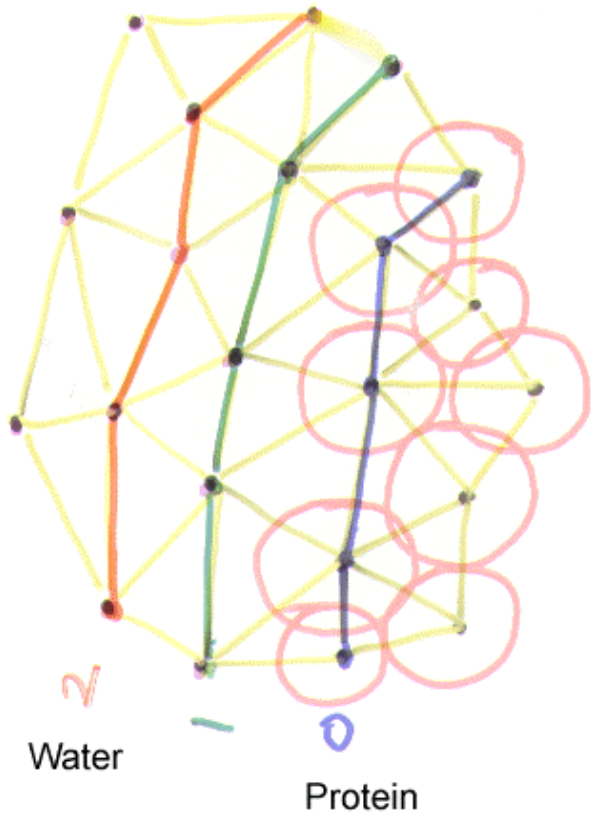


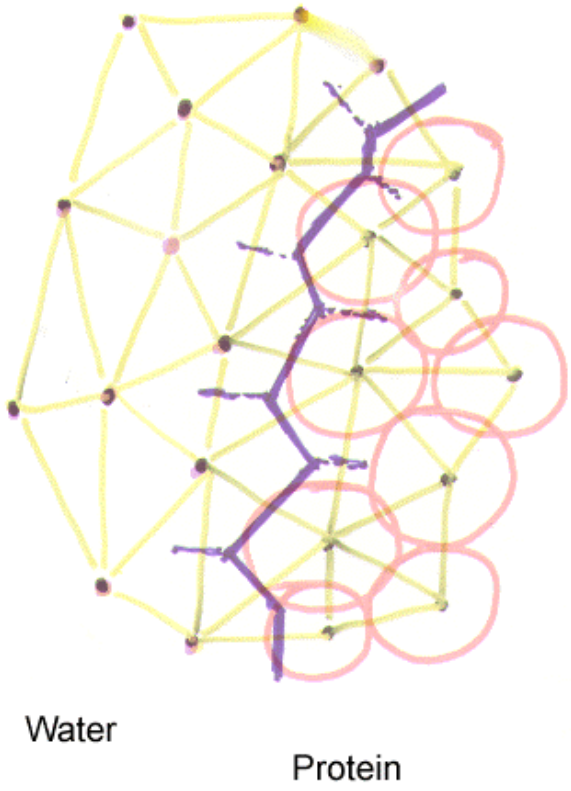
Defining Surfaces from Packing: Convex Hull and Layers of Waters



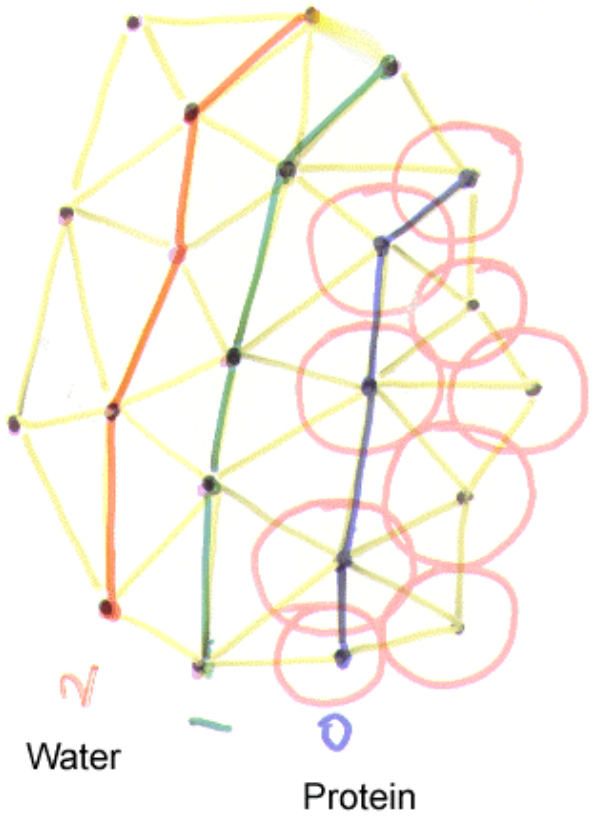


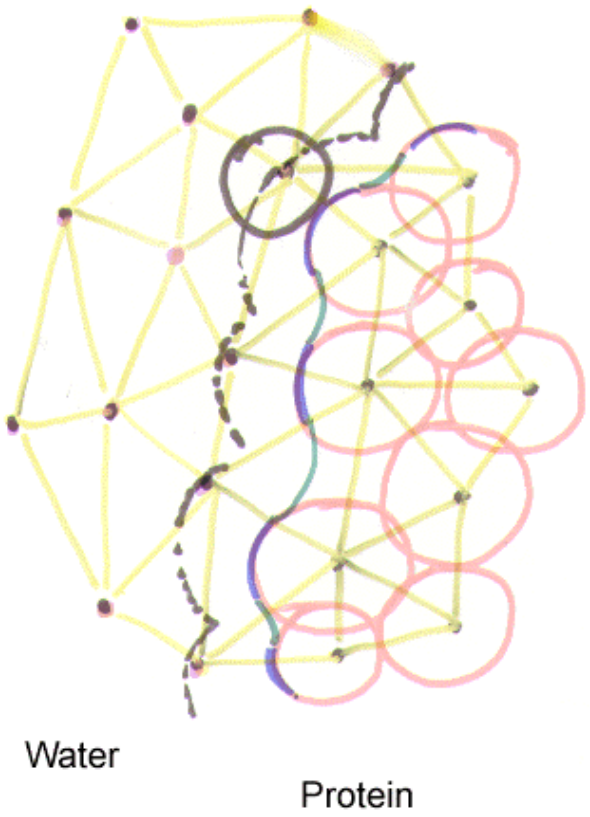
Defining a Surface from the Faces of Voronoi Polyhedra



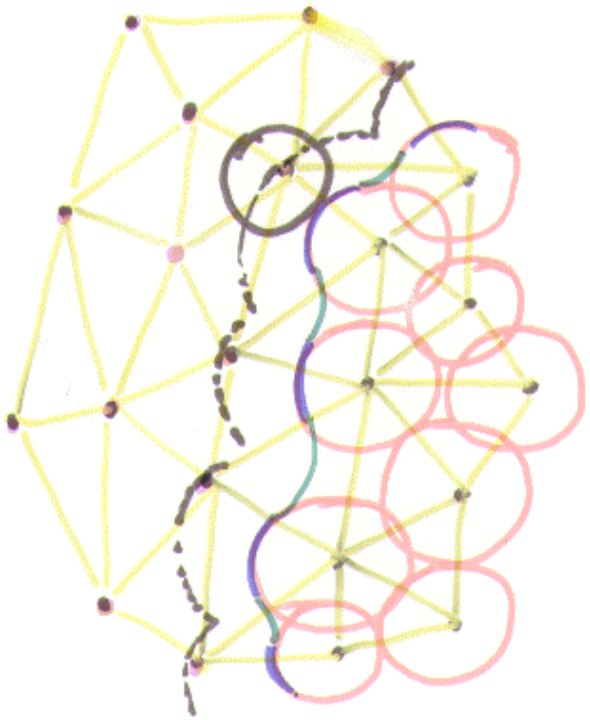


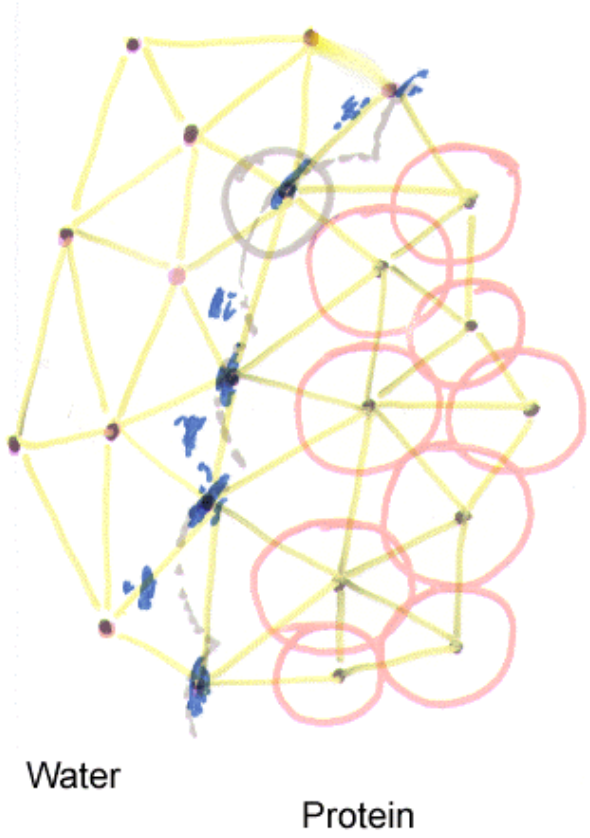
Accessible Surface as a Time-averaged Water Layer





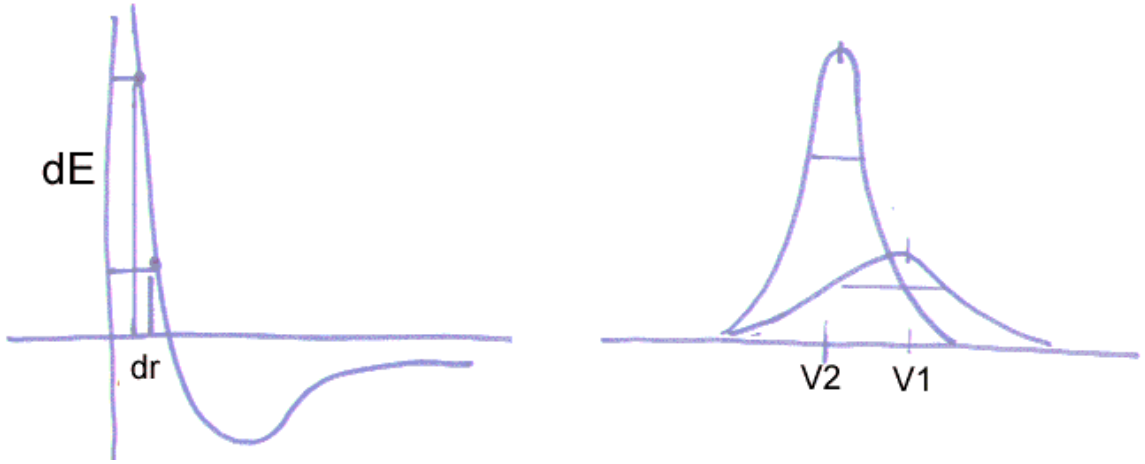
The Hydration Surface: Trying to Model Real Water





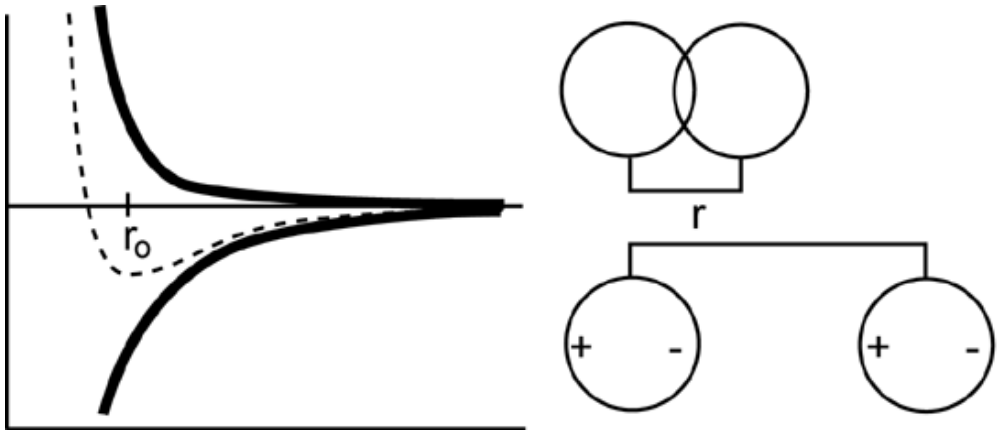
Small Packing Changes Significant

- 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



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