Informative Length Scale in Protein Folding

Nima Hamedani Radja
IPM, Ordibehesht 85

Collaborators:
M. R. Ejtehadi
R. R. Farzami

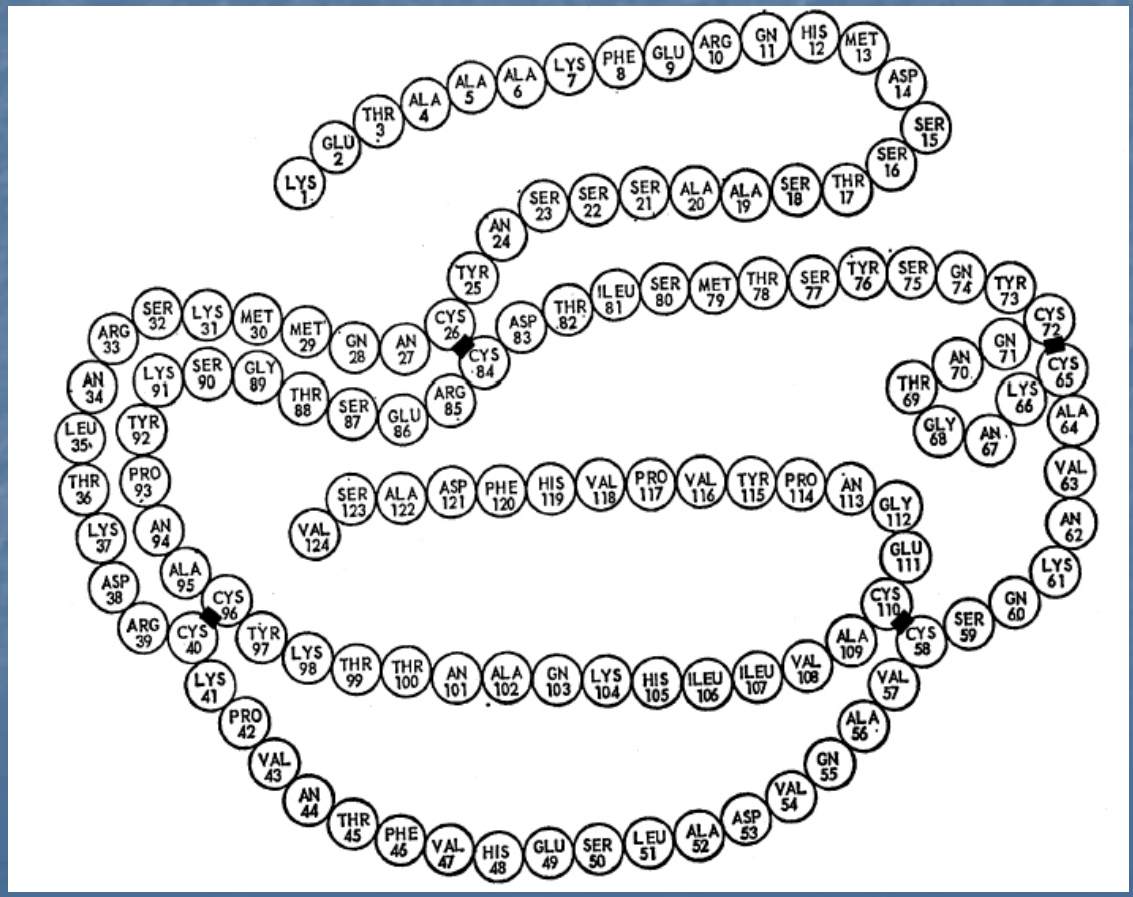
Outline

- Introduction
- Model
- Results

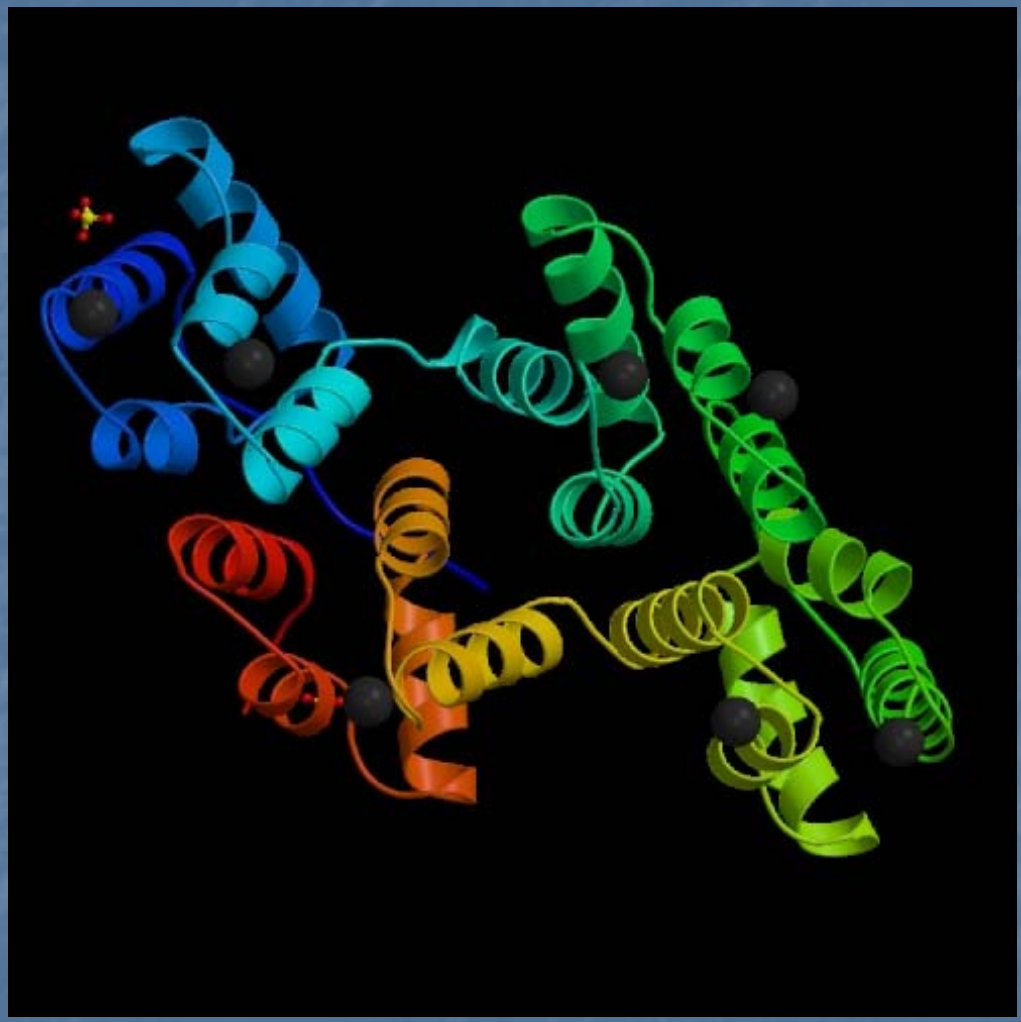
- Protein
- Amino Acid
- Nondegenerate contact matrices
- Side chains
- Cutoff
- Mean field approximation
- Contact
- Diamond lattice
- Analysis of PDB-like self-avoiding
- Coarse-graining
- PDB analysis
- random walks
- First to find structure coding
- PDB-like structures
- 6.5 angstrom

Protein → Polymer

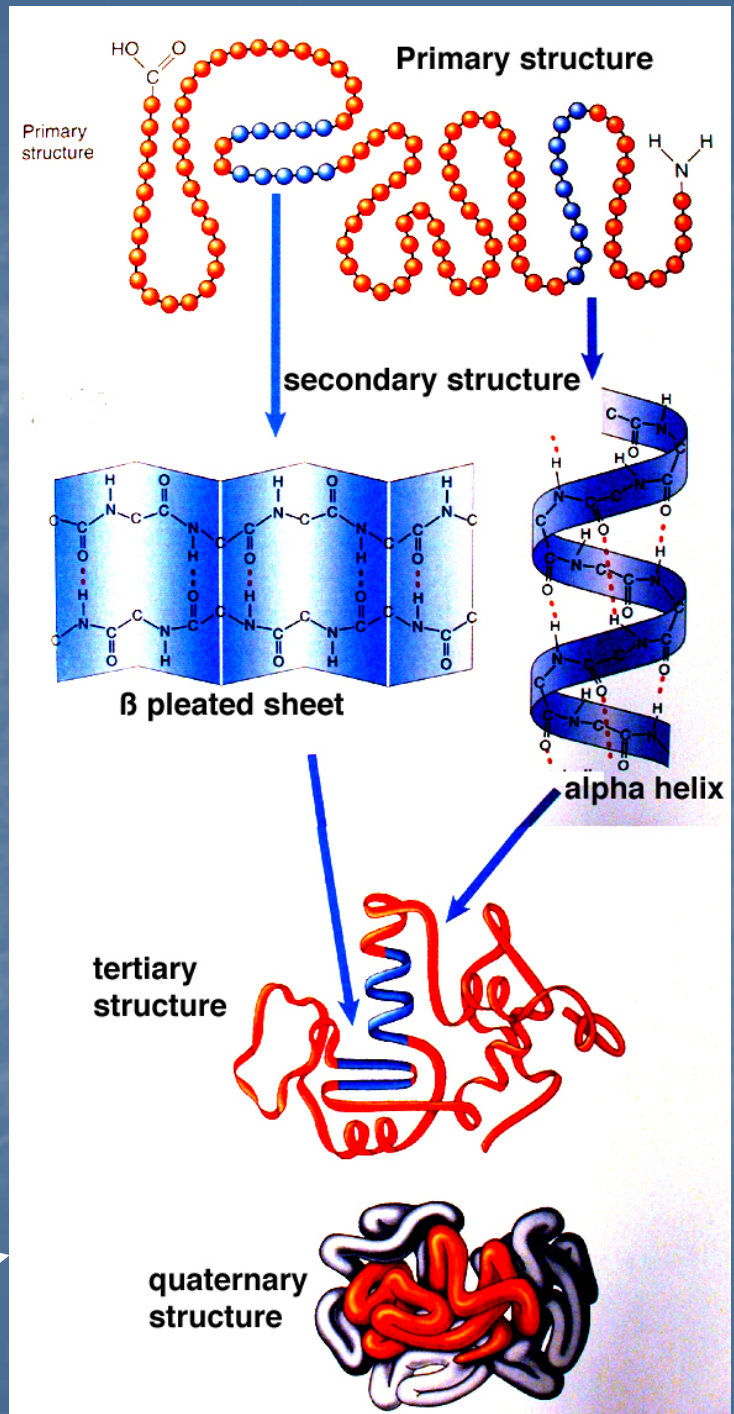
Amino acid → Monomer



Sequence

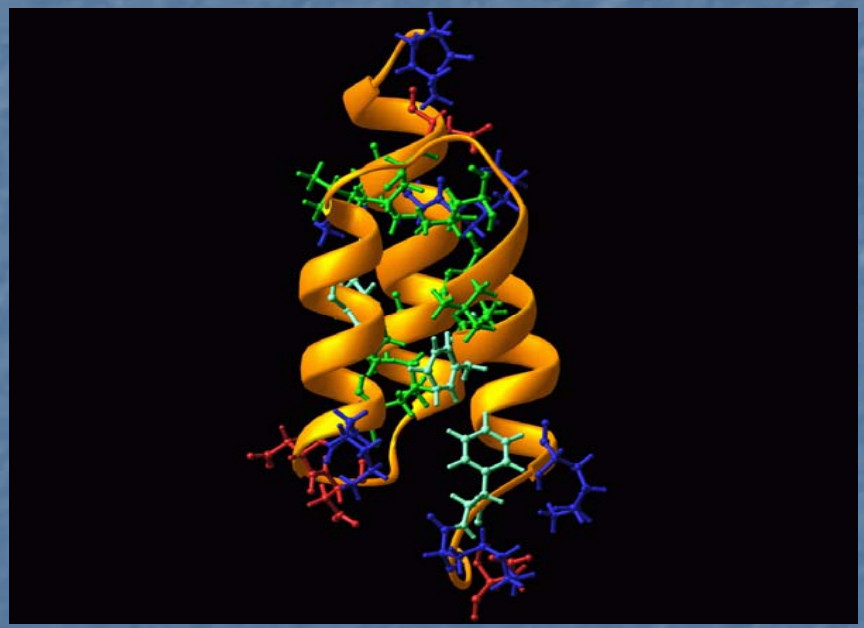


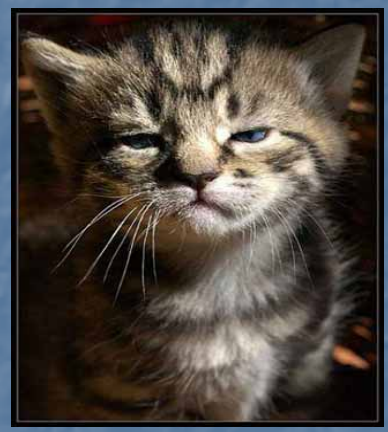
Structure, Configuration

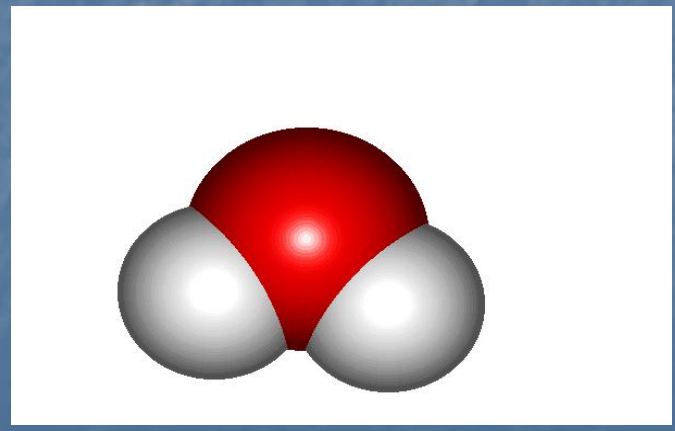


Proteins are something in between.

Biology

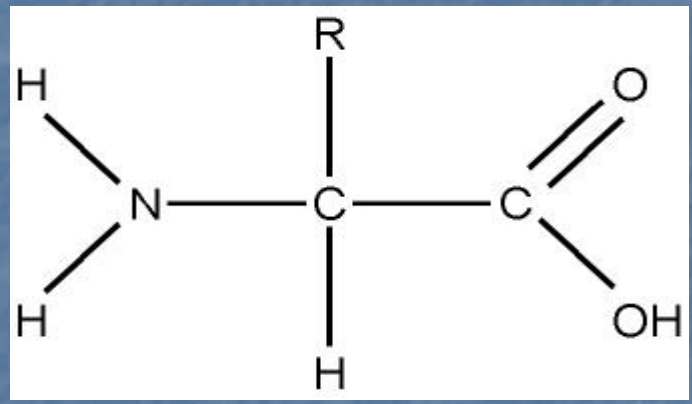






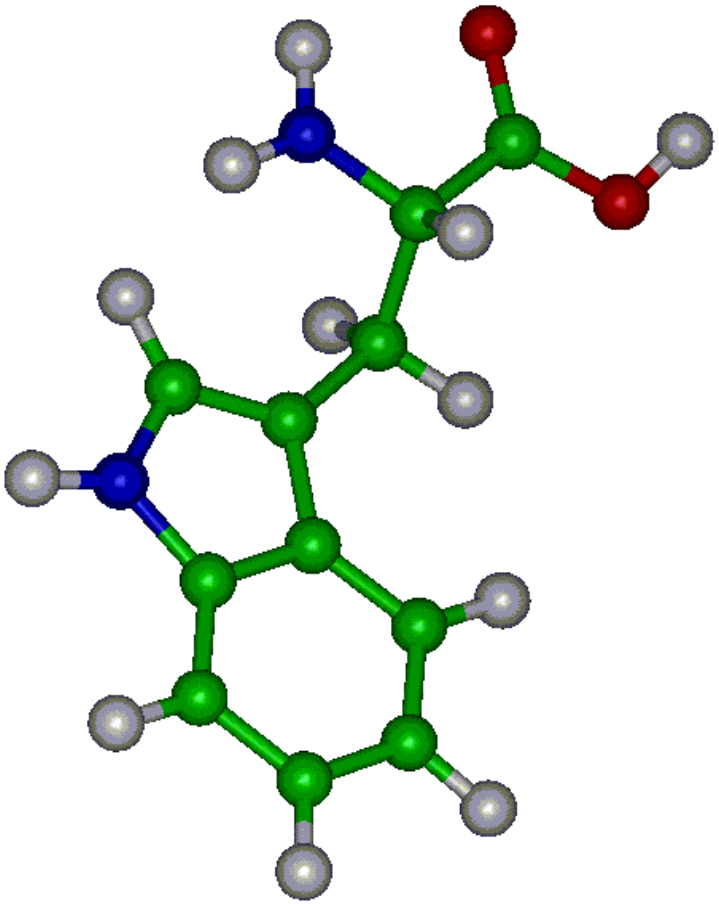
Chemistry

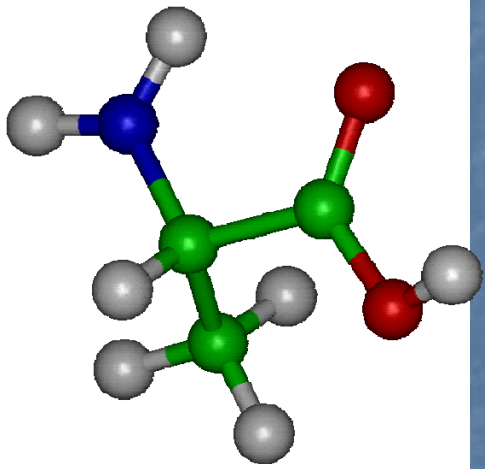
Amino acids



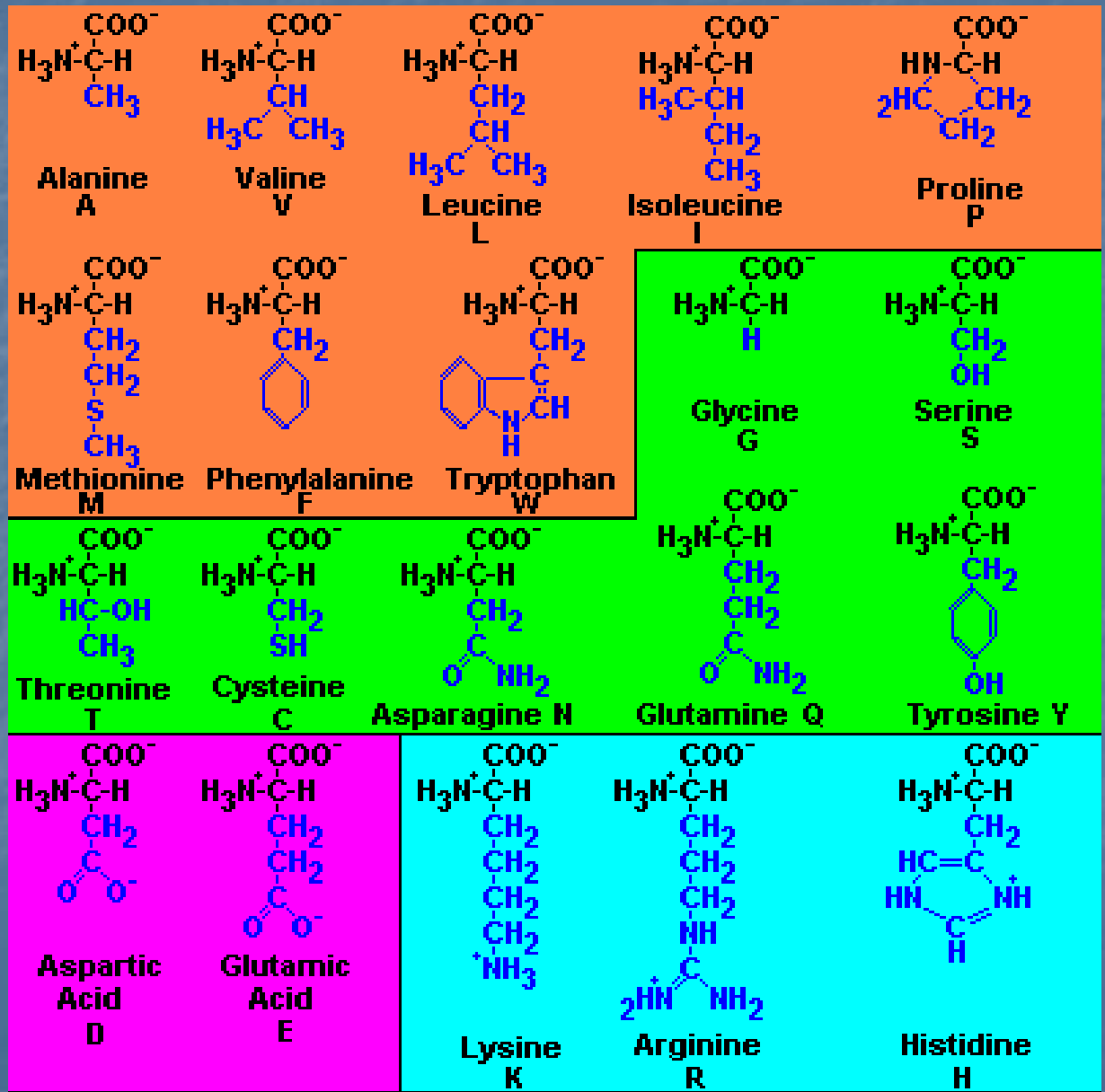
Amino

Acid

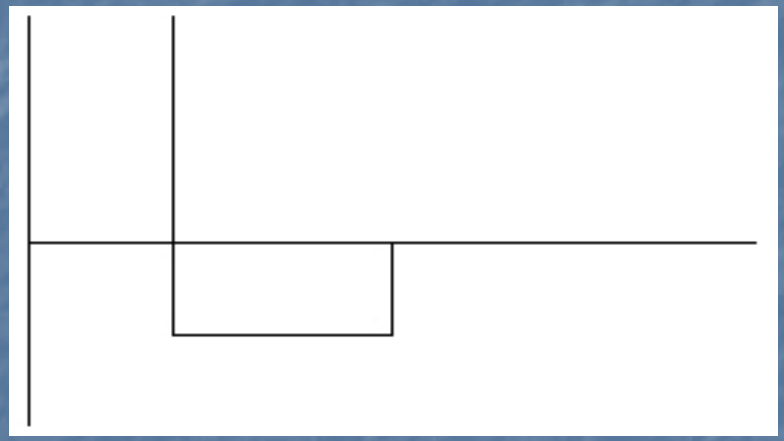


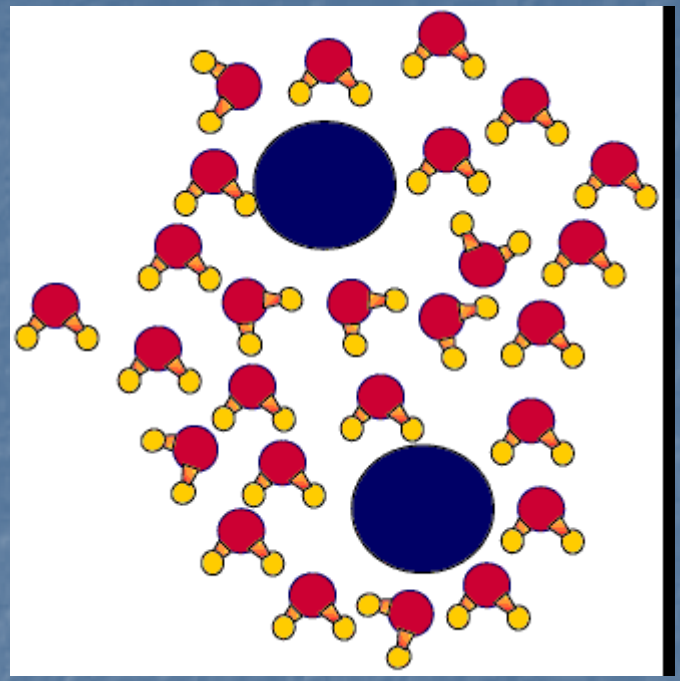


Side chains


Alanine A
Valine V
Leucine L
Isoleucine I
Proline P
Methionine M
Phenylalanine F
Tryptophan W
Glycine G
Serine S
Threonine T
Cysteine C
Asparagine N
Glutamine Q
Tyrosine Y
Aspartic Acid D
Glutamic Acid E
Lysine K
Arginine R
Histidine H

Cutoff





This picture works properly because of the short range behavior of interactions due to screening.

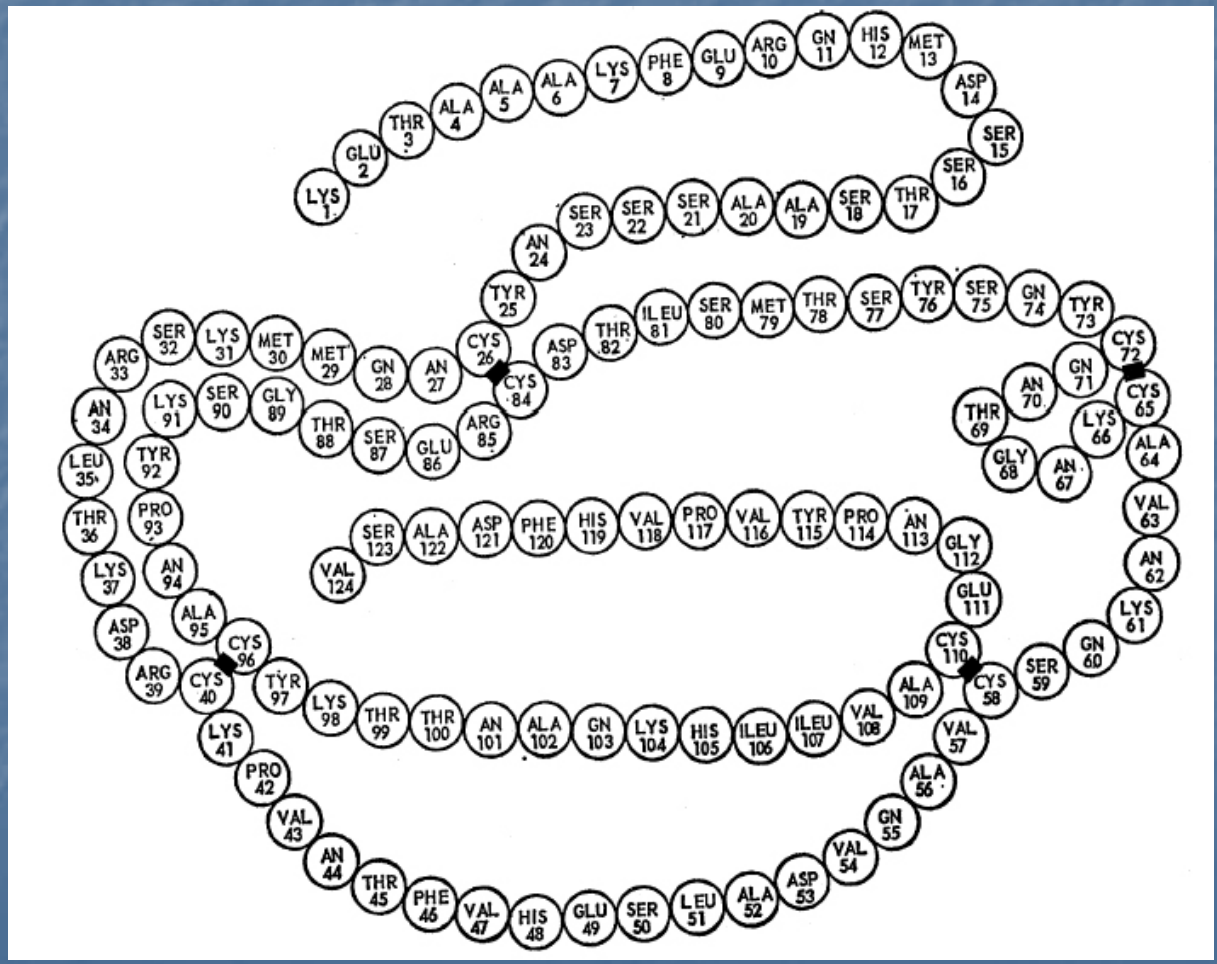
Contact

There is no unique and well defined definition for cutoff.

1. The spatial distance of the main chain carbons be less than 6.5 Angstrom.
2. The spatial distance of any of the heavy atoms (all but Hydrogens) be less than 4.5 Angstrom.

The general properties is found to be almost the same.

Coarse-Graining

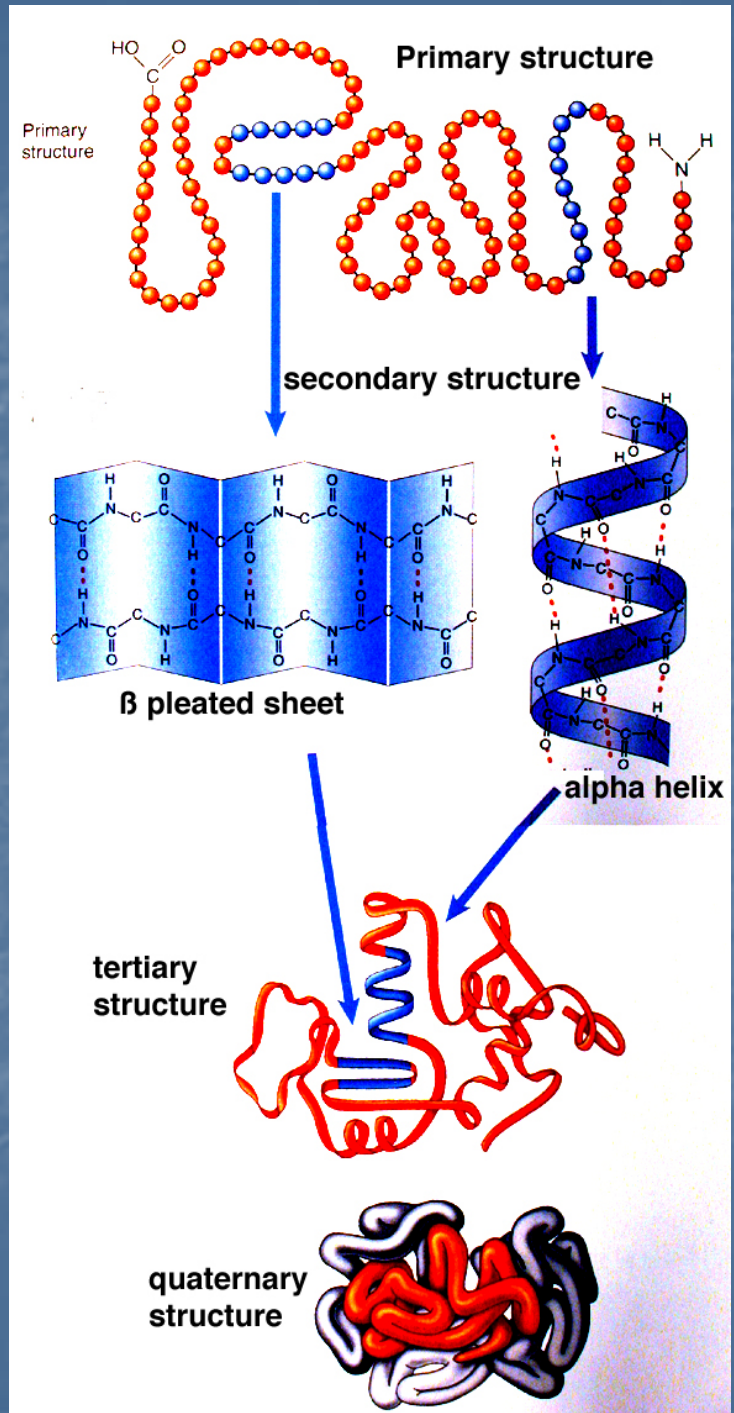


Contact matrix

$$S_{ij} = \begin{cases} 1 & \text{if } |\mathbf{r}_i - \mathbf{r}_j| < R_c \\ 0 & \text{otherwise.} \end{cases}$$

Protein Folding is a coding problem. Enough information to design a 3D structure for a protein exists in a short code in DNA.

The coding has more information in, when it chooses the target structure from a larger set of structures.



Miyazawa-Jernigan interaction

Table 3. Contact energies in RT units; e_{ij} for upper half and diagonal and e'_{ij} for lower half

		Cys	Met	Phe	Ile	Leu	Val	Trp	Tyr	Ala	Gly	Thr	Ser	Asn	Gln	Asp	Glu	His	Arg	Lys	Pro	
	Cys	−5.44	−4.99	−5.80	−5.50	−5.83	−4.96	−4.95	−4.16	−3.57	−3.16	−3.11	−2.86	−2.59	−2.85	−2.41	−2.27	−3.60	−2.57	−1.95	−3.07	Cys
	Met	0.46	−5.46	−6.56	−6.02	−6.41	−5.32	−5.55	−4.91	−3.94	−3.39	−3.51	−3.03	−2.95	−3.30	−2.57	−2.89	−3.98	−3.12	−2.48	−3.45	Met
	Phe	0.54	−0.20	−7.26	−6.84	−7.28	−6.29	−6.16	−5.66	−4.81	−4.13	−4.28	−4.02	−3.75	−4.10	−3.48	−3.56	−4.77	−3.98	−3.36	−4.25	Phe
	Ile	0.49	−0.01	0.06	−6.54	−7.04	−6.05	−5.78	−5.25	−4.58	−3.78	−4.03	−3.52	−3.24	−3.67	−3.17	−3.27	−4.14	−3.63	−3.01	−3.76	Ile
	Leu	0.57	0.01	0.03	−0.08	−7.37	−6.48	−6.14	−5.67	−4.91	−4.16	−4.34	−3.92	−3.74	−4.04	−3.40	−3.59	−4.54	−4.03	−3.37	−4.20	Leu
	Val	0.52	0.18	0.10	−0.01	−0.04	−5.52	−5.18	−4.62	−4.04	−3.38	−3.46	−3.05	−2.83	−3.07	−2.48	−2.67	−3.58	−3.07	−2.49	−3.32	Val
	Trp	0.30	−0.29	0.00	0.02	0.08	0.11	−5.06	−4.66	−3.82	−3.42	−3.22	−2.99	−3.07	−3.11	−2.84	−2.99	−3.98	−3.41	−2.69	−3.73	Trp
	Tyr	0.64	−0.10	0.05	0.11	0.10	0.23	−0.04	−4.17	−3.36	−3.01	−3.01	−2.78	−2.76	−2.97	−2.76	−2.79	−3.52	−3.16	−2.60	−3.19	Tyr
	Ala	0.51	0.15	0.17	0.05	0.13	0.08	0.07	0.09	−2.72	−2.31	−2.32	−2.01	−1.84	−1.89	−1.70	−1.51	−2.41	−1.83	−1.31	−2.03	Ala
	Gly	0.68	0.46	0.62	0.62	0.65	0.51	0.24	0.20	0.18	−2.24	−2.08	−1.82	−1.74	−1.66	−1.59	−1.22	−2.15	−1.72	−1.15	−1.87	Gly
	Thr	0.67	0.28	0.41	0.30	0.40	0.36	0.37	0.13	0.10	0.10	−2.12	−1.96	−1.88	−1.90	−1.80	−1.74	−2.42	−1.90	−1.31	−1.90	Thr
	Ser	0.69	0.53	0.44	0.59	0.60	0.55	0.38	0.14	0.18	0.14	−0.06	−1.67	−1.58	−1.49	−1.63	−1.48	−2.11	−1.62	−1.05	−1.57	Ser
	Asn	0.97	0.62	0.72	0.87	0.79	0.77	0.30	0.17	0.36	0.22	0.02	0.10	−1.68	−1.71	−1.68	−1.51	−2.08	−1.64	−1.21	−1.53	Asn
	Gln	0.64	0.20	0.30	0.37	0.42	0.46	0.19	−0.12	0.24	0.24	−0.08	0.11	−0.10	−1.54	−1.46	−1.42	−1.98	−1.80	−1.29	−1.73	Gln
	Asp	0.91	0.77	0.75	0.71	0.89	0.89	0.30	−0.07	0.26	0.13	−0.14	−0.19	−0.24	−0.09	−1.21	−1.02	−2.32	−2.29	−1.68	−1.33	Asp
	Glu	0.91	0.30	0.52	0.46	0.55	0.55	0.00	−0.25	0.30	0.36	−0.22	−0.19	−0.21	−0.19	0.05	−0.91	−2.15	−2.27	−1.80	−1.26	Glu
	His	0.65	0.28	0.39	0.66	0.67	0.70	0.08	0.09	0.47	0.50	0.16	0.26	0.29	0.31	−0.19	−0.16	−3.05	−2.16	−1.35	−2.25	His
	Arg	0.93	0.38	0.42	0.41	0.43	0.47	−0.11	−0.30	0.30	0.18	−0.07	−0.01	−0.02	−0.26	−0.91	−1.04	0.14	−1.55	−0.59	−1.70	Arg
	Lys	0.83	0.31	0.33	0.32	0.37	0.33	−0.10	−0.46	0.11	0.03	−0.19	−0.15	−0.30	−0.46	−1.01	−1.28	0.23	0.24	−0.12	−0.97	Lys
	Pro	0.53	0.16	0.25	0.39	0.35	0.31	−0.33	−0.23	0.20	0.13	0.04	0.14	0.18	−0.08	0.14	0.07	0.15	−0.05	−0.04	−1.75	Pro
$e_{rr}-2.55$	e_{tr}	−3.57	−3.92	−4.76	−4.42	−4.81	−3.89	−3.81	−3.41	−2.57	−2.19	−2.29	−1.98	−1.92	−2.00	−1.84	−1.79	−2.56	−2.11	−1.52	−2.09	
$e_r-3.60$	e_t	−4.29	−4.73	−5.57	−5.29	−5.71	−4.72	−4.41	−3.87	−3.17	−2.53	−2.63	−2.27	−2.14	−2.35	−2.02	−2.07	−2.94	−2.43	−1.82	−2.53	
$f_r-3.60$	f_t	−5.58	−6.14	−7.39	−7.09	−7.88	−6.15	−5.34	−4.60	−3.24	−2.22	−2.48	−1.92	−1.74	−1.93	−1.54	−1.49	−2.91	−2.07	−1.17	−1.97	
N_{tr}/N_t	2.096	2.723	2.722	2.780	2.811	2.893	2.728	2.537	2.493	2.143	1.840	1.973	1.771	1.699	1.720	1.598	1.508	2.075	1.787	1.343	1.629	
q_t 7.162	6.281	6.646	6.137	5.870	6.042	6.087	6.155	5.793	6.037	6.334	6.284	6.486	6.582	6.574	6.469	6.487	6.235	6.241	6.318	6.569	5.858	

Statistical method on real proteins in data bank. They used 6.5 Angstrom because maximum number of contacts occur in this distance.

The question

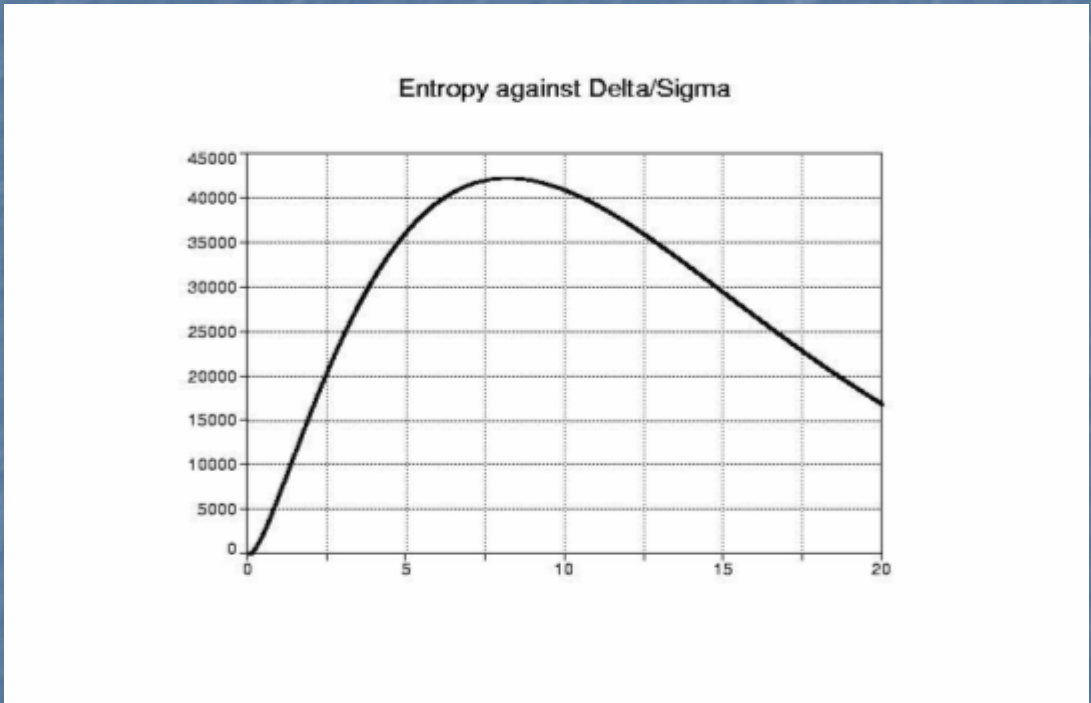
For cutoff=0 number of matrices = 0.

How many nondegerate contact matrices may exist for a specific cutoff distance?

For cutoffs larger than the protein size, number of matrices = 0.

Is there any maximum?

Are we looking for these kind of graphs?



Mean field approximation

- Assuming a FJC the RMSD between all residues is found.
- Number of different possible matrices is assumed to be proportional to the entropy of the matrix.

Analysis of PDB-like self-avoiding walks

- The frequency of contact lengths and angles of real proteins is found and the same distribution is applied to the SAWs.
- Number of nondegenerate contact matrices is found for different cutoffs.

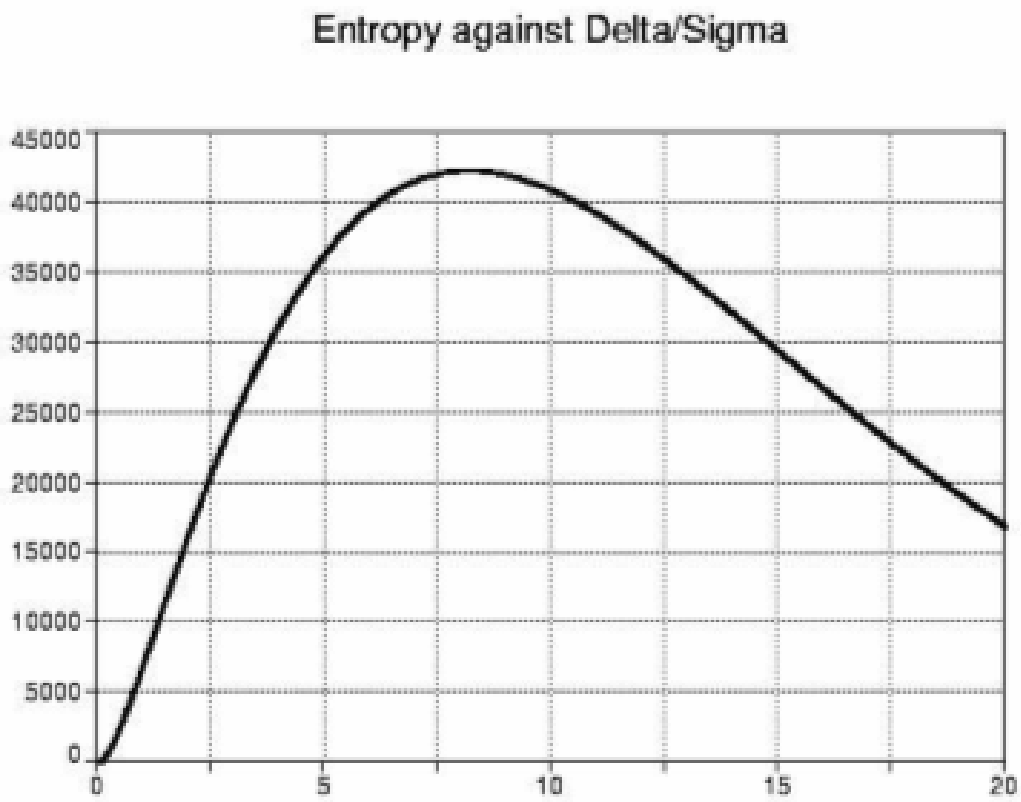
Probability of finding to residues at a specific distance

$$P(m,\Delta)=\int_0^{\infty}4\pi r^2 dr\,\frac{1}{\left(2\pi m\sigma^2\right)^{3/2}}e\frac{-r^2}{2m\sigma^2}\theta(r-\Delta)$$

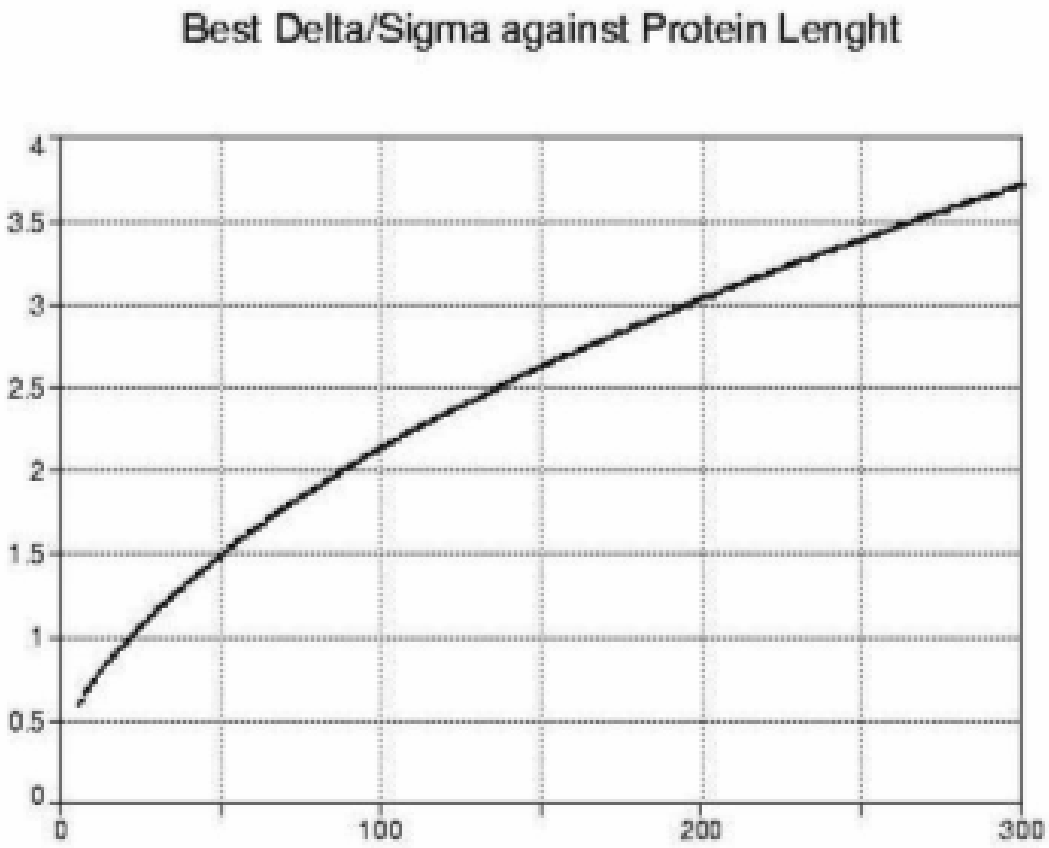
Entropy of the matrix

$$F(\Delta)=\sum_{i=1}^{N-3}\sum_{j=i+2}^{n}-P(i-j,\Delta)\ln p(i-j,\Delta)-(1-p)\ln(1-p)$$

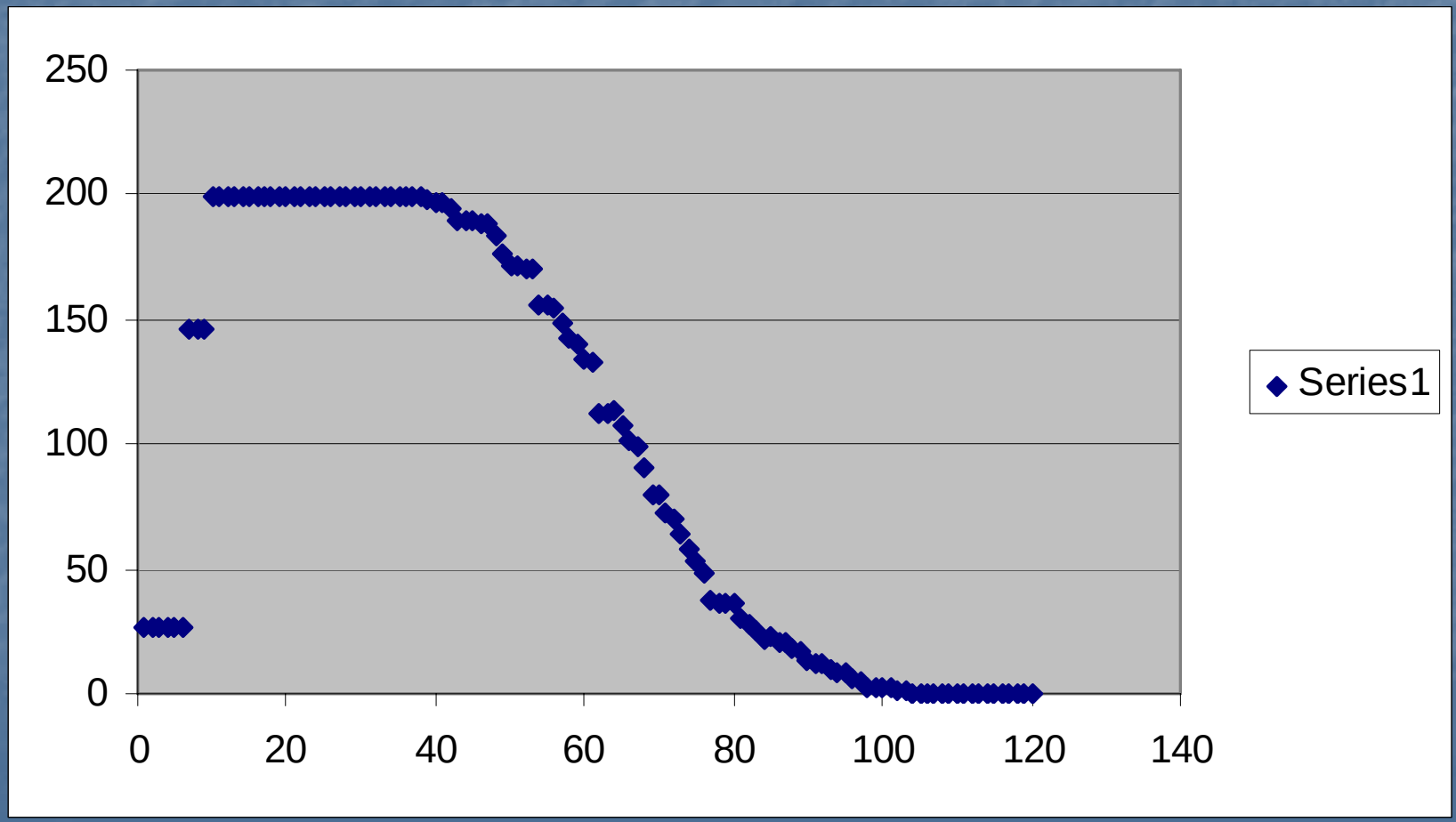
Entropy *vs* Cutoff



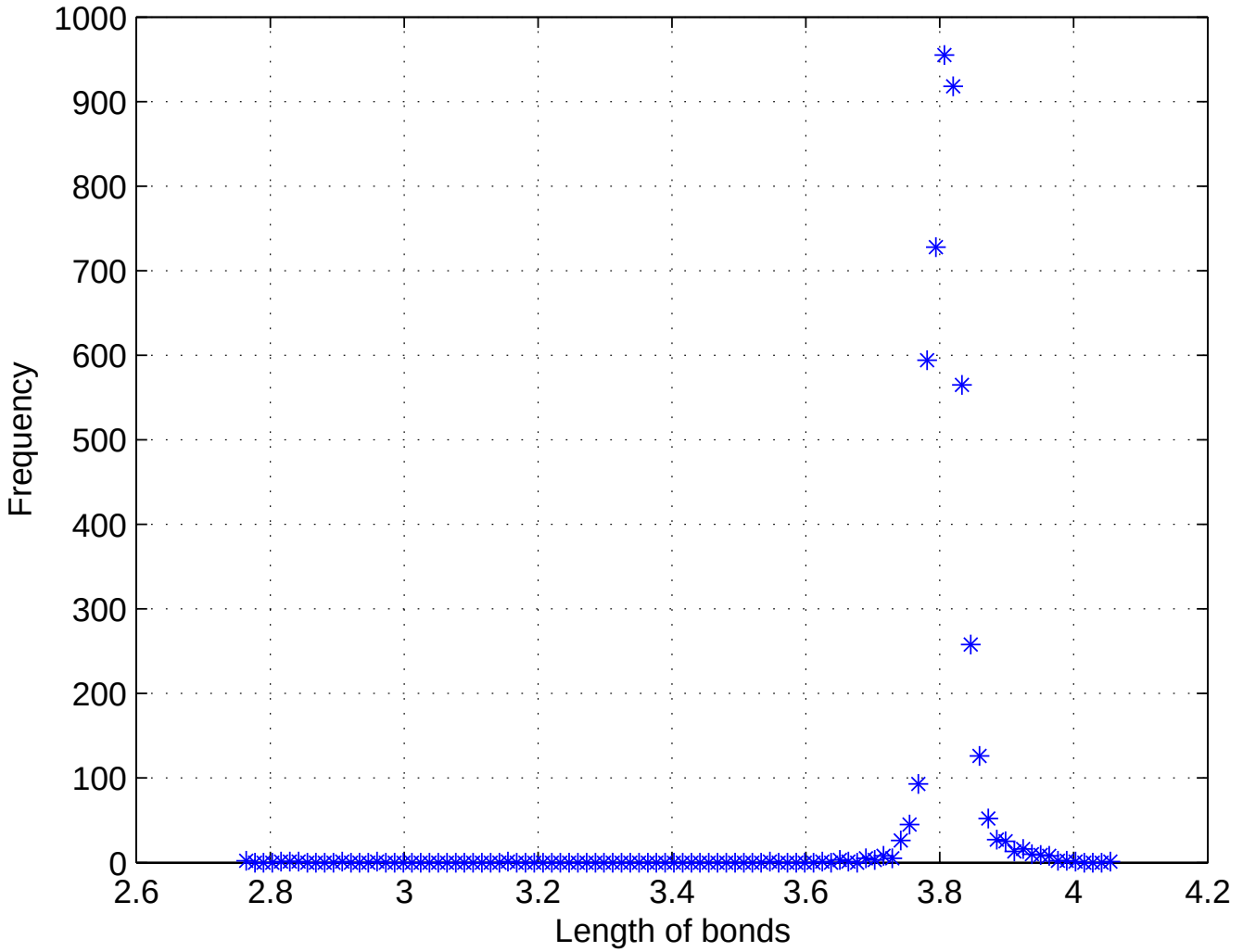
Optimum cutoff for different lengths



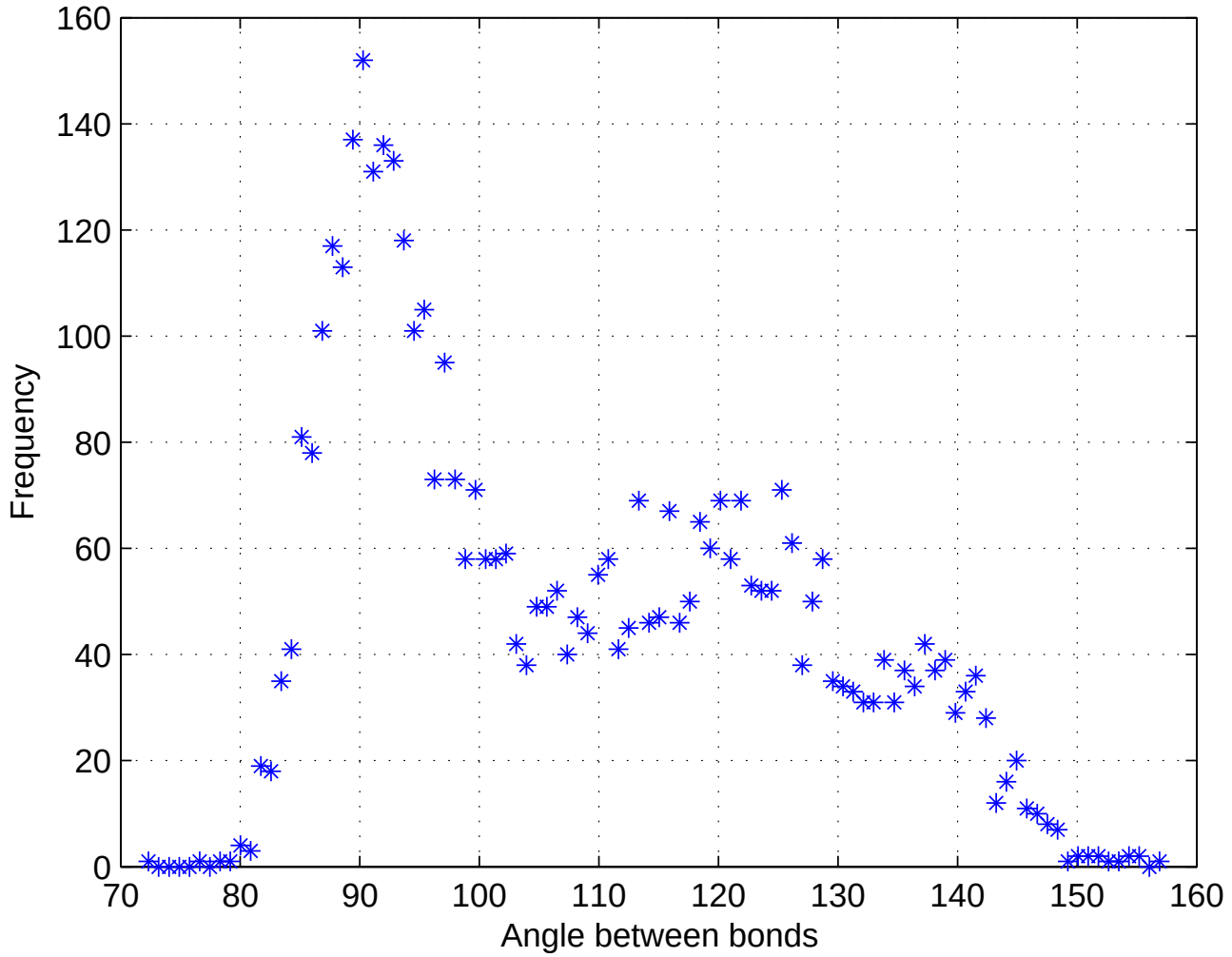
Number of nondegenerate matrices *vs* cutoff (in 0.1 Angstrom) in a diamond lattice



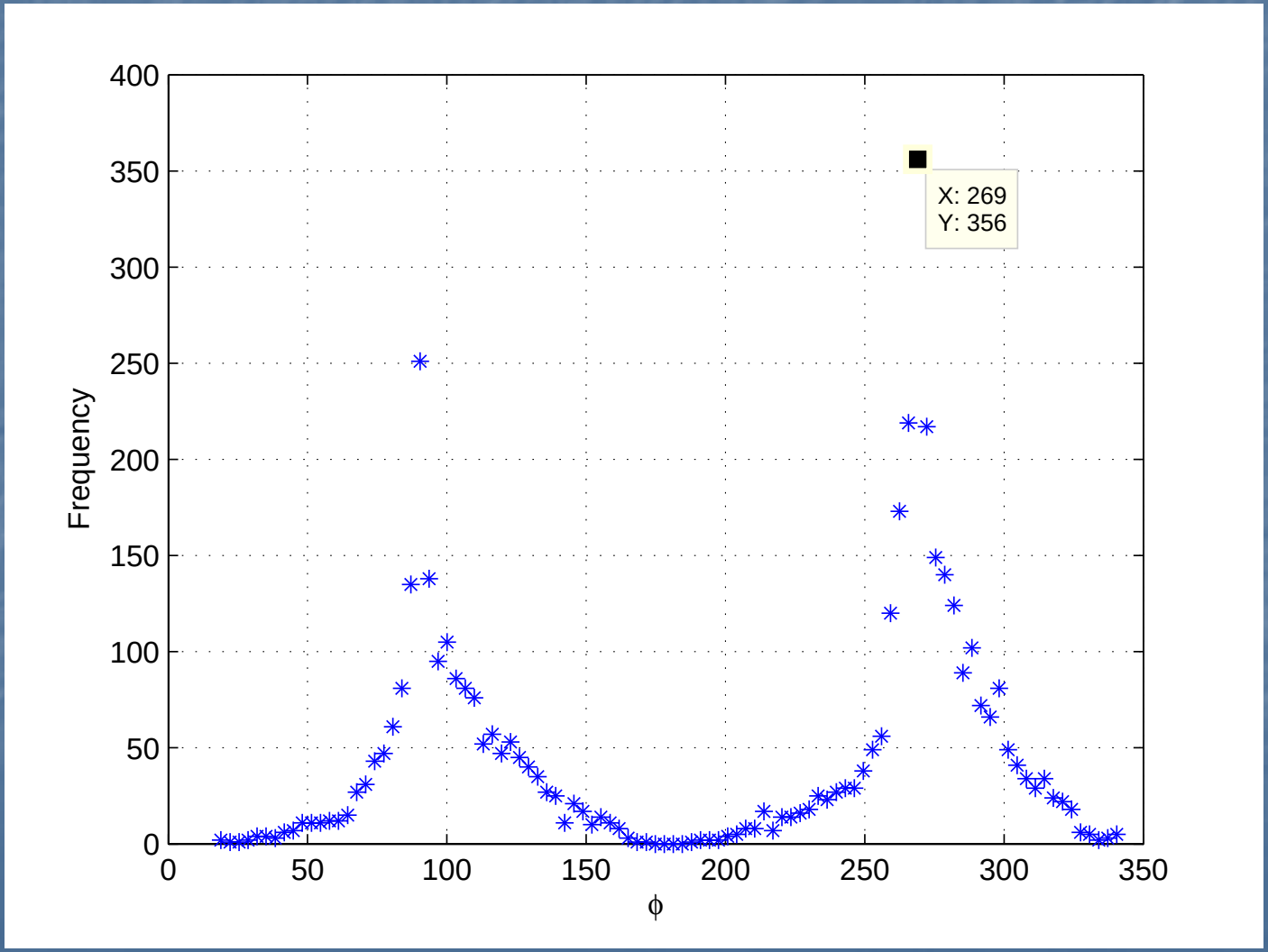
Contact Length Distribution in PDB



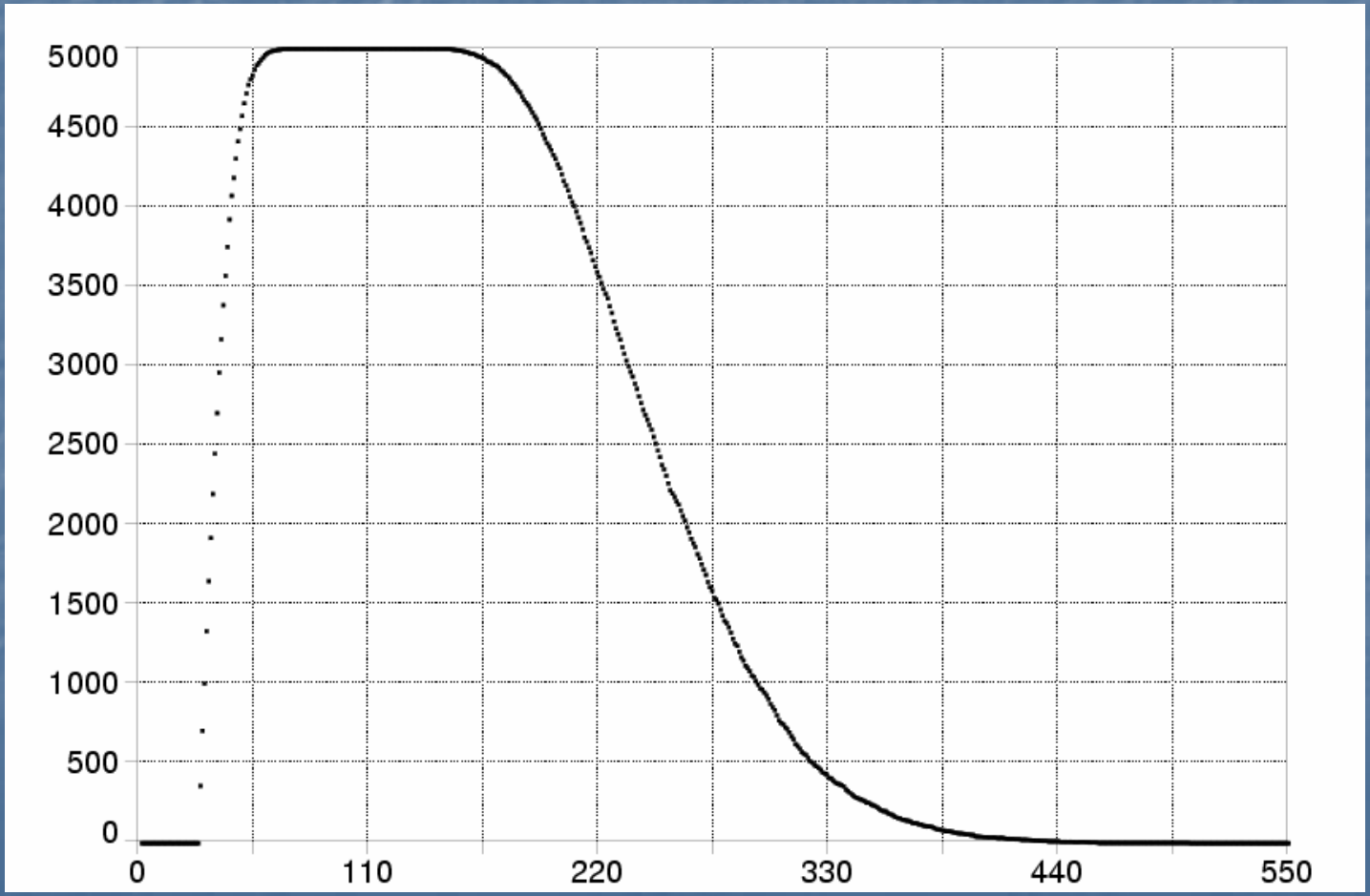
Contact Angle Distribution in PDB



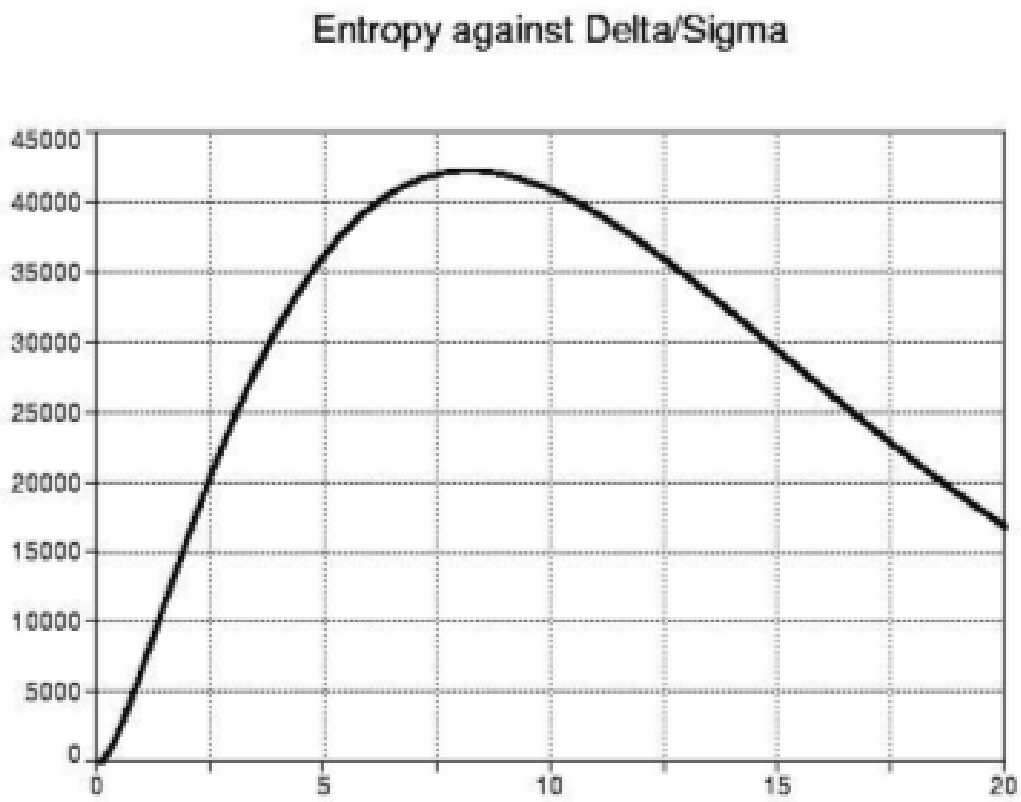
Angle between the contact planes distribution in PDB



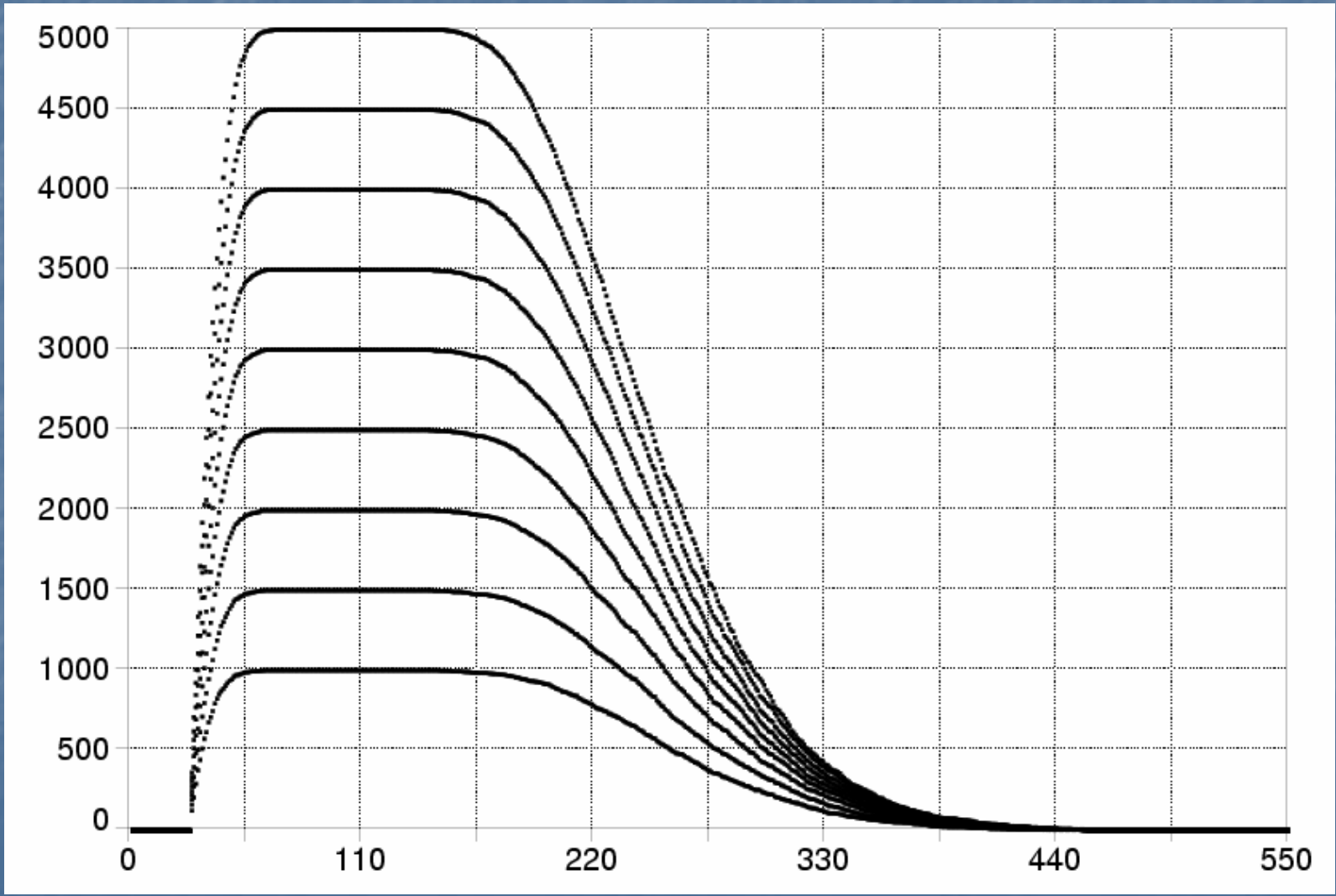
Number of nondegenerate matrices *vs* cutoff for PDB-like structures



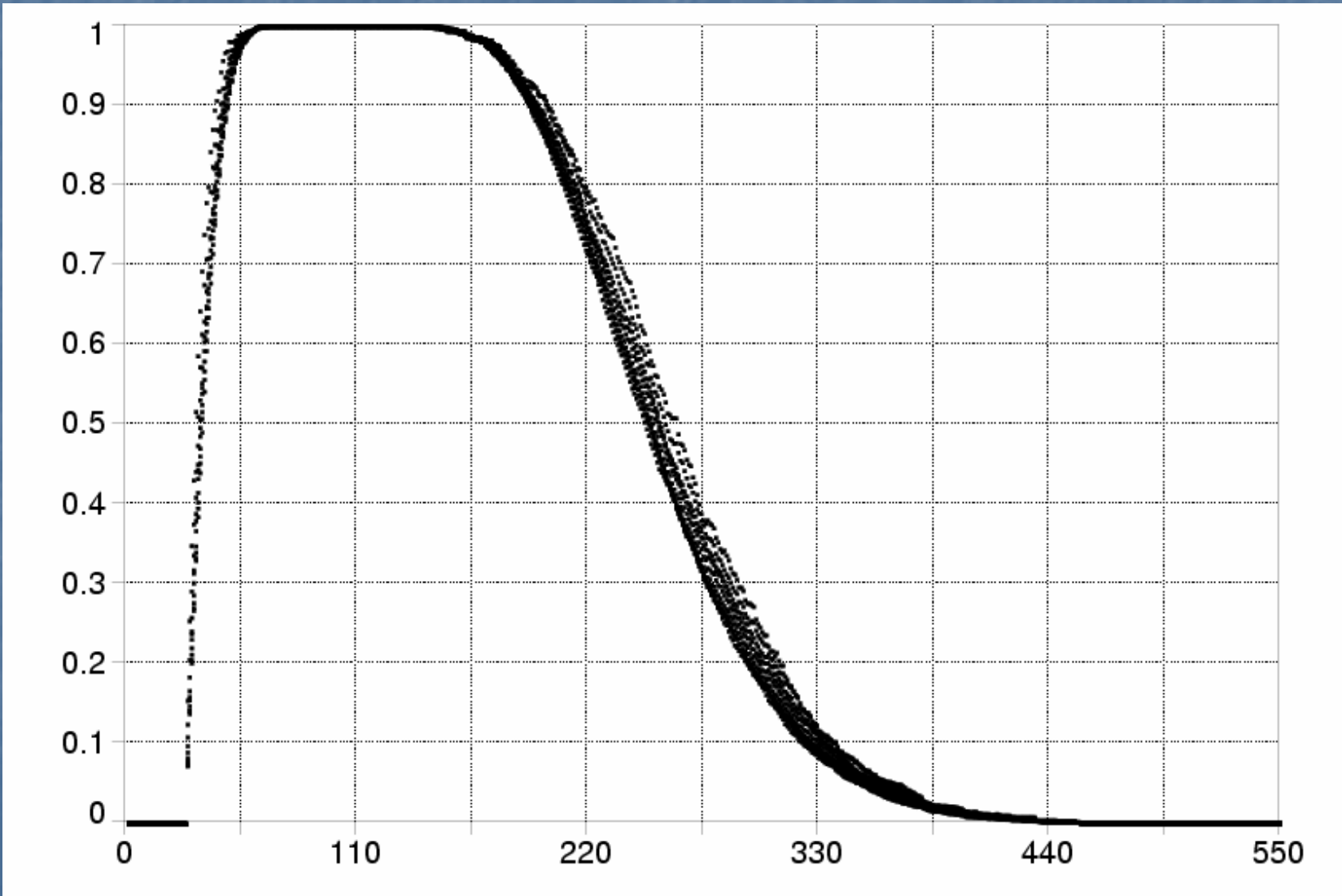
Entropy *vs* Cutoff



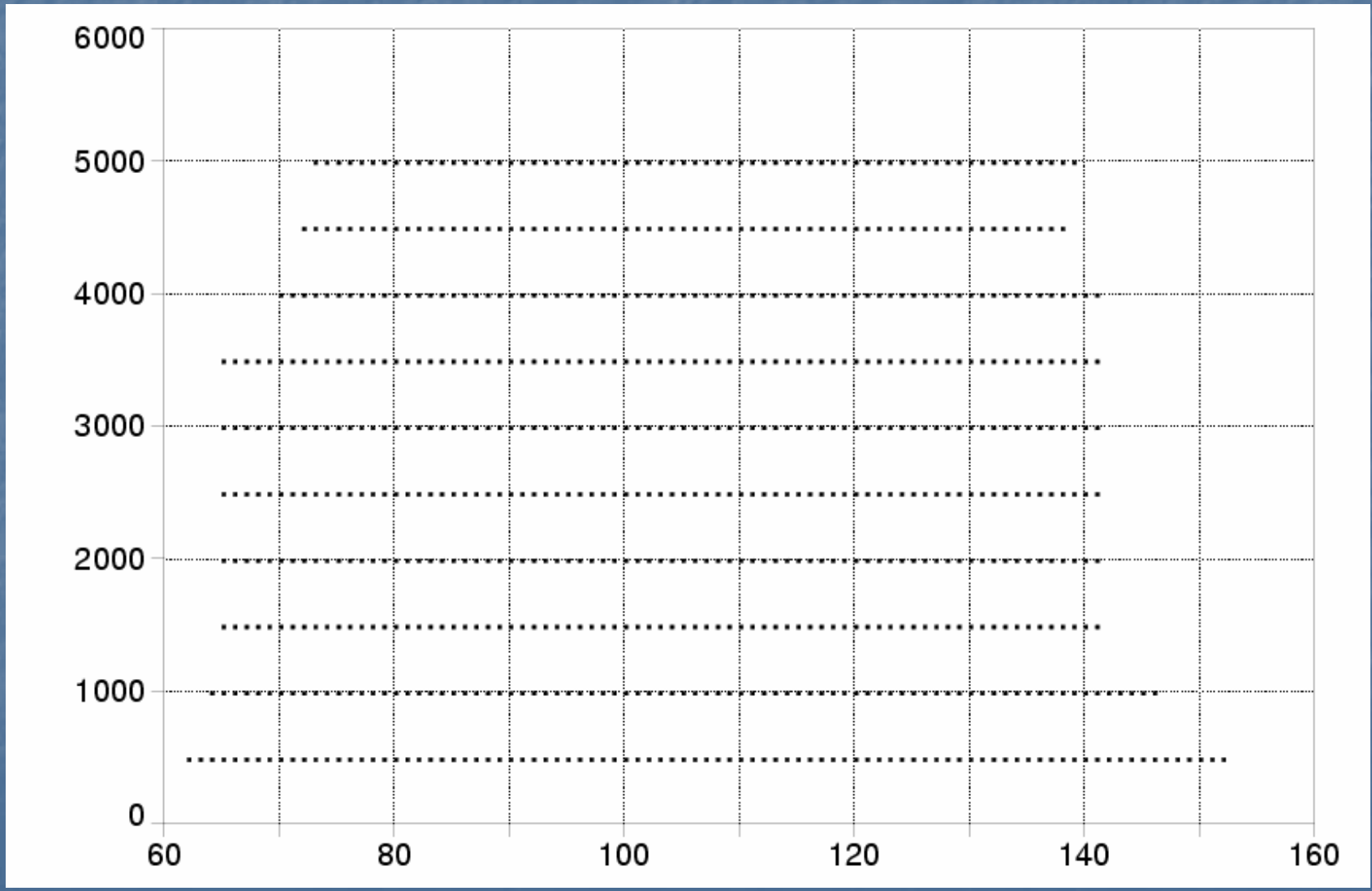
Different lengths



Scaled



Maximum?



Conclusion

- The maximum number of nondegenerate contact matrices is found to be a function of protein length in theoretic approach and is at the same order of 6.5 angstrom
- The number of nondegenerate contact matrices has its maximum value in range which contains 6.5 angstrom for PDB-like structures with 20 monomers.