Title
Protein structural families and the contribution of compactness

Author
Yee, David Paul

Publication Date
1994

Peer reviewed|Thesis/dissertation

Protein structural families and the contribution of compactness

by

David Paul Yee

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Chemistry

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA

San Francisco

Date **University Librarian**

Degree Conferred: . . . . . .

To my parents, Chet and Daisy, for their limitless love and support.

Acknowledgements

A dissertation is a story of one's research efforts. It provides a record of the process by which a student of science becomes a *scientist*. This dissertation not only stands as a record of my research accomplishments, but also marks a closure to an incredible period of my life. Thinking back upon my "graduate student" years, I am amazed by the number and variety of people I have met and interacted with. There are too many people to thank and acknowledge to construct an exhaustive list. I will surely forget many; if you are one, please accept my apologies.

First and foremost, I would like to thank Ken Dill, my research advisor, who provided me with a wonderful research environment. Sarina Bromberg, who gave me priceless advice & support when I needed it. I will always remember the wonderful Tassajara journal clubs and inspiring "mandatory beach days" with Brian Shoichet, Lydia Gregoret & Chuck Wilson. Chris Carreras has been a great housemate with whom I could always share new ideas. Thanks go to Mark Yee, my brother, for his presence as a housemate and friend. Finally, an extra special thanks goes to Catherine Ager for her patience through this period, for her guidance & advice, and for her Love.

Authorship

The text of chapter 2 contains work originally published in Protein Science (1993), volume 2, pages 884 - 899 with the title ''Families and the Structural Relatedness among Globular Proteins''. The coauthor listed on this paper directed and supervised the research which forms the basis for chapter 2.

The text of chapter 3 contains work originally published in the Journal of Molecular Biology (1994), volume 241, number 4, pages 557 - 573 with the title ''Does Compactness Induce Secondary Structure in Proteins - A Study of Poly-alanine Chains Computer by Distance Geometry''. This paper includes Dr. Hue Sun Chan, Dr. Timothy F. Havel, and Professor Ken A. Dill as coauthors. The work was supervised by Professor Dill. Dr. Chan initiated the investigation and performed the preliminary research and analysis. Dr. Havel performed the distance geometry calculations to generate random, compact poly-alanine chains. David Yee coordinated the project, performed virtually all the analysis, and was solely responsible in seeing the project through to its completion.

Ken Dill

Protein structural families and the contribution of compactness.

David Paul Yee

Abstract

Proteins are fundamental to life as we know it. They are the microscopic machines which catalyze essential chemical reactions. They act as critical structural supports in our cells. They are the end product of the "central dogma" of molecular biology; nucleic acids *code* for proteins. The problem of predicting the structure of a protein from its amino acid sequence is known as the protein folding problem. Thousands of biophysical characterizations and theoretical studies have been published in an attempt to understand how the amino acid sequence of a polypeptide chain carries the information necessary for a protein to fold into its precise 3 dimensional structure. In the nearly four decades that have passed since the first crystal structure of a protein was solved, hundreds of high resolution protein crystal structures have been determined. Yet the protein folding problem remains unsolved.

The work in this thesis represents an attempt to understand a small part of the protein folding puzzle. I first describe the development of a measure of protein structural dissimilarity. The dissimilarity measure is used to search databases for the presence of specific substructures and structural motifs. It is then used to compare, pairwise, a large dataset of diverse protein structures. Clustering methods are utilized to automatically partition proteins into unique structural classes. An analogy to points randomly distributed in an d-dimensional Euclidean space is used to ask if protein families are tightly-knit or loosely-knit entities.

Next, I analyze ensembles of random, compact poly-alanine conformations to explore the relationship between compactness and secondary structures in protein. The work shows that there is an entropy which stabilizes ordered structures in compact ensembles of polymers. I conclude by describing a method to map protein structures onto a cubic lattice. Distance and bond projection correlation functions are used to characterize the differences between real proteins and simplified models of proteins.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

If nucleic acids represent the genetic code, then proteins surely represent life's machinery. In introductory biology courses, one is taught that the information which codes for life is present in deoxyribonucleic acid (DNA). Genes, specific sequences of DNA, are transcribed into special ribonucleic acids known as messenger ribonucleic acids (or mRNA). mRNA is in turn translated by a complicated protein/nucleic acid complex into proteins. The massive effort to sequence the human genome is driven by our desire to identify the genes that code for the tiny molecular machines of which we are comprised.

Proteins are polymers of amino acids which, by virtue of specific sequences of amino acids, can fold into highly defined three-dimensional shapes. These structures are so precise that proteins can recognize and modify molecules with incredible precision and specificity.

There are 20 different naturally occurring amino acids. All amino acids consist of an amino group, a carboxyl group, a hydrogen atom, and a side chain bonded to a central carbon which is designated as the α-carbon. The side chains differ widely in their chemical composition. They can be polar or non-polar. They can be large and bulky or small. A series of amino acids linked together via *peptide bonds* forms a polypeptide. If the polypeptide consists of an amino acid sequence with certain properties, it can fold into a protein.

The experiments of Christian Anfinsen in the early 1970's on ribonuclease demonstrated that the information required to fold a polypeptide into a protein is contained in the amino acid sequence itself [1]. By denaturing ribonuclease with urea and β-mercaptoethanol, ribonuclease was unfolded and activity was destroyed. Anfinsen observed that upon dilution of the denaturants, activity was restored. Since denatured proteins consist of fully unfolded polypeptides, the observation that activity could be

restored upon dilution of denaturant provided strong evidence that proteins fold to their thermodynamically most stable form. The question of how a particular sequence of amino acids codes for a protein's structure is known as the protein folding problem.

The first crystal structure of a protein, myoglobin, was published in 1960 [2]. After nearly four decades, there are several hundred high resolution crystal structures in the Brookhaven Protein Data Bank (PDB) [3] [4]. Despite the wealth of structural data, thousands of biophysical characterizations, and thousands of theoretical studies, the protein folding problem remains unsolved.

The work in this thesis represents an attempt to understand a small part of the protein folding puzzle. In the next section, I describe the development of a measure of protein structural dissimilarity. In the second part of this thesis, I analyze ensembles of random poly-alanine conformations to gain an understanding of the relationship between compactness and secondary structures in protein. I conclude by describing work mapping protein structures onto a cubic lattice and discuss ways of characterizing the difference between real proteins and simplified models of proteins.

For further reading, introductory biochemistry texts by Stryer [5] and Zubay [6] are very good. Introductions to protein structure can be found in books by Schulz and Shirmer [7], Brandon and Tooze [8], and Lesk [9].

References

[1] C. B. Anfinsen, "Principles that Govern the Folding of Protein Chains," *Science*, 181(4096):1973.

[2] J. C. Kendrew, R. E. Dickerson, B. E. Strandberg, R. G. Hart, D. R. Davies, D. C. Phillips, and V. C. Shore, "Structure of Myoglobin," *Nature*, 185:422- 427, 1960.

[3] E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng, in: *Crystallographic Databases - Information Content, Software Systems, Scientific Applications*, F. H. Allen, G.Bergerhoff, and R. Seivers, ed., p. 107 - 132, Data Commission of the Int'l Union of Crystallography, Bonn/Cambridge/Chester, 1987.

[4] F. C. Bernstein, T. F. Koetzle, G. Williams, E. Meyer, M. Brice, J. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures," *Journal of Molecular Biology*, 112:535-542, 1977.

[5] L. Stryer, *Biochemistry*, Freeman, San Francisco, CA, 1983.

[6] G. L. Zubay, *Biochemistry*, Addison-Wesley, Reading, MA, 1981.

[7] G. E. Schulz and R. H. Schirmer, *Principles of Protein Structure*, Springer-Verlag, New York, 1979.

[8] C. Branden and J. Tooze, *Introduction to Protein Structure*, Garland, New York, 1991.

[9] A. M. Lesk, *Protein Architecture*, (*Practical Approach Series*), IRL Press, Oxford New York Tokyo, 1991.

Chapter 2

Structural Relatedness

&

Protein Families

Preface

A most remarkable treatise is Jane Richardson's work entitled "The Anatomy And Taxonomy Of Protein Structure" [1]. Published in 1981, Richardson reviews the basic elements of protein structure: helices, sheets, turns, etc. More impressively, she classified all known protein structures into broad classes of antiparallel α domains, parallel α/β domains, antiparallel β domains, and small domains containing either bound metal or disulphides. To support her classification, she provided beautifully drawn ribbon diagrams which highlighted the similarities between the proteins. These diagrams, which have become the *de facto* standard way of displaying a protein structure, allow one to clearly see the packing arrangement of the secondary structures within each protein. Richardson's contribution was invaluable because it allowed one to speak of protein families and, therefore, consider the protein folding problem in terms of how and why proteins adopt specific topologies. Although earlier work introduced the concept of protein families [2], Richardson's work allowed the researcher to *visualize* them.

This chapter describes the development of a general measure of protein dissimilarity. The measure, named CONGENEAL (for CONformational GENEALogy), was tested by searching a dataset of protein structures for the presence of related proteins and specific substructures. CONGENEAL was then used to compare pairwise a dataset of 158 protein structures. The pairwise dissimilarity data was used in conjunction with several clustering algorithms to automatically partition proteins into unique structural classes. This work represents a simple attempt to automatically partition proteins into families. An analogy to points randomly distributed in an d-dimensional Euclidean space was used to determine if protein families are tightly-knit or loosely-knit entities.

References

[1] J. S. Richardson, "The Anatomy and Taxonomy of Protein Structure," *Advances in Protein Chemistry*, 34:167 - 339, 1981.

[2] M. Levitt and C. Chothia, "Structural Patterns in Globular Proteins," *Nature*, 261:552 - 558, 1976.

March 9, 1993

Families and the Structural Relatedness among Globular Proteins

David P. Yee & Ken A. Dill

Department of Pharmaceutical Chemistry, Box 1204

University of California

San Francisco, California 94143-1204

U.S.A

Ken Dill

415-476-9964

415-476-1508 (FAX)

dill@maxwell.ucsf.edu

David Yee

415-476-8910

415-476-1508 (FAX)

yee@maxwell.ucsf.edu

Page Count:

Manuscript:		25 pages
Tables:	4 tables,	8 pages
Figure legends:		2 pages
Figures:	16 figures,	16 pages
Supplementary Material:		14 pages

TOTAL: 65 pages

A macintosh disk is included which contains a copy of the manuscript, supplementary material, and 2 potential kinemages.

Abstract

Protein structures come in families. Are families "closely-knit" or "loosely-knit" entities? We describe a measure of relatedness among polymer conformations. Based on weighted distance maps, this measure differs from existing measures mainly in two respects: (i) it is computationally fast, and (ii) it can compare any two proteins, regardless of their relative chain lengths or degree of similarity. It does not require finding relative alignments. The measure is used to determine the dissimilarities between all 12,403 possible pairs of 158 diverse protein structures from the Brookhaven Protein Data Bank (PDB). Combined with minimal spanning trees and hierarchical clustering methods, this measure is used to define structural families. It is also useful for rapidly searching a dataset of protein structures for specific substructural motifs. By using an analogy to distributions of Euclidean distances, we find that protein families are not tightly-knit entities.

Keywords

protein family, structural comparison, relatedness, substructure searches

Introduction

Pioneering work over the past 20 years has shown that proteins fall into families of related structures [1-4]. How many families are there? Are the families "tightly-knit" or "loosely-knit"? That is, do two proteins within a family have much greater structural similarity than two proteins from different families? If so, they are tightly-knit. What can we learn about the forces of protein folding and evolution from observing how proteins cluster into families?

In order to address these questions, it is necessary to have a suitable measure of the structural similarity between proteins, since a "family" relationship can only be defined in terms of some degree of similarity. Several measures of structural similarity have been developed [5-8]. There is no underlying fundamental principle dictating that one similarity measure is better than others. Ultimately, the concept of "similarity" is based upon some criterion arbitrarily chosen for a particular purpose [9]. For example, a common measure of structural similarity is the root-mean-square deviation of atomic positions after superposition (RMS). RMS is a useful distance metric for comparing structures that are nearly identical; for example, when refining or comparing structures obtained from x-ray crystallography or NMR experiments. However, RMS is of limited value as a general measure of similarity since it is a "maximum likelihood estimator" of the standard deviation between two structures only if the individual errors are Gaussian distributed with zero mean [10]. The Gaussian distribution assumption can be reframed as an assumption that the differences between two compared structures arise from fluctuations which obey a square-law potential. A square-law potential is only a good approximation for small conformational deviations. If two structures are not in the same energy well, or if errors are large, RMS will lose its underlying justification. In addition, the use of an RMS distance criterion to compare two protein structures requires making assignments in which atom *i* of protein 1 "is equivalent to" atom *j* of protein 2. When

comparing proteins with little sequence identity or unequal chain lengths, this requires making arbitrary decisions.

Some similarity measures require making structural alignments of one protein with the other [6, 7]. When there is a biological or evolutionary basis for making these alignments, such methods have the advantage of allowing a high degree of structural discrimination among highly similar proteins. For proteins that are not highly similar, however, making alignments requires making certain arbitrary choices about the possible locations of insertions and deletions and the choices of gap penalties. These decisions can be computationally intensive.

Rackovsky has developed a similarity measure that compares distributions of conformations of chain segments up to 4 residues in length [8]. Whereas it captures structural information of residues close together in sequence, our interest here is to capture information about contacting residues at all separations along the chain.

Our purpose here is better served by yet a different measure of structural relatedness. The following questions motivate the need for a different measure. What is the shape of protein conformational space? What is a useful "reaction coordinate" along which a protein folds to its native state? In models of proteins, such as those involving chains on lattices, how similar is a model conformation to the true native conformation? To address these questions, we need a similarity measure for which the two most important criteria are: (i) that it must be able to compare any two conformations, no matter how different, and (ii) that it must entail making the fewest possible arbitrary decisions. Furthermore, the measure must avoid comparing structures based on microscopic details such as hydrogen bond angles, since these are not appropriate for some low-resolution models. Many such problems do not involve insertions, deletions, or gaps, and therefore do not require that a similarity measure have

sophisticated alignment machinery.

If an algorithm that measured structural relatedness were computationally efficient enough, it could also be put to other uses. For example, since the number of known protein structures is $N \approx 100$ - 1000, (depending on whether we choose all known structures, or whether they are selected in some way to avoid repeats of nearly identical molecules), the number of pairwise comparisons involved is $(N \times (N-1))/2 \approx 10^4 - 10^6$. If we could compute all these pairwise "distances", we could measure the interrelatedness among proteins to learn how they cluster into families. Different similarity measures make different trade-offs between speed, number of arbitrary decisions, and discrimination. By choosing a measure that is as simple, fast to compute, and non-arbitrary as possible, we trade off the degree of discrimination among highly similar proteins obtained by other measures, but the latter is less important for our purposes.

The outline of this paper is as follows. We first introduce the algorithm for measuring dissimilarity. (It measures "*dis*similarity" because it is 0 for identical structures, and increases as the structural similarity between two proteins diverges.) We call it CONGENEAL (CONformational GENEALogy) because it compares conformations and can generate family trees describing their relatedness. Much of this paper is devoted to showing that CONGENEAL is a reasonable measure of relatedness. For example, in one test we show that it is a useful tool for searching databases of protein structures to locate specified substructures within proteins. We then apply this measure to the pairwise comparison of 158 diverse protein structures and use clustering algorithms to identify families. Finally, we compare the dissimilarity distribution of protein structures to simulations of points distributed in n-dimensional Euclidean spaces to explore the tightness with which proteins cluster into families.

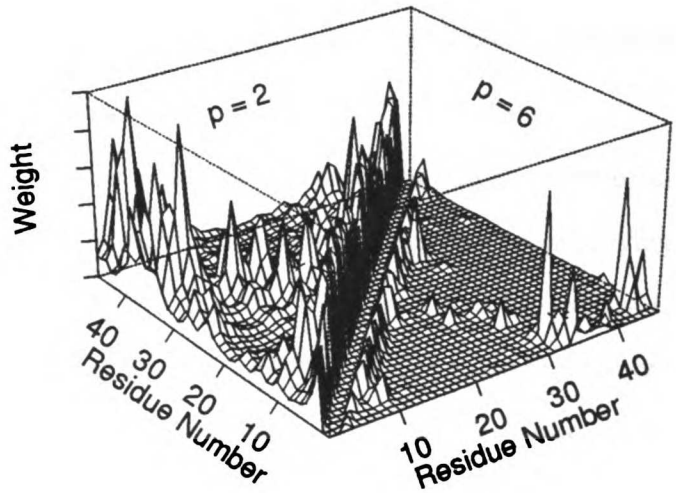


Figure 2.1: Weighted distance maps of crambin with $p = 2$ on the left and $p = 6$ on the right. The height of the peaks corresponds to the magnitude of the weight, w.

CONGENEAL: A Dissimilarity Measure

The CONGENEAL dissimilarity measure compares the *weighted distance maps* of two polymer conformations (see figure 2.1). The weighted distance map of a protein chain conformation that has N residues is an $N \times N$ matrix in which each matrix element (i,j) is a weight, w, equal to the distance, $d_{i,j}$, between the α-carbons of residues i and j, raised to a power $-p$ ($p > 0$):

$$w_{i,j} = d_{i,j}^{-p} \tag{2.1}$$

Two residues which are adjacent in space are assigned a large weight, while two residues which are far apart in space have a small weight. Since the matrix is symmetric, it is only necessary to compute the upper (or lower) triangle of the matrix. The difference between the weighted distance map and the contact map, first introduced by Liljas and

Rossmann [11], is that, in the former, the weights are assigned from a continuous range of values whereas, in the latter, only weights of 0 or 1 are used to indicate whether or not a pair of residues are adjacent.

The distance dependence in equation 2.1 resembles that of intermolecular forces. For the purpose of dissimilarity measures, however, there are no underlying principles that direct us to choose a particular value of p. We have investigated $p = 1$, 2, 4, and 6. When $p = 6$, only the closest neighbors contribute to the weighted distance map. When $p = 2$, pairs of residues separated in space by greater distances also contribute. We compare different values of p below, but the qualitative results are found to be sensibly independent of p. We mainly use $p = 2$.

We now describe how the dissimilarity score is obtained from the weighted distance maps for two conformations. Given two proteins, R and S, let r_{ij} be the distance between residues i and j in protein R and let s_{ij} be the distance between residues i and j in protein S. First, consider a simple case. When R and S have the same chain length, N, and have a direct residue-to-residue alignment, the dissimilarity between the two proteins is given by:

$$d(R,S) = \frac{\sum_{i=1}^{N}\sum_{j=i+2}^{N} \left| r_{ij}^{-p} - s_{ij}^{-p} \right|}{\frac{1}{2}\left[\sum_{i=1}^{N}\sum_{j=i+2}^{N} r_{ij}^{-p} + \sum_{i=1}^{N}\sum_{j=i+2}^{N} s_{ij}^{-p}\right]} \qquad (2.2)$$

If two proteins have identical weighted distance maps, then $d(R,S) = 0$.

Now, in order to compare proteins with different chain lengths and unknown alignments, we define a score based upon sliding one map across another, similar to a correlation function. That is, if two proteins, R and S, have chain lengths M and N respectively where $M \leq N$, then we calculate a series of dissimilarities as follows:

$$d'(R,S,\tau) = \frac{\sum_{i=1}^{M}\sum_{j=i+2}^{M} \left| r_{i,j}^{-p} - s_{i+\tau,j+\tau}^{-p} \right|}{\frac{1}{2}\left[\sum_{i=1}^{M}\sum_{j=i+2}^{M} r_{ij}^{-p} + \sum_{i=1}^{M}\sum_{j=i+2}^{M} s_{i+\tau,j+\tau}^{-p}\right]} \tag{2.3}$$

where the range of "offsets", τ, of one weighted distance map relative to the other varies from $-M/2$ to $N - M/2$ for a total of N different alignments. The dissimilarity between the proteins R and S is then obtained by finding the offset for which the similarity is greatest:

$$d(R,S) = \min\left\{ d'(R,S,\tau) \right\} \tag{2.4}$$

This procedure alone, however, is not sufficient to specify a score, since the sliding of distance maps means that some (i,j) pairs of one conformation will sometimes go unpaired with (i',j') pairs in the other. For example, if $\tau = -5$, then the residue pair $(1,3)_R$ of protein R will be compared to a non-existent pair $(-4,-2)_S$ of protein S. Therefore, there are two additional steps in the scoring method. First, the weighted distance maps are made periodic (i.e., "wrapped around") so that residue pairs are defined for all offsets. In this way, the number of compared residue pairs is the same for all alignments. Second, since a "wrapped-around" weighted distance map may imply some structural features which are not present in the actual conformation (e.g., a helix can move from the N-terminus end of a conformation to the C-terminus end), a randomization procedure is used to ensure that the dissimilarity score and alignment do not contain artifacts from using periodic weighted distance maps. The randomization procedure is as follows: For any comparison of residue pairs, $(i,j)_R$, $(i',j')_S$, involving a wrapped-around residue pair, a difference weight is not added directly to the dissimilarity score. Instead, these weights are collected in separate bins based on contact order (the contact order for residue pair

(i,j) is defined as $|j-i|$, i.e., the separation of the residues along the chain). The binned distance weights from one conformation are then randomly matched with the binned distance weights from the other conformation and then added to the dissimilarity score. In practice, the offset which gives rise to the best alignment of two proteins contain few references to non-existent (i.e., wrapped-around) residue pairs and several different methods that we tried for treating them gave similar results.

When comparing a larger protein with a smaller one, CONGENEAL finds the part of the large protein that is most similar to the weighted distance map of the smaller protein. This feature makes CONGENEAL useful for rapidly finding specific substructures within different proteins in a structural database. To search a database for a specific substructure, CONGENEAL is used to generate scores between the substructure and each protein in the database: a small score for some alignment with a given protein locates that motif within the protein.

Table 2.1

Key to 158 protein set

code	# residues	protein name
451c	82	cytochrome c_{551}
155c	134	cytochrome c_{550}
256b	106	cytochrome b_{562}
1aat	288	aspartate aminotranferase
1abp	306	L-arabinose binding protein
2abx	74	α-bungarotoxin
2act	218	actinidin
1acx	107	actinoxanthin
6adh_a	374	alcohol dehydrogenase
3adk	194	adenylate kinase
2ait	74	tendamistat
1alc	122	α-lactalbumin
2alp	198	α-lytic protease
4ape	330	endothiapepsin
7api	339	α_1-antitrypsin
3app	323	penicillopepsin
2apr	325	rhizopuspepsin

2atc_c	305	aspartate transcarbamylase (regulatory subunit)
2atc_r	152	aspartate transcarbamylase (catalytic subunit)
2aza_a	129	azurin (*A. denitrificans*)
3b5c	85	cytochrome b_5
1bds	43	sea anemone antiviral protein
3blm	257	β-lactamase
1bp2	123	phospholipase A_2
3c2c	112	cytochrome c_2
2ca2	256	carbonic anhydrase
8cat_a	498	beef liver catalase
1cbp	86	cucumber basic protein
1cc5	83	cytochrome c_5
1ccr	111	rice cytochrome c
2ccy_a	127	cytochrome c'
2cdv	107	cytochrome c_3
2ci2	65	barley chymotrypsin inhibitor
3cln	143	calmodulin
1cms	323	chymosin B
2cna	237	concanavalin A
5cpa	307	carboxypeptidase A
2cpp	405	cytochrome P450 CAM
5cpv	108	carp parvalbumin
1crn	46	crambin
1cro_o	66	λ cro
2cro	65	434 cro
1cse	274	subtilisin carlsberg
1ctf	68	C-terminal domain of ribosomal protein L7/L12
1ctx	71	α-cobratoxin
5cyt	103	tuna cytochrome c
2cyp	293	cytochrome c peroxidase
3dfr	162	dihydrofolate reductase
5ebx	62	erabutoxin A
1ecd	136	erythrocruorin
1efm	130	elongation factor TU
2enl	436	enolase
2est	240	porcine elastase
1etu	141	elongation factor TU
2fb4_h	229	F_{ab} KOL (heavy chain)
2fb4_l	216	F_{ab} KOL (light chain)
1fc1	206	F_c fragment of immunoglobulin
4fd1	106	ferredoxin

3fxn	138	flavodoxin
3gap_c	208	catabolite gene activator protein (closed form)
3gap_o	205	catabolite gene activator protein (open form)
2gbp	309	galactose-binding protein
1gcr	174	γ-crystallin
1gd1_o	334	glyceraldehyde-3-phosphate dehydrogenase
2gls_a	468	glutamine synthetase
1gp1_a	184	glutathione peroxidase
3grs	461	glutathione reductase
1hho_a	141	human hemoglobin
1hip	85	high potential iron protein
1hkg	457	hexokinase
2hla_h	270	human class 1 histocompatibility complex (heavy)
2hla_m	99	human class 1 histocompatibility complex (β-2-microglubulin)
2hmg_1	328	influenza hemeagglutinin (HA1)
2hmg_2	175	influenza hemeagglutinin (HA2)
1hmq_a	113	hemerythin
1hoe	74	α-amylase inhibitor
3hvp	99	HIV protease
2i1b	153	interleukin-1β
3icb	75	intestinal calcium-binding protein
4ins_a	21	2Zn insulin
1kga	173	2-keto-3-deoxy-6-phosphogluconate aldolase
2lbp	346	leucine-binding protein
3ldh	329	dogfish lactate dehydrogenase
2lh4	153	lupin leghemoglobin
2liv	344	leucine/isoleucine/valine-binding protein
1lrd	87	λ-repressor
1lyz	129	hen egg white lysozyme
1lz1	130	human lysozyme
3lzm	164	T4 lysozyme
1mbd	153	sperm whale myoglobin
4mdh	334	malate dehydrogenase
2mev_vp1	268	mengo virus VP1
2mev_vp2	249	mengo virus VP2
2mev_vp3	231	mengo virus VP3
4mlt	26	mellitin
1mon_a	44	monellin (A chain)
1nxb	62	neurotoxin B
2ovo	56	ovomucoid, third domain
2pab	114	prealbumin

9pap	212	papain
2paz	123	pseudoazurin
1pcy	99	plastocyanin
4pep	326	pepsin
1pfk_c	320	phosphofructokinase (closed form)
1pfk_o	320	phosphofructokinase (open form)
3pgk	416	phosphoglycerate kinase
3pgm	230	phosphoglycerate mutase
1phh	394	p-hydroxybenzoate hydroxylase
1phy	126	photoreactive yellow protein
2pka	232	kallikrein A
2plv_vp1	283	polio virus VP1
2plv_vp2	268	polio virus VP2
2plv_vp3	235	polio virus VP3
1pp2_r	122	snake venom phospholipase
1ppt	36	avian pancreatic polypeptide
1prc_c	332	photosynthetic reaction center *R. viridis* C subunit
1prc_l	273	photosynthetic reaction center *R. viridis* L subunit
1prc_m	323	photosynthetic reaction center *R. viridis* M subunit
1prc_h	258	photosynthetic reaction center *R. viridis* H subunit
2prk	279	proteinase K
1pte	348	carboxypeptidase/transpeptidase
5pti	58	bovine pancreatic trypsin inhibitor
4ptp	223	trypsin
1pyp	280	pyrophosphatase
1r69	63	434 repressor (N-terminal domain)
1rbb_a	124	ribonuclease B
1rei	107	immunoglobulin V_κ domain
1rhd	293	rhodanese
2rhe	114	immunoglobulin V_λ domain
4rhv_vp1	273	rhinovirus VP1
4rhv_vp2	255	rhinovirus VP2
4rhv_vp3	236	rhinovirus VP3
3rn3	124	ribonuclease A
1rns	72	ribonuclease S
2rnt	104	ribonuclease T_1
3rp2	224	rat mast cell protease
7rsa	124	ribonuclease A
5rub_a	260	rubisco
5rxn	54	rubredoxin
4sbv_a	199	southern bean mosaic virus

2sga	181	*S. griseus* proteinase A
3sgb	185	*S. griseus* proteinase B
1sn3	65	scorpion neurotoxin
2sns	141	staphylococcal nuclease
2sod_o	151	superoxide dismutase (orange subunit)
1srx	108	thioredoxin
2ssi	107	streptomyces subtilisin inhibitor
2stv	184	satellite tobacco necrosis virus
2taa	478	taka-amylase
2tbv	286	tomato bushy stunt virus
1tec	279	thermitase
1thi	207	thaumatin I
1tim	247	chicken triosephosphate isomerase
2tmv	154	tobacco mosaic virus
4tnc	160	troponin C
1tnf	152	tumour necrosis factor
1ubq	76	ubiquitin
1utg	70	uteroglobin
9wga	171	wheat germ agglutinin
1wrp	102	*trp* repressor
4xia	393	xylose isomerase
2yhx	457	hexokinase

Validation of the Dissimilarity Measure

How can one validate a dissimilarity measure? For any two proteins, different measures can predict different degrees of relatedness. As noted before, there is no fundamentally correct measure of relatedness. Therefore, the validation of a dissimilarity measure ultimately depends on whether it seems sensible in light of other knowledge. In the section below, we characterize CONGENEAL in the following ways:

Pairwise tests. *(a) Finding sequence alignments:* When two different proteins contain the same substructure, a dissimilarity measure should find the sequence alignment for which the structures most closely superimpose. *(b) Using a probe protein structure to*

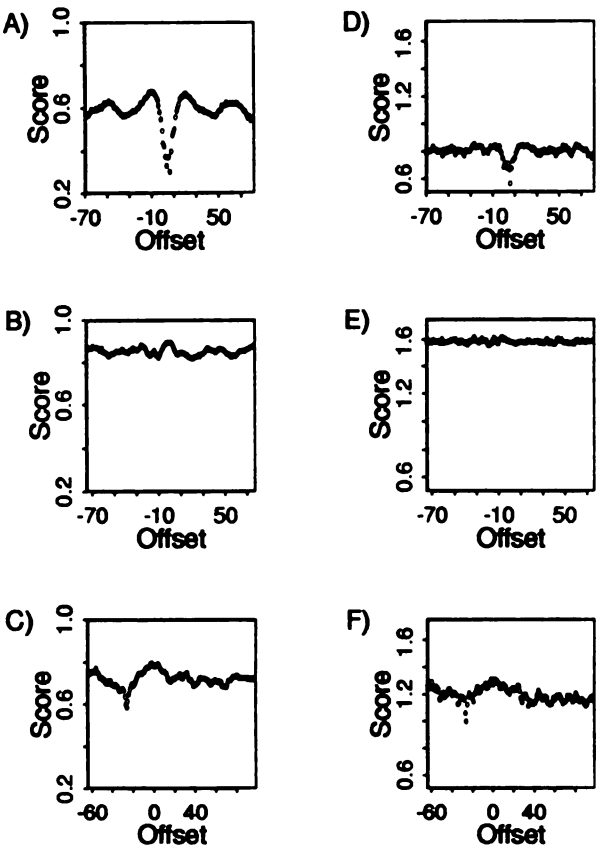


Figure 2.2: Plots of dissimilarity score versus alignment. For figures A-C, $p = 2$. For figures D-F, $p = 6$. Figures A & D show the comparison of sperm whale myoglobin with human hemoglobin. B & E show the comparison of two unrelated proteins: sperm whale myoglobin and superoxide dismutase. C & F show the comparison of two weakly similar proteins: T4 (bacteriophage) lysozyme and hen egg white lysozyme.

search the database: A dissimilarity measure should find related proteins or substructures in a search of a structural database.

Cluster Analysis. We compare 158 proteins pairwise and apply clustering algorithms to ask whether the dissimilarity measure finds sensible family relationships among them. We use two types of clustering methods: minimal spanning trees and hierarchical trees based on agglomerative clustering.

All protein coordinates were obtained from the Brookhaven Protein Data Bank (PDB) [12, 13]. The set of 158 protein structures was derived from Appendix 3 in "Protein Architecture" by Arthur Lesk [14]. Table 2.1 lists the proteins and their PDB filenames.

Pairwise tests: (a) Finding Sequence Alignments.

To first choose a few examples, it is reasonable to believe that sperm whale myoglobin (1mbd) and the A chain of human hemoglobin (1hho_a) are closely related proteins; that sperm whale myoglobin and the orange subunit of superoxide dismutase (2sod) are unrelated; and that the lysozymes from T4 bacteriophage (3lzm) and from hen egg white (1lyz) are only distantly related. Figure 2.2 shows the dissimilarity score as a function of alignment for these three comparisons using CONGENEAL with either $p = 2$ or $p = 6$. The point at which the score dips to a minimum, (i) indicates the degree of similarity between the two proteins, and (ii) gives the offset (i.e., shift) of one sequence starting position relative to the other sequence for which the structures bear closest resemblance.

For the three pairwise protein comparisons mentioned above, CONGENEAL finds the expected relationships. Sperm whale myoglobin is found to be similar to the A chain of human hemoglobin with an offset of -6 residues. On the other hand, figure 2.2B indicates that there is no similarity between sperm whale myoglobin and the orange subunit of superoxide dismutase. The dissimilarity measure finds hen egg white

lysozyme and T4 bacteriophage lysozyme to have only a small degree of similarity. In this case, the best score is obtained at an offset of -26 residues, in agreement with the observations of Remington and Matthews [5] and Rossmann and Argos [15], who noted that when residues 1-80 of the phage lysozyme are aligned with residues 27 - 106 of the hen egg white lysozyme, there is overlap of the active sites.

(b) Using a probe to search the database.

When a probe protein or substructure is scanned through a protein databank, a dissimilarity measure should properly rank order them by their similarity to the probe. Below we show three examples -- the helix-turn-helix DNA binding motif, the EF hand calcium binding motif, and the globin fold -- for which the dissimilarity score identifies closely related protein conformations.

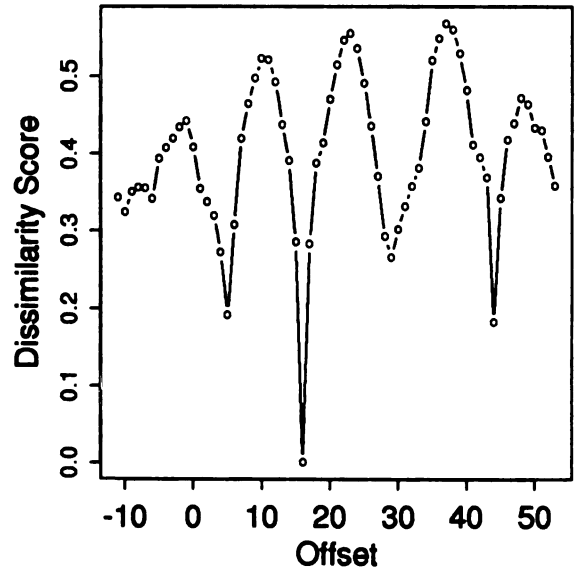


Figure 2.3: Dissimilarity score versus alignment of the 434 cro DNA binding substructure with the complete structure of 434 cro.

Table 2.2

Proteins with Helix-Turn-Helix motif

Protein Name	PDB filename	CONGENEAL	RMS
434 Cro	2cro	0.000	0.000
434 repressor (N-terminal domain)	1r69	0.052	0.380
λ cro	1cro_o	0.068	0.585
λ repressor	1lrd	0.099	0.830
enolase	2enl	0.110	1.634
catabolite gene activator protein (open form)	3gap_o	0.120	1.135
catabolite gene activator protein (closed form)	3gap_c	0.126	1.068
cytochrome P450 CAM	2cpp	0.135	1.742
C-terminal domain of ribosomal protein L7/L12	1ctf	0.170	1.938
xylose isomerase	4xia	0.177	2.047
cytochrome c peroxidase	2cyp	0.178	2.053
photosynthetic reaction center *R. viridis* M subunit	1prc_m	0.178	2.963
photosynthetic reaction center *R. viridis* L subunit	1prc_l	0.189	2.835
trp repressor	1wrp	0.197	1.729
beef liver catalase	8cat_a	0.203	2.652
erythrocruorin	1ecd	0.205	2.978
proteinase K	2prk	0.205	2.543
glutamine synthetase	2gls_a	0.207	3.694
hemerythin	1hmq_a	0.209	4.206
sperm whale myoglobin	1mbd	0.210	4.531

(1) DNA Binding Motif: A number of proteins are known to have similar helix-turn-helix substructures that bind DNA. λ Cro, λ repressor, 434 cro, 434 repressor, *trp* repressor, and catabolite gene activator protein (CAP) all have sequence similarity in a region of 22 amino acids, corresponding to the helix-turn-helix structural motif [16]. How widely distributed is the helix-turn-helix motif throughout the protein database? We use CONGENEAL to search the dataset for the helix-turn-helix conformation. In our search, the helix-turn-helix substructure is defined as the 23-residue stretch from 434 Cro starting with methionine 15 and ending with glycine 37. As a simple test, figure 2.3 shows the result of aligning the 434 Cro helix-turn-helix substructure with the full 434 Cro protein. The deepest minimum in figure 2.3 correctly identifies the proper alignment

with itself, and the score of 0 indicates that it is an exact match. The other three minima correspond to the three other turns between helices found in Cro.

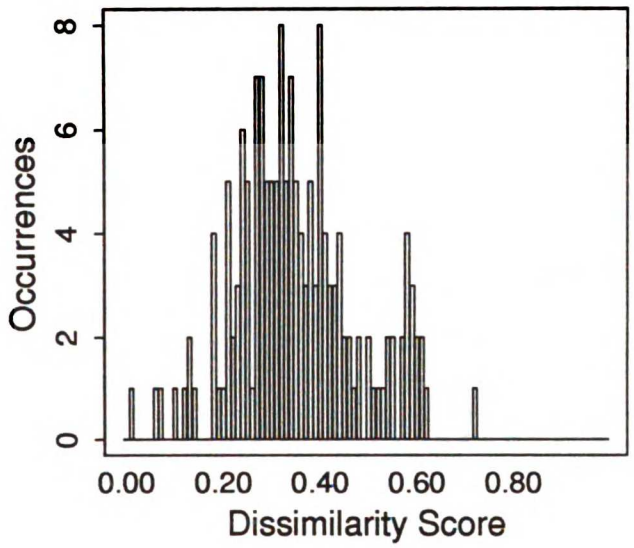


figure 2.4: Histogram showing the distribution of dissimilarity scores when the 434 cro DNA binding substructure is compared with a dataset of 158 proteins.

The 434 Cro helix-turn-helix DNA binding substructure was then scanned across the dataset of 158 proteins. The distribution of dissimilarity scores is shown in figure 2.4. Several proteins are found to have helix-turn-helix substructures similar to that of 434 Cro. Table 2.2 lists the twenty proteins with the greatest similarities to the target substructure. All seven proteins known to have the DNA binding helix-turn-helix substructure are in this group. In all seven cases, the predicted alignment of the substructure with the protein is identical to the alignment produced by sequence analysis [16]. The protein with the most similar substructure (other than 434 Cro) is 434 repressor. This is consistent with the observation of Mondragon et al. that the amino terminal domain of 434 repressor is remarkably similar to 434 Cro [17] and that the

substructures are virtually identical. The DNA binding protein that is least similar to 434 Cro is *E. coli trp* repressor. The latter differs from the other DNA binding proteins in two respects: (i) the end of the first helix is more open, and (ii) the orientation of the second helix in the helix-turn-helix substructure is constrained by the binding of L-tryptophan [18].

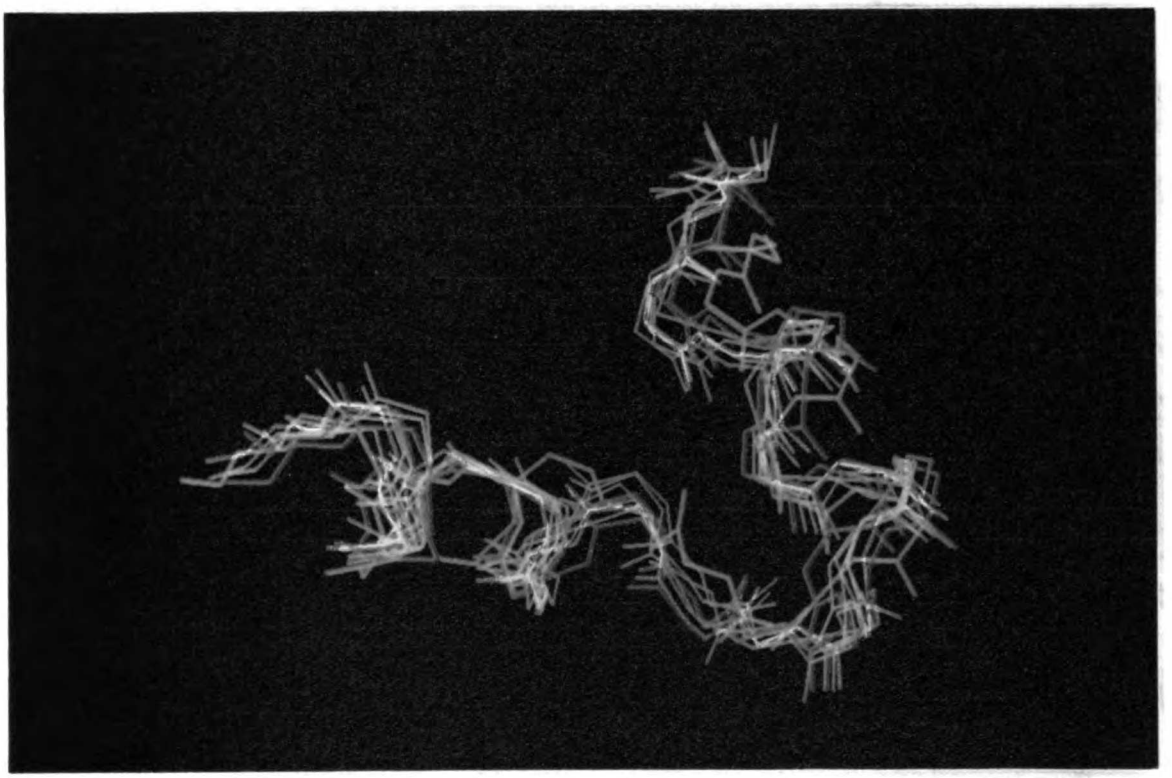


figure 2.5: Structural alignment of 434 Cro DNA binding motif to top 8 matches identified by CONGENEAL and *trp* repressor which scored 14th. 434 Cro DNA binding substructure is shown in green. DNA binding proteins are shown in red. Non-DNA binding proteins are shown in blue. *trp* repressor, the DNA binding protein least similar to the target 434 Cro substructure, is shown in cyan.

The seven DNA binding proteins rank 1, 2, 3, 4, 6, 7, and 14 in similarity to the probe helix-turn-helix structural motif. Some non-DNA binding proteins also score well for the presence of the helix-turn-helix substructure (see table 2.2). In many cases, the best match of the substructure to a protein occurs when the helix-turn-helix substructure is aligned with the last half of a long helix, a turn, and the first few residues of the following helix. Two of the proteins identified here as having a substructure similar to that of the probe substructure have been previously noted by Richardson and Richardson [19]. They found that cytochrome c peroxidase and ribosomal L7/L12 protein contain conformations similar to the DNA-binding helix pairs in gene activator and repressor proteins. In the present analysis, the non-DNA binding protein which has a substructure most similar to the probe DNA binding substructure is yeast enolase. Enolase has two domains consisting of (i) a 3 stranded β meander and 4 α- helices and (ii) an 8-fold $\beta + \alpha$ barrel [20]. The helix-turn-helix substructure of 434 Cro aligns with enolase near the end of the N-terminal domain. Figure 2.5 shows the top eight alignments found by CONGENEAL for substructures from DNA binding proteins or non-DNA binding proteins with the DNA binding substructure of 434 Cro (see also kinemage 1).

(2) Calcium Binding: the EF Hand: Another well characterized substructural motif is the EF hand calcium binding conformation, first described by Kretsinger and Nockolds [21] from carp muscle calcium-binding parvalbumin (5cpv). The EF hand is also found in several other proteins that bind calcium, including calmodulin (3cln), troponin C (4tnc), and intestinal calcium binding protein (3icb). Although CONGENEAL finds relatively few substructures identical to EF hands, many proteins contain substructures that are fairly similar.

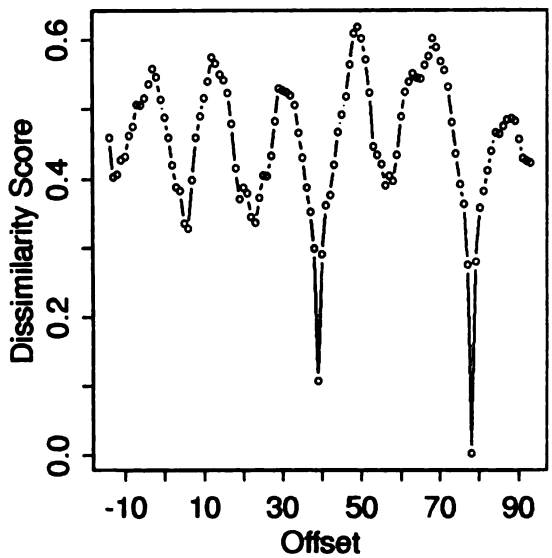


Figure 2.6: Dissimilarity score versus alignment of the carp parvalbumin EF hand substructure with the complete structure of carp parvalbumin.

We define the EF hand substructure in carp parvalbumin as the 29 residues from asparagine 79 to lysine 107. The E helix is 12 residues (79 - 90); the loop is 8 residues (91 - 98); and the F helix is 8 residues long (99 - 107). Figure 2.6 shows the result of aligning this substructure with the complete structure of carp parvalbumin. There are 5 minima, corresponding to the joining regions between the 6 helices of parvalbumin (labeled A - F): AB, BC, CD, DE, EF. Strong matches are found at two positions, corresponding to C- loop-D and E-loop-F. Both of these substructures are in the EF hand conformation. While Kretsinger and Nockolds suggested that A-loop-B is related to the EF hand, our results do not find significant structural similarity between A-loop-B and the EF hand substructure. In fact, the A and B helices are oriented nearly parallel to one another whereas the helices in an EF hand are nearly perpendicular.

We then use the EF hand substructure as a probe to search the dataset of 158 proteins. Figure 2.7 shows the distribution of dissimilarities. The four best scoring

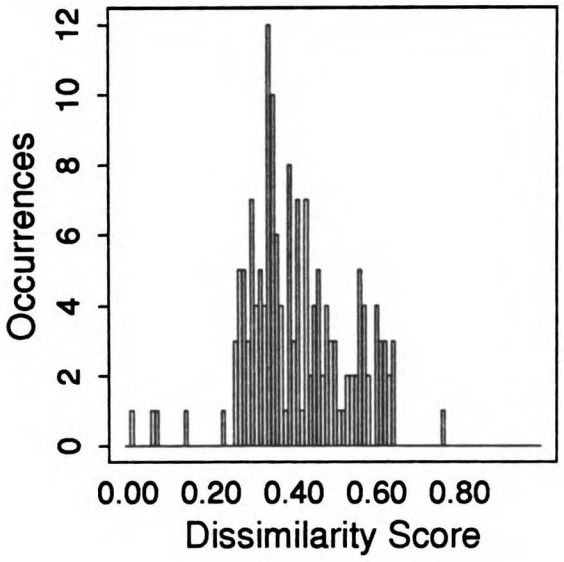


Figure 2.7: Histogram showing the distribution of dissimilarity scores when the EF hand substructure is compared with a dataset of 158 proteins.

proteins are all calcium binding proteins (see table 2.3). The EF hand in troponin C was found to be the most similar to the parvalbumin structure; the dissimilarity scores as a function of alignment are shown in figure 2.8. Minima identify the four EF hand substructures in troponin C. The two minima on the right in figure 2.8 indicate two substructures which are the most similar to the parvalbumin substructure and correspond to the EF hands nearest the C- terminal end of the protein. The two minima on the left identify two N-terminal EF hands which are less similar to the parvalbumin EF hand. Interestingly, the N-terminal EF hands do not bind calcium [22].

The method correctly identifies bovine intestinal calcium binding protein as being similar to the EF hand. On the other hand, given that carp parvalbumin and bovine intestinal calcium binding protein are related, the RMS deviation of C_α positions between their EF hands seems to be surprisingly large. The large RMS deviation arises

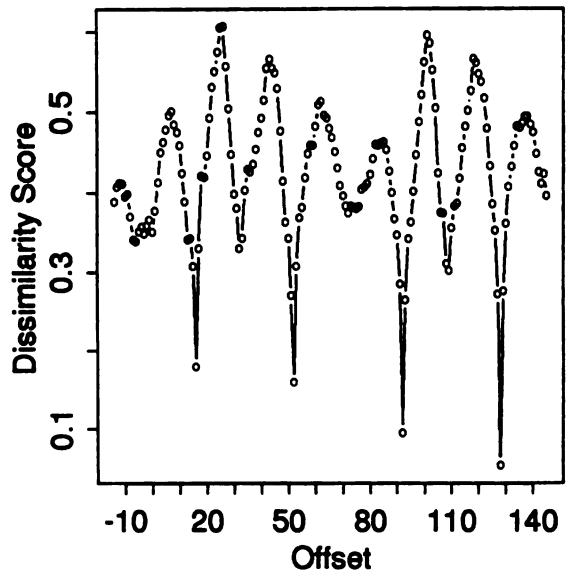


Figure 2.8: Dissimilarity score versus alignment of the EF hand substructure with troponin C. The 4 minima correspond to the 4 EF hand substructures in troponin C.

because the residues at both ends of the substructure have different conformations in the two proteins. The next most similar substructure to the EF hand is in T4 lysozyme (see kinemage 2); this similarity was first noted by Tufty and Kretsinger [23].

(3) Searching protein structures for functional subunits: The ability to perform fast searches for substructures within proteins allows for searching a database of protein structures for specific functional subunits. We show an example of using CONGENEAL to find possible calcium binding proteins in the dataset. In the EF hand substructure, the calcium is bound to residues within the loop region. Therefore, we search the dataset for the presence of the E helix, the F helix, and the loop. E and F are α-helices, so all proteins with α-helices score well for the presence of the E and F helix substructures (data not shown). Figure 2.9 shows the distribution of dissimilarities with the calcium binding loop. While most proteins are predicted to have at least one short loop

Table 2.3

Proteins with EF hand

Protein Name	PDB code	CONGENEAL	RMS
carp parvalbumin	5cpv	0.002	0.000
troponin C	4tnc	0.054	0.644
calmodulin	3cln	0.064	0.987
intestinal calcium-binding protein	3icb	0.135	2.868
T4 lysozyme	3lzm	0.227	2.994
sperm whale myoglobin	1mbd	0.251	5.199
human hemoglobin	1hho_a	0.253	5.143
Subtilisin carlsberg	1cse	0.255	4.107
cytochrome P450 CAM	2cpp	0.261	6.142
erythrocruorin	1ecd	0.264	4.991
lupin leghemoglobin	2lh4	0.265	4.612
hemerythin	1hmq_a	0.266	5.050
enolase	2enl	0.269	4.524
thermitase	1tec	0.270	4.159
cytochrome c peroxidase	2cyp	0.275	4.565
trp repressor	1wrp	0.277	4.272
cytochrome b_{562}	256b	0.277	4.474
proteinase K	2prk	0.280	5.069
leucine-binding protein	2lbp	0.282	3.518
malate dehydrogenase	4mdh	0.282	5.295

conformation similar to the calcium binding loop, five proteins are clearly distinct as being more similar than the other proteins. They include the four calcium binding proteins identified above. Hence, the weighted distance map for this loop region is a good identifier of the calcium binding motif. In addition, galactose-binding protein scores well for the presence of a calcium binding loop. Consistent with this finding, galactose-binding protein was reported to have a calcium binding site [24] which resembles the EF hand without the helices. T4 Lysozyme, which scored fifth for the presence of an EF hand scores 69th for the presence of the calcium binding loop. While T4 lysozyme has two helices similar in orientation to the EF hand helices in parvalbumin, the intervening loop is clearly not in the calcium binding conformation.

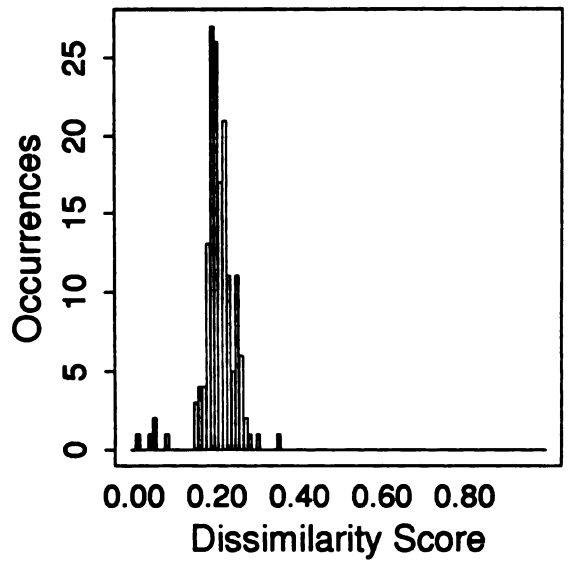


Figure 2.9: Histogram showing the distribution of dissimilarity scores when the calcium binding loop from carp parvalbumin is compared with a dataset of 158 proteins.

(4) Globins: Figure 2.10 shows the dissimilarities of sperm whale myoglobin (1mbd) to the set of 158 protein structures. The four most similar structures are all globins: sperm whale myoglobin (1mbd), erythrocruorin (1ecd), human hemoglobin (1hho_a), and leghemoglobin (2lh4). The next most similar proteins are all dominated by α-helices. They include uteroglobin (1utg), *trp* repressor (1wrp), calcium binding protein (3icb), and cytochrome b_{562} (256b). The proteins least similar to myoglobin are all β-sheet proteins: they include immunoglobulin fragments (2fb4_h, 2fb4_l, 1fc1), tumour necrosis factor (1tnf), monellin (1mon_a), and Cu,Zn superoxide dismutase (2sod). The dissimilarity distribution from CONGENEAL resembles an earlier comparison made by Bowie et al. [25] of sperm whale myoglobin versus a protein dataset based on their 3D profiling method. Their method shows the degree to which the *sequences* of other globins are compatible with the *structure* of sperm whale myoglobin. Our method shows the degree to which the *structures* of other globins are similar to the *structure* of sperm

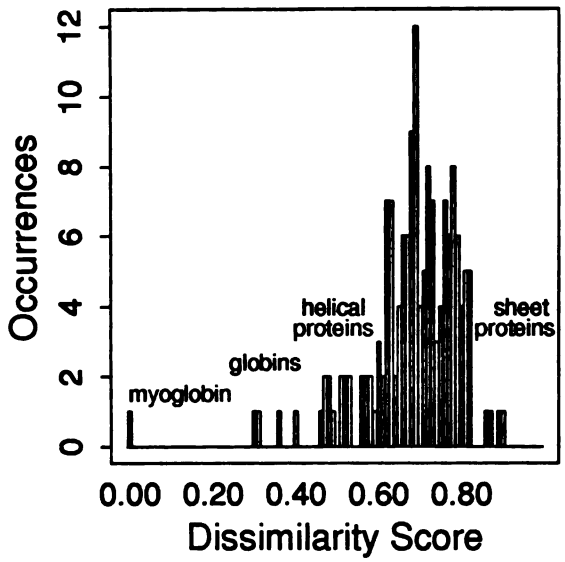


Figure 2.10: Histogram showing the distribution of dissimilarity scores when sperm whale myoglobin is compared with a dataset of 158 proteins.

whale myoglobin. At least for sperm whale myoglobin, the distributions from 3D profiling and CONGENEAL bear considerable resemblance.

Most closely related proteins

As another test, we compare the 158 proteins pairwise ($(N\times(N-1))/2 = 12{,}403$ tests), and ask which pairs are the most closely related. Of course, since this is not a "selected" set of unrelated conformations, some of these proteins are quite similar; these highly similar pairs are controls that we study here. Table 2.4 lists the 20 most similar protein pairs. Not surprisingly, several of these pairs represent the same protein in two different conformations. For example, six of the closest structural similarities are pairs of ribonucleases. The dataset contains four ribonuclease structures: two independently determined structures of ribonuclease A (3rn3, 7rsa), ribonuclease B (1rbb_a), and ribonuclease S (1rns). Each pair of structures involves only a small

structural variation. For example, ribonucleases A and B have identical amino acid sequences, but differ by a polysaccharide moiety which is attached to asparagine 34 of ribonuclease B.

Ribonuclease B is about as similar to ribonuclease A as the two ribonuclease A's are to each other. This result is consistent with the conclusion of Williams et al. [26] that the conformation of ribonuclease B is not significantly different than that of ribonuclease A. The small variability occurs mostly in the β-sheet regions.

CONGENEAL also finds the correct alignment of ribonuclease S with the other ribonucleases (i.e., at an offset of 21 residues). All the ribonuclease structures are quite similar to each other, but ribonuclease S is the least similar among them. The deviations of ribonuclease S relative to the other ribonuclease structures are attributable to the contacts formed by residues 21 - 23 with the rest of the protein. Since ribonuclease S is formed by cleavage of ribonuclease A between alanine 20 and serine 21, the conformations of residues 21 - 23 presumably readjust in response to the cleavage of the peptide bond between residues 20 and 21.

CONGENEAL finds other highly similar pairs. It finds rice cytochrome c to be very similar to tuna cytochrome c. The rice structure has 8 additional residues at the N-terminus and there are 43 substitutions in the other 103 residues. Despite these sequence differences, the structures are found to be nearly identical [27]. Other sets of proteins that are found to be highly similar by CONGENEAL include DNA binding proteins (434 cro, 434 repressor, λ-repressor), neurotoxins (erabutoxin A, neurotoxin B), immunoglobulins (F_{ab} KOL, immunoglobulin V_λ domain), and viral VP3 domains (rhinovirus VP3, polio virus VP3).

CONGENEAL has limitations. First, it does not treat insertions, deletions, or gaps. An example of this limitation is in the comparison of α/β barrels such as triose phosphate

isomerase (TIM). It has been suggested that all the known α/β barrels may have diverged from a common ancestor [28]. If so, and if the process of evolutionary divergence involves changing loop lengths while retaining secondary structural domains, then evolutionary "distance" requires a similarity measure that carries only weak penalties for changing lengths of loops between domains. While some similarity methods do this [6], CONGENEAL does not, and would therefore not be useful as a measure of evolutionary divergence by this mechanism. Hence, again we caution that different similarity measures will find different degrees of relatedness among proteins, and will find different family clusters, but there is no unique right way to do this. And we note that the approach taken in CONGENEAL, while it is disadvantageous for measuring evolutionary divergence by this mechanism, is advantageous for other purposes, since it is based on making no assumptions about mechanisms of how one conformation is caused to differ from another. Such a need arises in the comparison of conformations of a given sequence, in which case there are no gaps, insertions, or deletions, or in the comparison of very different conformations that may not be related by a known evolutionary mechanism, in which case we believe it may be often preferable to measure similarity with an algorithm having a minimum number of degrees of freedom.

Second, when comparing sets of proteins with different alignments and chain lengths, the dissimilarity measure is not a true distance metric. That is, as with many other similarity measures, the triangle inequality:

$$d(a,b) + d(b,c) \geq d(a,c) \tag{2.5}$$

can be violated. For example, in the pairwise comparison of a sheet (S), a helix (H), and a protein consisting of both a sheet and a helix (P):

Table 2.4

Most Closely Related Proteins in Dataset

Pair Number	Protein Name	PDB code	score
1	ribonuclease A ribonuclease A	7rsa 3rn3	0.014
2	phosphofructokinase (open form) phosphofructokinase (closed form)	1pfk_o 1pfk_c	0.035
3	rice cytochrome c tuna cytochrome c	1ccr 5cyt	0.052
4	ribonuclease A ribonuclease B	3rn3 1rbb_a	0.056
5	ribonuclease A ribonuclease B	7rsa 1rbb_a	0.057
6	434 cro 434 repressor (N-terminal domain)	2cro 1r69	0.072
7	erabutoxin A neurotoxin B	5ebx 1nxb	0.076
8	fab KOL (light chain) immunoglobulin V_λ domain	2fb4_l 2rhe	0.079
9	ribonuclease A ribonuclease S	3rn3 1rns	0.103
10	ribonuclease A ribonuclease S	7rsa 1rns	0.103
11	hexokinase hexokinase	2yhx 1hkg	0.114
12	ribonuclease B ribonuclease S	1rbb_a 1rns	0.116
13	CAP (closed form) CAP (open form)	3gap_c 3gap_o	0.125
14	tendamistat α-amylase inhibitor	2ait 1hoe	0.136
15	leucine-binding protein leu/ile/val-binding protein	2lbp 2liv	0.140
16	human lysozyme hen egg white lysozyme	1lz1 1lyz	0.222
17	rhinovirus VP3 polio virus VP3	4rhv_vp3 2plv_vp3	0.231
18	λ-repressor 434 repressor (N-terminal domain)	1lrd 1r69	0.237
19	λ-repressor 434 cro	1lrd 2cro	0.289
20	elongation factor TU elongation factor TU	1etu 1efm	0.277

$$d(S,P) = 0 \text{ and } d(P,H) = 0, \tag{2.6}$$

but

$$d(S,H) > 0$$

Third, as with other contact-map based approaches, CONGENEAL does not distinguish structures by their chiralities. A molecule is indistinguishable from its mirror image. For comparing molecules with consistent chiralities, such as two real proteins, this is not a limitation. For comparing a lattice model and a real protein, however, chiral errors will not be detected. In a most general way, CONGENEAL only attempts to characterize distances pertinent to non-local interactions. In this sense, right-handed and left-handed helices are similar. When it is important to distinguish them, CONGENEAL is not appropriate.

Protein clustering into families

CONGENEAL is a measure that computes the structural similarity between any two compact polymer conformations. We have shown a few tests indicating where it is sensibly consistent with other knowledge. We now use this measure to study how it divides proteins into families. We define a *family* as a set of structures that collectively share a high degree of similarity to one another. The concept of family carries the implication that there are relatively sharp boundaries between families. Given a measure of similarity, there are several different methods for identifying clustering. As with similarity measures, there are no right or wrong clustering methods. In order to determine whether the families obtained are sensitive to the choice of clustering method, we study the clustering of protein structures by two different methods: a minimal spanning tree and a hierarchical method. Different similarity measures and clustering methods can lead to different, but equally valid, divisions of proteins into families.

Clustering by minimal spanning trees

First, we construct a minimal spanning tree, which is a graph that provides one way to describe relatedness among proteins. Consider a graph in which each one of the N protein structures is represented by a node. Every possible pair of nodes is connected by an edge. Each edge is weighted by the dissimilarity score relating the two proteins. Hence, there are $(N\times(N-1))/2$ edges. A spanning tree is a subgraph in which there are only $N-1$ edges connecting the N vertices (proteins). A minimal spanning tree is a spanning tree in which the sum of the weights of the edges is as small as possible. Thus the only connectivity is among the most similar proteins. We construct a minimal spanning tree using Kruskal's algorithm [29], as follows. First, the pairwise scores are sorted from most similar to least similar. The tree is then constructed edge by edge. The first edge is defined as the protein pair with the lowest score (highest similarity). The second edge is chosen to be the next lowest score that does not lead to a cycle in the graph. If a cycle were formed, then there would be more than one path between two vertices, and at least N edges for N vertices, thus violating the criterion of a spanning tree. The process continues until there are $N-1$ edges.

Figure 2.11 shows the minimal spanning tree for the set of 158 protein structures based on the CONGENEAL dissimilarity measure. A tree is unique provided that no two edges have the same weight. Note that no meaning should be attributed to the edge lengths shown in the figure, because they are not drawn in proportion to their respective dissimilarity weights. An edge connecting two proteins implies structural similarity between the two proteins.

By this clustering method, proteins are found to collect around *hubs* which may be thought of as *consensus family structures* or structural paradigms. For example, monellin (1mon_a), uteroglobin (1utg), crambin (1crn), and λ repressor (1lrd) are all hubs. Many

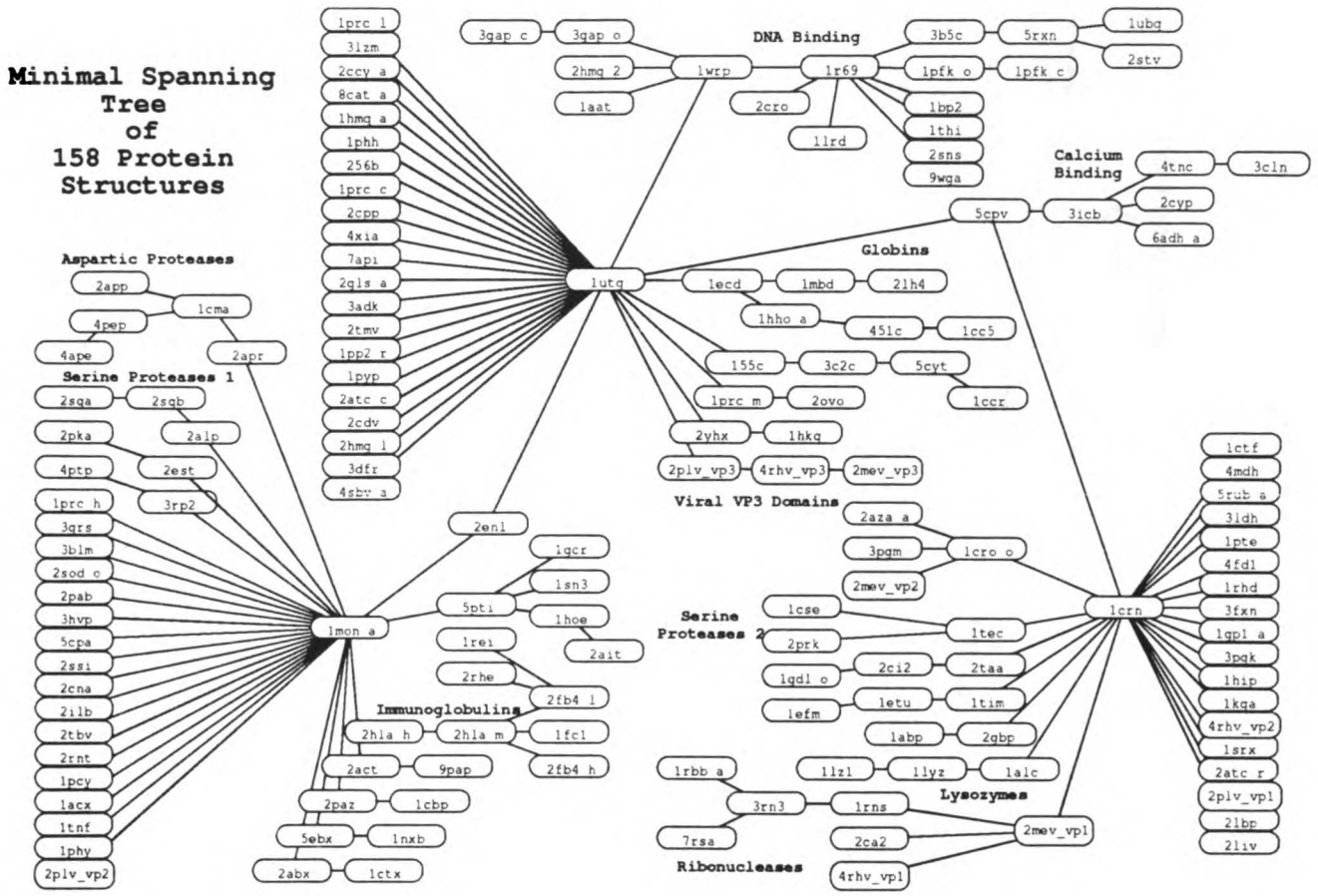


Figure 2.11: Minimal spanning tree of 158 protein dataset. Proteins are referenced by the codes listed in table 2.1. An edge connecting two proteins implies structural similarity between them. Edge lengths are not proportional to the dissimilarity between proteins. As a guide, the general location of some major family relationships are indicated in bold.

other proteins are connected to each hub. Each hub represents some characteristic topological feature (i.e., some specific protein fold). For example, the A chain of monellin forms three strands of an antiparallel β-sheet. Any protein in the dataset which has three strands in a similar conformation will score well when compared to monellin, and may be connected to the monellin hub. Similarly, uteroglobin, a progesterone-

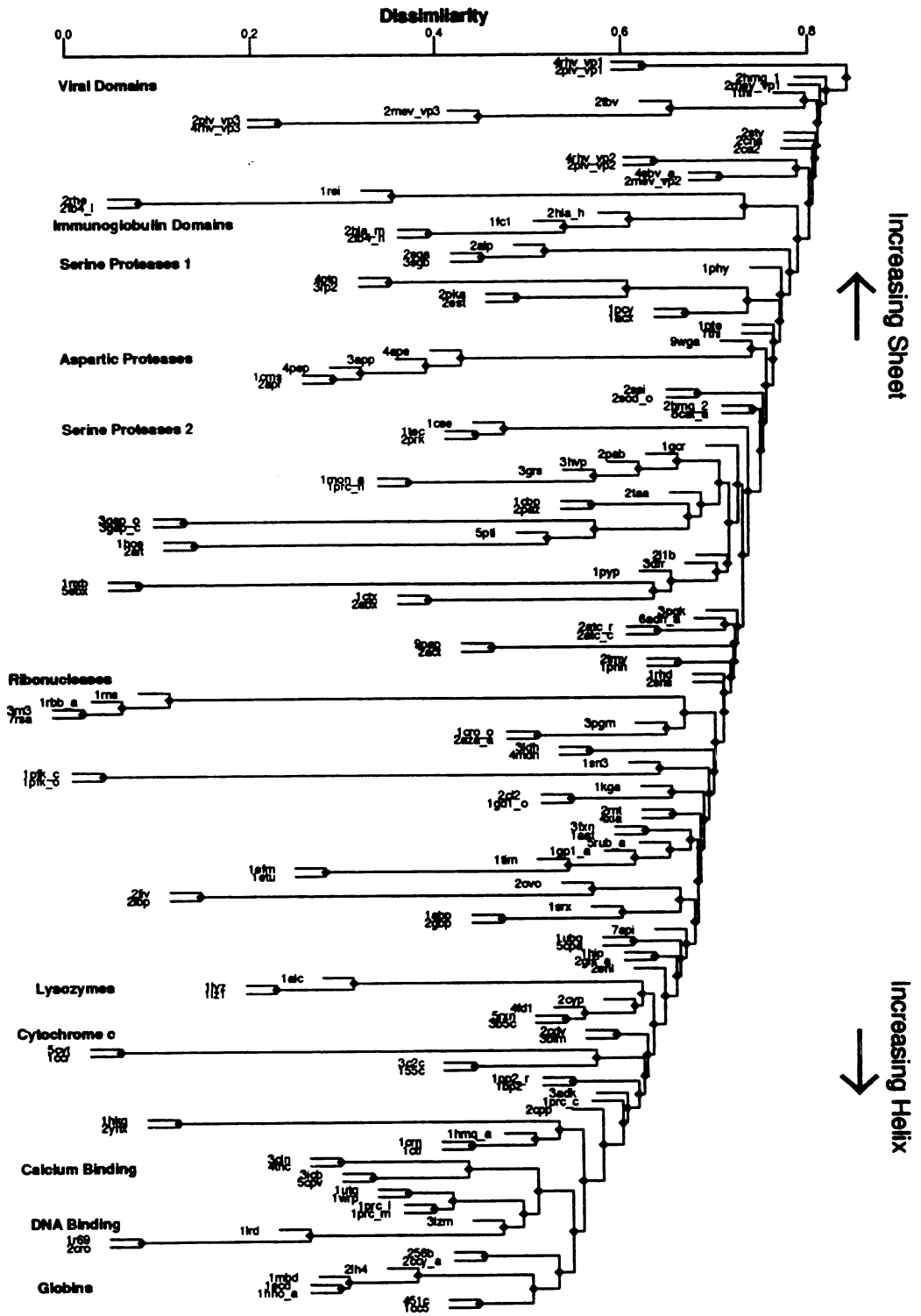


Figure 2.12: Hierarchical clustering of 158 proteins into a relatedness tree. The mean dissimilarities between the group members are indicated by diamonds at the branch points of the tree. Proteins are referenced by the codes listed in table 2.1. As a guide, some major family relationships are indicated.

binding protein consisting of 4 α-helices, is a hub for structures with similar helical and turn features. Crambin has both β-sheet and α-helix and serves as a hub for proteins with similar secondary structural features.

Proteins are found to cluster into families, often around hubs. For example, the globins cluster together. Lysozymes from hen egg white and humans cluster with α-lactalbumin. Other protein clusters include (i) viral VP3 domains, (ii) cytochrome c structures, (iii) immunoglobulin domains, (iv) aspartic proteases, (v) trypsin-like serine proteases and (vi) subtilisin-like serine proteases. Interestingly, T4 bacteriophage lysozyme is separated by 4 nodes from the other lysozymes. In this case, despite the structural similarity of the active sites, the remainder of the structure of T4 phage lysozyme is different from that of the other lysozymes.

Hierarchical clustering

In order to learn whether the family partitions found by CONGENEAL depend on the clustering method, we now consider a different clustering algorithm for collecting proteins into families. Here we use hierarchical clustering, which successively groups proteins into increasingly larger sets. At first, there are N proteins in N groups. Step 1 is to combine the two most similar proteins to form the first group; there are now $(N-2)$ single-protein groups and one 2-protein group. This is recorded as the first decision. Step 2 is to combine the two groups that now have the greatest similarity. To determine group similarities, all pairwise dissimilarities between groups are calculated; this generates $(M\times(M-1))/2$ average dissimilarities for M groups. The group dissimilarity is the average of the dissimilarities between the members of one group with respect to the members of another group. The merging process is repeated until all groups are combined into a single group. The merging process is a sequence of decisions that can be represented as a tree; see figure 2.12. Nodes at a given level in this tree represent a

given degree of dissimilarity.

As with the spanning tree, the hierarchical method finds that immunoglobulins, serine proteases, ribonucleases, globins, aspartic proteases, and viral VP domains form families. Hence the general division into these families appears to be relatively independent of the clustering method, although the details differ.

One interesting consequence of the hierarchical clustering is evident from figure 2.12. It leads to a partitioning of families in which sheet-structures are concentrated at the top of figure 2.12 and helix-structures are concentrated at the bottom. According to this partitioning, sheet structures are less related to one another than helical structures. Helical structures are more related to one another because of the regular pattern of close contacts formed by residues in the helical conformation.

Are proteins tightly clustered?

Are protein families tight or loose forms of organization? "Tight" organization means that any two proteins within a family are much more similar than any two proteins from different families. There would be a sharp boundary between protein families. "Loose" organization means that two proteins within a family may be only slightly more similar than two proteins taken from different families. The boundaries between protein families would not be sharp and might even overlap.

We assess tightness of protein families by studying the shape of the histogram of pairwise dissimilarities (figure 2.13). Qualitatively, if proteins are tightly clustered, the histogram in figure 2.13 would have 2 peaks: one representing the high similarities within families and the other representing the low similarities between families. The distribution of dissimilarities for the 158 protein dataset structures, however, shows mainly a single broad peak which indicates a wide range of relatedness among proteins.

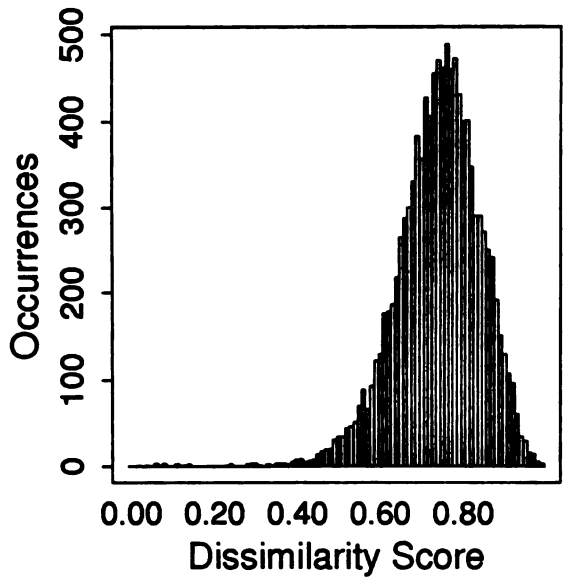


Figure 2.13: Histogram showing distribution of 12,403 pairwise comparisons of the 158 protein dataset.

There is only a very small peak indicating high similarities and tight families. The mean dissimilarity is 0.737, indicating that two arbitrary proteins are relatively unrelated. By this qualitative criterion, protein structural families are only loose entities.

A second way to assess the tightness of clustering draws on the analogy between (i) N proteins as points separated by their pairwise dissimilarities and (ii) a set of N points distributed in a d-dimensional space separated by their Euclidean distances. To pursue this analogy, we generated points in Euclidean spaces with varying degrees of clustering in several different dimensionalities, d. The distribution of distances between the points is compared to the distributions of pairwise protein dissimilarities shown in figure 2.13. We created varying degrees of clustering as follows. First, we assume there are f families of points in a d-dimensional space. We randomly generate f points which represent the family centers. Within each family, we then generate N/f points which are Gaussian-distributed around each family center. That is, the probability distribution for a

point x within a family is given by:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma k} \exp\left[\frac{-d(x,c)^2}{2\sigma^2 k^2}\right] \tag{2.7}$$

where c is a family center, $d(x,c)$ is the distance between x and c, k is the average distance between any two family centers, and σ is the parameter that controls the degree of clustering. When σ equals 1, the standard deviation of points around a family center is equal to the average distance between the family centers. As σ decreases, the tightness of the clustering increases. As an example, figure 2.14 shows scatter plots of points distributed in 2 dimensions around 15 "families" with three different values of σ.

After randomly generating points as described above, we calculate all the pairwise distances between the points within each set. Figure 2.15 shows histograms for $N = 200$, $d = 7$, $f = 25$, and varying σ. When σ is small (tight clustering) the histograms have two peaks, as expected. The leftmost peak is due to intrafamily distances and the rightmost peak is due to interfamily distances.

One method to compare the shapes of two distribution functions is the Quantile-Quantile (Q-Q) plot [30]. A Q-Q curve plots the sorted values of one distribution against the sorted values of a second distribution. If two distributions have the same shape, and differ only by a multiplicative factor that scales the width, or if they differ by a constant factor that shifts the mean values, then a Q-Q plot gives a straight line. Figure 2.16A shows a Q-Q plot of the histogram of figure 2.13 versus a non-clustered distribution of points in Euclidean space. The deviation at both extremes of the plot indicate that protein structures are more broadly distributed than would be predicted by the completely nonclustered distribution. Hence, a non-clustered uniform distribution of points is not a good model for the distribution protein dissimilarities.

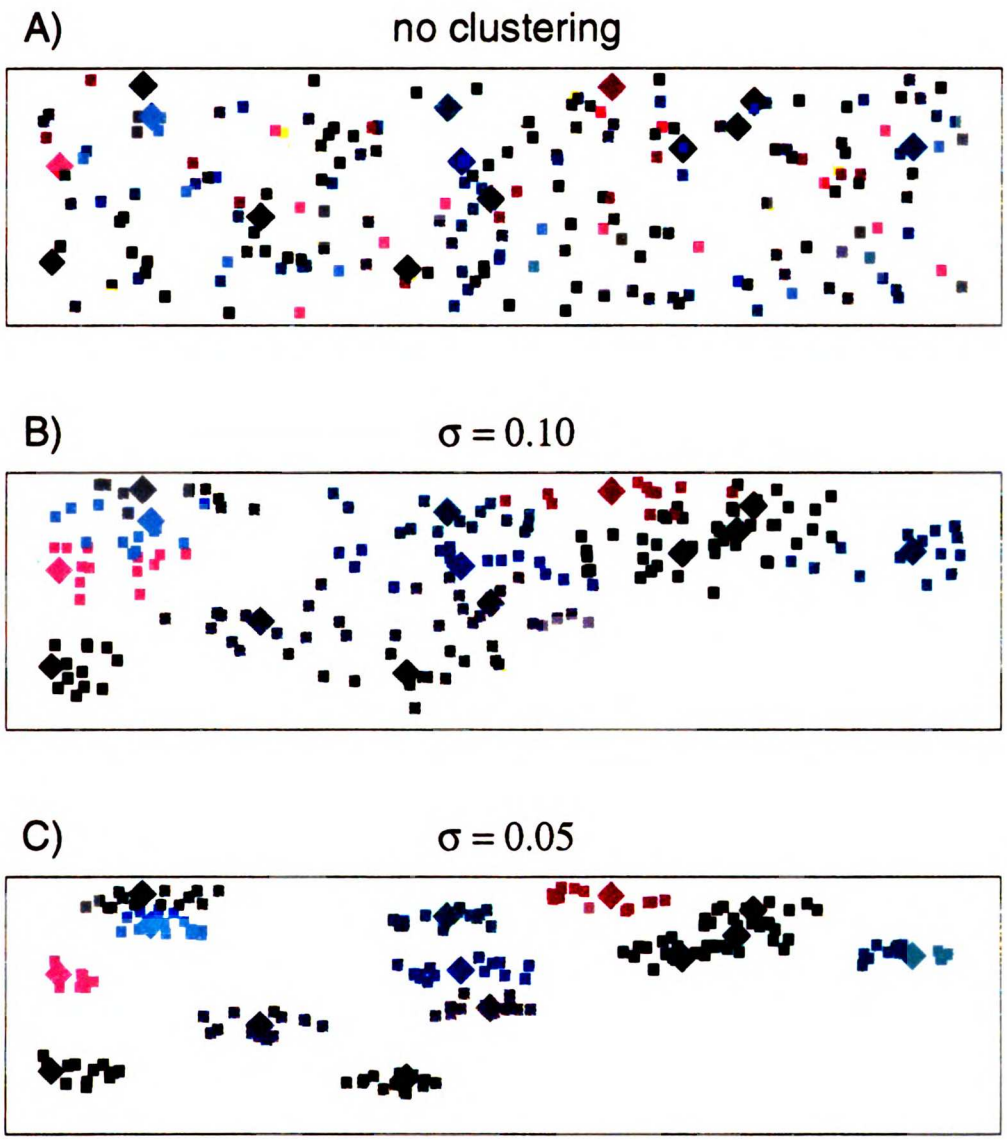


Figure 2.14: Scatter plots of points randomly distributed around family centers. Family centers are represented by large diamonds. Points associated with a family center are represented as small squares. In A-C, the number of family centers is 15 and the total number of points is 200.

On the other hand, Figure 2.16C shows that proteins are also not well represented as being very tightly clustered ($\sigma \leq 0.05$). When the distributions are tightly clustered, the Euclidean distribution underestimates the number of similar protein pairs in the range of

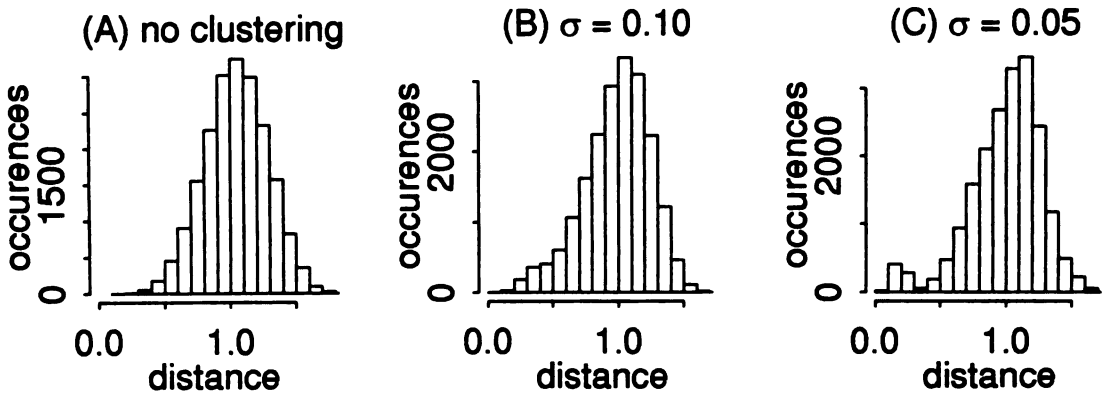


Figure 2.15: Distribution of pairwise distances between points randomly distributed between 25 families within a 7-dimensional sphere.

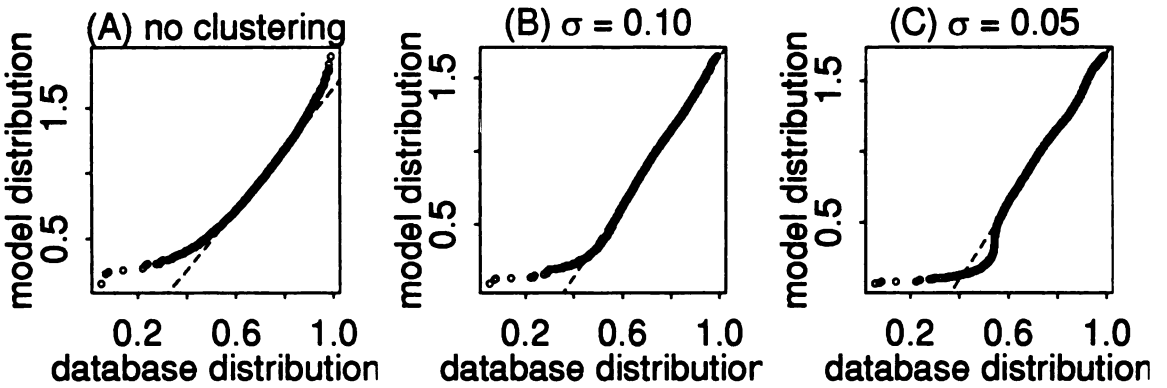


Figure 2.16: Q-Q plots comparing protein dataset pairwise dissimilarity distribution with distributions of random point pairwise distances. For the random point distributions, $f = 25$ and $d = 7$.

dissimilarites between 0.40 and 0.60, and overestimates them for distances less than 0.40. The closest correspondence between the Euclidean distances and protein similarities is obtained for values around $\sigma = 0.10$. This is the case for which the Q-Q plot is most linear; see figure 2.16B. While this value of σ implies that family members are

considerably closer together than are interfamily centers, protein families are not sufficiently tightly knit to avoid considerable overlap between families and individuals cannot be unambiguously assigned to families. This degree of clustering is indicated schematically in figure 2.14B for a 2-dimensional Euclidean space. By systematically varying the clustering parameter, σ, and the dimensionality, d, we found the optimal dimensionality to be about $d = 7$. It is not clear to us if this dimensionality in the Euclidean space analogy has any physical meaning since our dissimilarity measure is not a true metric. This comparison should be viewed simply as an analogy.

Our results are consistent with those of Rackovsky [8]. His similarity measure, which is based on local conformational preferences, also orders proteins from helices to sheets and finds families to be only loosely-knit entities.

Conclusions

We have described a simple quantity for characterizing the structural similarity between any two compact polymer or protein conformations. Based on differences between weighted distance maps, it requires no alignments or gap penalties and makes few assumptions or arbitrary decisions about polymer structure. It is computationally fast. The only parameter is the exponent p in the distance-dependence of the weights. The results are not very sensitive to this parameter.

The method can compare any two conformations, no matter how similar or different, and does not require identical chain lengths. It is intended for the purpose of comparing diverse polymer conformations. Several tests show that the relatedness among proteins reported by this measure are sensibly consistent with existing knowledge. This method can be used to rapidly search through a database of protein structures to find specified substructures, and to seek given functional components in other proteins. For

example, it searches the protein dataset in minutes to find possible calcium-binding motifs similar to the EF hand.

Such a similarity measure can be used to test algorithms of protein folding, for which generated conformations may be distant from the native structure. Hence, the CONGENEAL measure can serve as a sort of "reaction coordinate" for nativeness. For such problems, gaps are unimportant.

We combine this measure with two different clustering methods to identify protein families. We then ask how tightly clustered are families by drawing an analogy with points distributed in Euclidean space. The analogy indicates that protein families are only loosely-knit entities, and that individual proteins may often not be unambiguously assignable to a unique family.

Acknowledgements

We thank Sarina Bromberg, Fred Cohen, and Chris Carreras for many helpful discussions, and the DARPA University Research Initiative program and the NIH for financial support. Molecular graphics images were produced using the MidasPlus software system from the Computer Graphics Laboratory, University of California, San Francisco.

Supplementary Material Available with Reprints

This work would not have been possible without the enormous amount of effort required to experimentally determine the protein structures used in this analysis. References for all the structures obtained from the PDB are included as appendix A of this thesis.

References

[1] J. S. Richardson, "The Anatomy and Taxonomy of Protein Structure," *Advances in Protein Chemistry*, 34:167 - 339, 1981.

[2] J. S. Richardson and D. C. Richardson, "Principles and Patterns of Protein Conformation," in: *Prediction of Protein Structure and the Principles of Protein Conformation*, G. D. Fasman, ed., p. 1 - 98, New York, 1989.

[3] M. Levitt and C. Chothia, "Structural Patterns in Globular Proteins," *Nature*, 261:552 - 558, 1976.

[4] C. Chothia and A. V. Finkelstein, "The Classification and Origins of Protein Folding Patterns," *Annual Review of Biochemistry*, 59:1007-1039, 1990.

[5] S. J. Remington and B. W. Matthews, *Proceeding of the National Academy of Science*, 75(5):2180 - 2184, 1978.

[6] W. R. Taylor and C. A. Orengo, "Protein Structure Alignment," *Journal of Molecular Biology*, 208:1 - 22, 1989.

[7] A. Sali and T. L. Blundell, "Definition of General Topological Equivalence in Protein Structures," *Journal of Molecular Biology*, 212:403-428, 1990.

[8] S. Rackovsky, "Quantitative Organization of the Known Protein X-Ray Structures. I. Methods and Short-Length Scale Results," *Proteins*, 7:378 - 402, 1990.

[9] G. M. Maggiora and M. A. Johnson, "Introduction of Similarity in Chemistry," in: *Concepts and Applications of Molecular Similarity*, G. M. Maggiora and M. A. Johnson, ed., p. 1 - 13, New York, 1990.

[10] Y. Beers, *Introduction to the Theory of Error*, Addison-Wesley, 1957.

[11] A. Liljas and M. G. Rossmann, "Recognition of Structural Domains in Globular Proteins," *Journal of Molecular Biology*, 85:177-181, 1974.

[12] F. C. Bernstein, T. F. Koetzle, G. Williams, E. Meyer, M. Brice, J. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures," *Journal of Molecular Biology*, 112:535-542, 1977.

[13] E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng, in: *Crystallographic Databases - Information Content, Software Systems, Scientific Applications*, F. H. Allen, G.Bergerhoff, and R. Seivers, ed., p. 107 - 132, Data Commission of the Int'l Union of Crystallography, Bonn/Cambridge/Chester, 1987.

[14] A. M. Lesk, *Protein Architecture*, (*Practical Approach Series*), IRL Press, Oxford New York Tokyo, 1991.

[15] M. G. Rossmann and P. Argos, "Exploring Structural Homology of Proteins," *Journal of Molecular Biology*, 105:75 - 95, 1976.

[16] D. H. Ohlendorf, W. F. Anderson, and B. W. Matthews, "Many Gene-Regulatory Proteins Appear to Have a Similar α-helical Fold That Binds DNA and Evolved from a Common Precursor," *Journal of Molecular Evolution*, 19:109 - 114, 1983.

[17] A. Mondragon, S. Subbiah, S. C. Almo, M. Drottar, and S. C. Harrison, "Structure of the Amino-terminal Domain of Phage 434 Repressor at 2.0A Resolution," *Journal of Molecular Biology*, 205:189 - 200, 1989.

[18] R. W. Schevitz, Z. Otwinowski, A. Joachimiak, C. L. Lawson, and P. B. Sigler, "The Three-dimensional Structure of *trp* Repressor," *Nature*, 317:782 - 786, 1985.

[19] J. S. Richardson and D. C. Richardson, "Helix Lap-Joints as Ion-Binding Sites: DNA-Binding Motifs and Ca-Binding "EF Hands" are Related by Charge and Sequence Reversal," *Proteins*, 4:229 - 239, 1988.

[20] L. Lebioda, B. Stec, and J. M. Brewer, "The Structure of Yeast Enolase at 2.25-A Resolution," *Journal of Biological Chemistry*, 264(7):3685 - 3693, 1989.

[21] R. H. Kretsinger and C. E. Nockolds, "Carp Muscle Calcium Binding Protein II. Structure Determination and General Description," *Journal of Biological Chemistry*, 248(9):3313 - 3326, 1973.

[22] K. A. Satyshur, S. T. Rao, D. Pyzalska, W. Drendel, M. Greaser, and M. Sundaralingam, "Refined Structure of Chicken Skeletal Muscle Troponin C in the Two-calcium Start at 2-Å Resolution," *Journal of Biological Chemistry*, 263(4):1628 - 1647, 1988.

[23] R. M. Tufty and R. H. Kretsinger, "Troponin and Parvalbumin Calcium Binding Regions Predicted in Myosin Light Chain and T4 Lysozyme," *Science*, 187:167 - 169, 1975.

[24] N. K. Vyas, M. N. Vyas, and F. A. Quiocho, "A Novel Calcium Binding Site in the Galactose-binding Protein of Bacterial Transport and Chemotaxis," *Nature*, 327:635 - 638, 1987.

[25] J. U. Bowie, R. Luthy, and D. Eisenberg, "A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure," *Science*, 253:164 - 170, 1991.

[26] R. L. Williams, S. M. Greene, and A. McPherson, "The Crystal Structure of Ribonuclease B at 2.5Å Resolution," *Journal of Biological Chemistry*, 262(33):16020 - 16031, 1987.

[27] F. S. Matthews, "The Structure, Function, and Evolution of Cytochromes," *Progress in Biophysics and Molecular Biology*, 45:1 - 56, 1985.

[28] G. K. Farber and G. A. Petsko, "The Evolution of α/β Barrel Enzymes," *Trends in Biochemical Sciences*, 15:228 - 234, 1990.

[29] E. Horowitz and S. Sahni, *Fundamentals of Computer Algorithms*, Computer Science Press, 1978.

[30] J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey, *Graphical Methods For Data Analysis*, Boston, 1983.

Chapter 3

Compactness

&

Protein Secondary Structure

Preface

The construction of a simple model of a complex system serves several purposes. First, it allows us to test our understanding of the most important features of a system. Second, by modeling a subset of forces, the complexity of the system is reduced. Hence, computation is more tractable and the relationship between the governing forces may be easier to understand and interpret. Third, it allows one to formulate hypotheses about the behavior of complex systems based on general principles. By describing a complex system in simple terms, one can gain insights into a system which may otherwise be obscured by its complexity. Finally, a simple model can provide a framework for building more complex models.

Ken Dill and his co-workers have utilized a simple model of protein structure for several years. In its simplest form, chains are represented as linear strings of monomers (i.e., amino acids). Excluded volume is modeled by never allowing two monomers to occupy the same lattice site. Bond angles are limited to 90 and 180 degrees. In some complicated studies, sequence effects are studied by defining two monomer types: hydrophobic and polar. Although it is simple, the model was designed to capture the dominant forces of protein folding (for a review, see [1]). Two of these forces are the hydrophobic effect and conformational entropy.

Protein folding is opposed by a large decrease in conformational entropy. Put simply, there are many more denatured (unfolded) polypeptide chain configurations than there are native state (folded) configurations. The lattice model captures certain physics involved in this process. By counting the total number of possible open and compact configurations, one can explicitly calculate the loss of conformational entropy upon

molecular collapse for the lattice model. Likewise, by classifying the 20 amino acids into 2 types - hydrophobic and hydrophilic - the lattice model can be used to explore the effect of varying sequence on protein folding.

In 1989, Hue Sun Chan and Ken Dill used the lattice model to study the conformational properties of compact polymers [2]. By restricting themselves to short chain lengths for which they could exhaustively enumerate all the possible ways of configuring a chain on the lattice, they were able to analyze the *complete* conformational space of the chain. One of the first applications of the model was to look at loop formation in polymers. Previous treatments of this problem were based on the random flight theory of Jacobson and Stockmayer (see Cantor & Schimmel [3] for an introduction).

Using the lattice model, the entropic free energy change of loop formation was calculated by counting the number of conformations with a specified contact, (i,j). The entropic cost of forming a loop is given by taking the logarithm of the ratio of the number of conformations which contain the contact, (i,j) and the total number of conformations possible. That is:

$$\Delta G_{entropic} = -kT\ln[\frac{Q(i,j)}{Q_o}], \tag{3.1}$$

where $Q(i,j)$ is the total number of conformations in which monomers i and j come together to form a contact and Q_o the the total number of possible conformations. For the formation of a single contact, it was determined that the small loops were favored over large loops [4]. This is because a small loop minimally reduces the number of possible conformations. Hence, the entropic cost of forming a small loop is less than the cost of forming of a large loop.

They then turned to the issue of multiple contacts and asked "given an association of two chain monomers, (i,j), what is the next most probable contact, (i',j')"? In order to address this question, a contact (i,j) was presumed to exist. Next, the total number of conformations which contained both (i,j) and (i',j') was determined for all possible combinations of i' and j'. As in equation 3.1, the entropic free energy can be expressed as:

$$\Delta G_{entropic} = -kT\ln[\frac{Q(i,j;i',j')}{Q_o}], \tag{3.2}$$

where $Q(i,j;i',j')$ is the total number of conformations which contain both (i,j) and (i',j'). A convenient way to display the result of this calculation is by using a topological free energy surface.

A topological free energy surface is the upper (or lower) triangle of an $N \times N$ matrix in which each matrix element corresponds to a different pair of residues. The value of each matrix element is the entropic free energy calculated using equation 3.2. Figure 3.1 shows the topological free energy surface for a 12 residue In the figure, residues **5** and **8** are presumed to form a contact, (i,j). The figure shows that contacts between some residue pairs are impossible given the formation of a contact between (5,8). (4,7) is an example of a prohibited contact. The darkest regions on the free energy surface represent the most favorable contacts, whereas the lightest regions are the most disfavored contacts. Given.the contact (5,8), it is apparent from the figure that the most favored contacts are (3,6), (7,10), and (4,9).

Chan and Dill realized that the most favorable contacts formed patterns which were similar to the patterns of contacts observed for secondary structures (i.e., helices and sheets) in proteins. They defined helix and sheet structures on the lattice and derived the patterns of contacts consistent with the formation of lattice secondary structure (figure

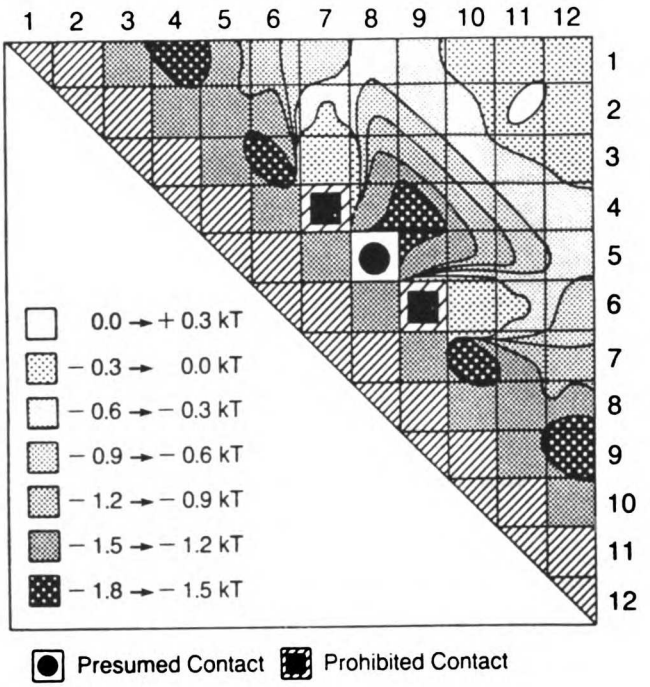


figure 3.1: Topological entropic free energy surface for 12 residue chains configured on a 3 dimensional cubic lattice. Figure reproduced from [5].

3.2). Having defined secondary structures on the lattice, Chan and Dill found that the amount of secondary structure increases dramatically as a function of compactness. Furthermore, the distribution of helices and sheets resembled the distribution obtained from analysis of structures in the PDB. The lattice work provided the basis for the hypothesis that a non-specific force, such as compactness, could provide a driving force for the formation of secondary structure. This hypothesis was unexpected and novel given the traditional view that the driving force for secondary structure formation in proteins comes from hydrogen bond interactions and intrinsic propensities of the polypeptide chain.

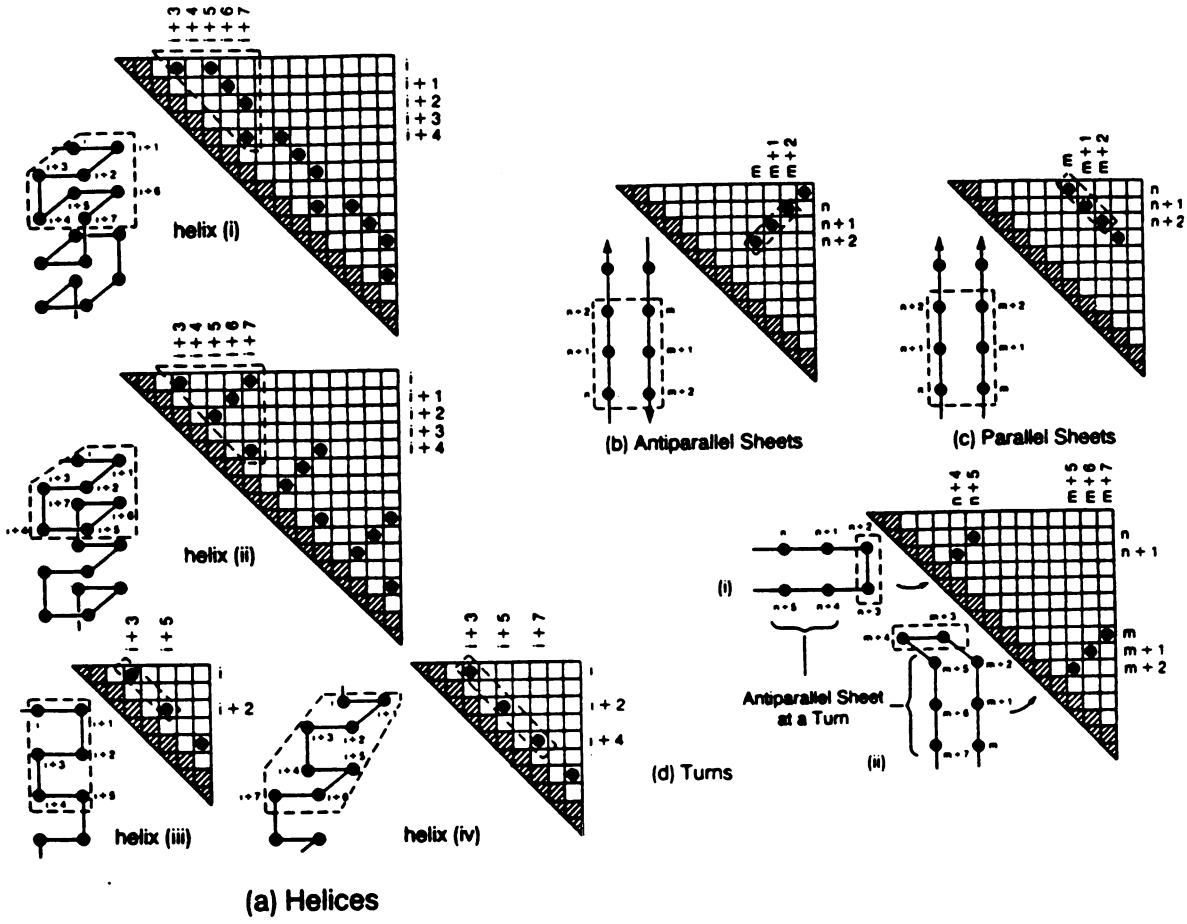


figure 3.2: Patterns of contacts defining secondary structures in lattice models. Reproduced from [4]

The lattice-based result that secondary structure formation in proteins could be driven by compactness turned out to be quite controversial. Many studies subsequently appeared which attempted to determine how the result extrapolated to off-lattice model systems. The next section contains a summary of these works, and describes work which examines the hypothesis that compactness drives secondary structure in an off-lattice system.

References

[1] K. A. Dill, "Dominant Forces in Protein Folding," *Biochemistry*, 29(31):7133 - 7155, 1990.

[2] H. S. Chan and K. A. Dill, "Compact Polymers," *Macromolecules*, 22:4559 - 4573, 1989.

[3] C. R. Cantor and P. R. Schimmel, *Techniques for the Study of Biological Structure and Function*, Freeman, San Francisco, 1980.

[4] H. S. Chan and K. A. Dill, "The Effects of Internal Constraints on the Configurations of Chain Molecules," *Journal of Chemical Physics*, 92(5):3118 - 3135, 1990.

[5] H. S. Chan and K. A. Dill, "Origins of Structure in Globular Proteins," *Proceedings of the National Academy of Science, USA*, 87:6388 - 6392, 1990.

Does compactness induce secondary structure in proteins? A study of poly-alanine chains computed by distance geometry.

Running Title: Does compactness induce secondary structure?

David P. Yee, Hue Sun Chan, Timothy F. Havel *, & Ken A. Dill

Department of Pharmaceutical Chemistry

University of California, San Francisco, CA 94143-1204

* Dept of Biological Chemistry and Molecular Pharmacology

Harvard Medical School

240 Longwood Ave,Boston, MA 02115

Keywords: secondary structure, compactness, distance geometry, protein folding

Latest Revision: May 11, 1994

Summary

A few years ago, lattice model studies indicated that compactness could induce polymer chains to develop protein-like secondary structures. Subsequent off-lattice studies have found the amounts of induced structure to be relatively small. Here we use distance geometry to generate random conformations of compact poly-alanine chains of various chain lengths. The poly-alanine chains are subjected only to compactness and excluded volume constraints; no other energies or conformational propensities are included in the chain generation procedure. We find that compactness leads to considerable stabilization of secondary structure, but the absolute amount of secondary structure depends strongly on the criteria used to define helices and sheets. By loose criteria, much secondary structure arises from compactness, but by strict criteria, little does. The stabilization free energy of secondary structure provided by compactness, however, appears to be independent of criteria. Since real helices and sheets in proteins can be identified by strict criteria, we introduced small energy perturbations to compact poly-alanine chains using the AMBER force field. Small refinements produced good α-helices. For β-sheets, however, larger refinements are necessary. Compactness appears to impart stability, but not much structural specificity, to secondary structures in proteins. Compactness acts more like diffusion as a force, a result of ensemble statistics, than like pair interactions such as hydrogen bonding.

1. Introduction

Nearly 50% of the residues in globular proteins are in either α-helix or β-sheet [1, 2]. What are the forces that stabilize secondary structures in globular proteins? Helices and sheets can be identified on the basis of their hydrogen bonding patterns [2], so it is reasonable to expect that hydrogen bonds play some role. But since the short helices that are predominant in globular proteins are not very stable when isolated in solution, other aspects of the surrounding protein must help stabilize them [3]. Furthermore, recent results on the refolding of cytochrome c suggest that the protein population refolds at the same rate as molecular collapse [4]. This result precludes the formation of substantial amounts of stable secondary structure prior to collapse. A few years ago, it was proposed that compactness in single polymer chains could contribute substantially to the formation of regular internal structure in proteins [5-7]. Based on exhaustive simulations of short chains on 2-dimensional square lattices and systematic conformational searches on 3-dimensional cubic lattices, it was found that the amount of secondary structure increases sharply with the compactness of a polymer chain. Those studies found roughly the same chain length distributions of helices and sheets in compact lattice chain conformations as in the protein structures in the Protein Data Bank (PDB) [8].

Subsequently other studies of lattice models [9] and off-lattice models [10-13] have explored in greater detail the role of compactness in inducing secondary structure. They are summarized briefly below.

Gregoret and Cohen [10] generated native-like random chains using an off-lattice rotational isomeric model of proteins. Chains were modelled as linear strings of residues. Each residue excluded a spherical volume. Random chain conformations were

constructed using a backtrace Monte Carlo procedure in which virtual bond and torsion angles were selected from the distribution observed in real protein structures. Chains were restricted to lie within ellipsoids defined by a formula derived from real proteins. Gregoret and Cohen observed an increase in secondary structure with increasing compactness, but only found significant amounts of secondary structure, by their criteria, when the chains are 30% more compact than real proteins.

Hao et al. [11] also used a Monte Carlo backtrace procedure to generate random chains. They too used a simplified representation of proteins where side chains are represented as spheres. The chains have random amino acid sequences. Compact chain conformations were generated with torsion angles selected either randomly or weighted to reflect local interactions imposed by the peptide bond. Hao et al. compared a bond vector correlation function for real proteins with the random chains and concluded that the observed bond vector correlations in real proteins can be reproduced best when (1) the chains are constrained to be compact and (2) intra-residue interactions are included.

Kolinski and Skolnick [9] used a high-resolution lattice model of poly-alanine and poly-valine chains. Local and nonlocal alanine-alanine and valine-valine interactions were derived from angular correlations of sidegroup vectors in proteins taken from the PDB. Hydrogen bonding was included in the form of two terms: (1) an attractive hydrogen bond term and (2) a cooperative interaction for adjacent hydrogen bonds. Conformational space was explored by a Monte Carlo dynamics algorithm. The poly-alanine chains readily formed helical structures at low temperatures and exhibited a cooperative helix-coil transition. Similar results were obtained when (1) local bond correlations and hydrogen bonding terms were used, (2) when a nonlocal hydrophobic term was added, and (3) when the hydrophobic and hydrogen bonding terms were used without local bond correlations. In no case did the poly-alanine chains collapse into compact states.

The poly-valine chains, however, collapsed into compact, β-sheet-like conformations when using bond correlation, hydrogen bonding, and hydrophobic interactions. When only the hydrogen bonding and hydrophobic terms were used in the absence of local bond correlations, they collapsed to compact states consisting of about 25% helix and no β-sheet. When only the hydrophobic interaction was used, the chains collapsed into compact states, but no appreciable secondary structure was found. They concluded that collapse alone was not sufficient to produce secondary structure.

Socci et al. [12] developed a simplified model in which each residue was represented by a single point. Random conformations were generated by minimizing a potential which included a covalent term for chain connectivity, an r^{12} term for excluded volume, and a radius of gyration term to drive compactness. Two constants in the potential which control (1) the balance between covalent and non-covalent forces and (2) strength of excluded volume were derived by examining real proteins. They looked for repeating structure by defining a function which measured correlations between dihedral angles between different points along the protein chain. They could not detect any well defined repeating patterns that resembled secondary structure if the chains were only constrained to adopt compact conformations.

Hunt et al. [13] extended the earlier work of Gregoret and Cohen [10]. They used an all-atom, off-lattice model of protein structure. Conformational space was searched using a Monte Carlo simulated annealing method in which the ϕ/ψ angles of a randomly selected residue were reassigned from a distribution of ϕ/ψ angles derived from the PDB. Simplified energy functions were used which included (1) a radius of gyration term to induce compactness in the chain, (2) an energy term for hydrogen bond effects, or (3) a combination of (1) and (2). Hunt et al. observe an increase in secondary structure when only the compactness term is used. When a combination of the compactness and hydrogen bond terms are used, they are able to generate highly compact conformations

with amounts of secondary structure comparable to real proteins.

Why is there a need for yet another study? While these studies have contributed considerable insight into the role of packing in secondary structure formation, they have also raised new questions. Although they represent proteins more accurately than the original lattice model, they, too, are simplified models. In order to avoid simplified side-chain and lattice models, we study here an all-heavy-atom model of poly-alanine, of chain lengths 50, 100, and 150, in a continuum representation. Chirality is taken into account, whereas it is not in some of the earlier studies. The use of backtrace Monte Carlo and constraining ellipsoids can introduce conformational bias, so here we use distance geometry to constrain the conformations. Our statistics indicate that the conformations generated by distance geometry are relatively unbiased. In lattice models, compactness and secondary structure can be defined with little ambiguity, but in off-lattice models it is not so simple. What conformations should be called helices and sheets? Does the amount of secondary structure observed depend on the criteria used to define it? Here we explore these issues.

2. Methods

Generating unbiased compact poly-alanine conformations

Poly-L-alanine conformations were generated using the DG-II distance geometry program [14], with sequential tetrangle inequality bound smoothing, randomized metrization using a uniform distribution, and embedding in four-dimensions followed by 10,000 steps of dynamical simulated annealing refinement in which the superfluous dimension was eliminated. The resulting convergence rate averaged about 80%, and the annealing was simply repeated (using different initial velocities) on nonconvergent conformations until they converged. In addition to the constraints needed to obtain the

proper covalent geometry and chirality, a uniform upper bound was imposed on all the nonbonded distances, thereby effectively packing the polypeptide chain into a sphere whose radius is half of that upper bound. Previous extensive studies of the randomized metrization method on unconstrained poly-L-alanine chains [15] have demonstrated that it yields coordinates that can readily be refined to polypeptide chains whose statistical properties are in accord with the statistical mechanics of chain molecules. It is therefore reasonable to expect that the compact, self-avoiding polypeptide chains generated in this paper, for which few theoretical predictions can be made, are likewise satisfactorily unbiased.

Determining compactness

Determining compactness is simple for lattice models, but is more difficult for more realistic models. For example, Gregoret and Cohen [10] calculated the radius, R, of a hypothetical ideal spherical protein based on average properties of amino acids and proteins using the expression:

$$R = \left[\frac{3 \times N \times m}{4\pi \times \rho_P \times 10^6 \times N_A} \right]^{1/3} \times 10^{10} , \qquad (3.3)$$

where N is the number of residues, m is the average molecular weight of an amino acid (110g / mol), ρ_P is the average density of globular proteins (1.4 g / mol), N_A is Avogadro's number, and $10^6 ml/m^3$ and $10^{10} A/m$ are unit conversion factors. Compactness was then varied by scaling the chain volume using the equation:

$$V = \varepsilon \times \frac{4\pi abc}{3,} \qquad (3.4)$$

where a, b, and c are the principal axes of the ellipsoid and $abc = R^3$. Gregoret and Cohen suppose the condition $\varepsilon = 1$ represents real native proteins, and $\varepsilon < 1$ represents higher densities.

However using their ideal values in equation 3.3, the volume occupied by a single amino acid is $130.4A^3$. But in their chain generation procedure, a residue is allowed to be placed as close as 4.25A from any other non-bonded residue. This implies that the volume excluded by one chain segment is at most only $4\pi/3 \times (4.25/2)^3 = 40.2A^3$. Hence, excluded volume is considerably underestimated, by $130.4 - 40.2 = 90.2A^3$ per amino acid. Whereas they estimate a native-like radius of gyration for their 64-mers of 11.0A, if we use their excluded volume of $40.2A^3$ per residue, then the radius that contains all the residues of the maximally compact state would be

$$R = \left[\frac{3(40.2A^3)(64)}{4\pi(0.74)}\right]^{1/3} = 9.4A, \qquad (3.5)$$

which is equivalent to a radius of gyration of $R_G = \sqrt{3/5}R = 7.3A$ [16]. The factor of 0.74 in the denominator of equation (3.5) approximates the maximal packing density for the model system: it is both the theoretical maximum packing density of close-packed spheres of the same size and the packing density observed in crystals of small organic molecules [17]. Note that if $130.4A^3$ is used in equation (3.5) instead of $40.2A^3$ for the volume excluded per amino acid, we obtain $R_G = 10.8A \approx 11.0A$. While their studies show less than 20% secondary structure with constraining radii of 11.0 A, their studies with radii of gyration of 7.3A show more than 35% secondary structure.

In our study, chains were constrained to be enclosed by spheres of specific radii by the distance geometry procedure. Specifying a large radius yields relatively open chains, while specifying a small radius gives compact chains. We generated several sets of random chains varying in both chain length and compactness. Table 3.1 lists the chain lengths, the constraining radii, and the average radius of gyration for each set of poly-alanine conformations. Figure 3.3 shows an example of a near maximally compact poly-alanine 100mer.

Table 3.1

Poly-Alanine Chains

# Residues	Constraining Radius	$<R_G>^*$	# Conformations
50	10.0	7.96	100
	11.0	8.62	100
	12.5	9.60	100
	15.0	11.14	100
100	12.5	10.11	100
	14.0	11.12	100
	16.5	12.72	50
	20.0	14.96	50
150	15.0	11.92	25
	17.0	13.26	25
	20.0	15.25	25
	25.0	18.39	25

*: R_G is based on C_α atom positions.

By two different criteria, the most compact sets of these chains are nearly maximally compact. First, we use the criterion of Chothia [18], by which the mean volume of alanine, when buried in the protein core, is $91.5A^3$. The minimal radius of a maximally compact chain is given by:

$$\frac{4}{3}\pi R^3 = N \times 91.5A^3 \tag{3.6}$$

where R is the radius in Angstroms and N is the chain length. Solving equation (3.6) for $N = 50$, 100, and 150 gives maximally compact radii of 10.3, 13, and 14.8-A respectively. The most compact sets of 50, 100, and 150mers have constraining radii of 10, 12.5, and 15 A, so by this criterion, the chains are nearly maximally compact.

Second, we use the criterion of Maiorov and Crippen [19] who derived an expression for the minimal radius of gyration for polypeptide chains:

$$R_{\min} = -1.26 + 2.79 \times N^{\frac{1}{3}} \tag{3.7}$$

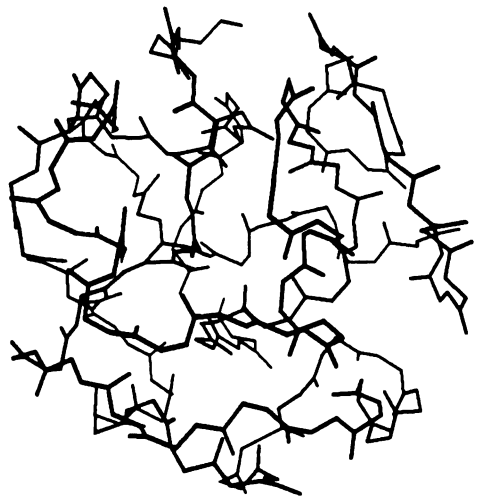


Figure 3.3: Conformation of near maximally compact poly- alanine 100mer generated by the distance geometry method.

According to equation (3.7), the radius of gyration of a maximally compact 50, 100, and 150-mer should be 9.0, 11.7, and 13.6A^3. Thus, by this second criterion too, the poly-alanine chains are nearly maximally compact.

Defining helices and sheets

There is no single correct way to determine secondary structure. There are several published methods for identifying secondary structures in proteins [1, 2, 20]. While they are largely consistent with one another, and differ mainly at the ends of helices and sheets, their differences, nevertheless, can be considerable [21]. In the present study, we use three very different methods for assigning secondary structures. We use Define [20] , based on inter-residue distances, DSSP [2] based on hydrogen bonding patterns, and a topological contact (TC) method [6, 7] based on patterns of inter-residue contacts. We are able to use DSSP effectively only in cases when we include hydrogen bonding

interactions in the AMBER refinements, since it is not applicable to models without defined hydrogen bonds. The methods are described below.

Define

Define [20] identifies secondary structure by comparing the inter-residue distance map of a chain segment with the inter-residue distance maps of an ideal α-helix and extended strand. If the difference distance map between the chain segment and the ideal helix or strand is below some threshold, then the residues in the segment are identified as participating in secondary structure. Error thresholds and mismatch limits were set to their default values. Since this procedure does not match strands to locate or identify sheets, Define is used only to identify helices and strands.

DSSP

DSSP [2] is the most stringent definition we used. It is based primarily on hydrogen bonding patterns. Helices, for example, are defined in terms of repeats of hydrogen bonds between residues separated by 3, 4, or 5 residues. Sheets are identified by locating hydrogen bonds between residues which are not close in sequence.

Topological Contact (TC)

One definition of secondary structure that does not depend on geometries or bond angles, but depends only on the topology of neighboring contacts was given by Chan and Dill [6, 7] It was implemented and applied to model and real protein structures by Gregoret and Cohen [10], who showed that it correctly identifies helices and sheets in proteins. By identifying secondary structures on the basis of specific contacts, the TC method is similar to the definition of secondary structure used by Levitt and Greer [1].

In this method, a helix is identified by patterns of specific contacts between residues. A contact between two residues is defined when the distance between the C_α atom positions is less than some cutoff value. In this work, we primarily use a cutoff of 5.5-A. A helix is defined if either of the following conditions hold:

(1) [(i, i + 3), (i + 2, i + 5)] and

(2) [(i, i + 3), (i, i + 5), (i + 1, i + 6), (i + 2, i + 7), (i + 4, i + 7)].

Since (2) is mainly used to identify helices on lattices, definition (1) was the main identifier of helices in the poly-alanine chains. Antiparallel sheets are defined by contacts between residues [(i, j + 2), (i + 1, j + 1), (i + 2, j)]. Parallel sheets are defined by contacts between residues [(i, j), (i + 1, j + 1), (i + 2, j + 2)]. All residues involved in a putative sheet must be in an extended conformation specified by certain lower bounds on the interior virtual bond angles.

Energy minimization

In some of the studies below, the constrained conformations were further refined using energy minimization. The AMBER [22, 23] potential as implemented in the InsightII 2.2.0 / Discover 2.9 molecular modeling package from Biosym Technologies was used in all energy minimizations. 1-4 non-bonded interactions were scaled by 0.5.

Structural comparison and clustering

To determine whether energy minimizations substantially perturbed the conformations, we used CONGENEAL [24], a computer algorithm that calculates the structural dissimilarity between conformations. CONGENEAL represents chains in terms of their weighted distance maps. The weighted distance map of a protein chain that has N residues is an $N \times N$ matrix in which each matrix element (i, j) is a weight, w, equal to the distance, $d_{i,j}$, between the α-carbons of residues i and j, raised to a power, -2 (i.e., $w_{i,j} = d_{i,j}^{-2}$). The essential feature of the weighting function is that residues

which are close together in space are given more weight than residues which are distant in space.

Comparisons are performed by superimposing two weighted distance maps and summing the absolute differences between corresponding distance weights. The dissimilarity is defined as the sum of the absolute differences between corresponding inter-residue distance weights normalized by the average of the summed distance weights for each of the two distance maps. So for two chain conformations R and S, the dissimilarity between the chains is given by:

$$d(R,S) = \frac{\sum\sum \mid r_{ij}^{-2} - s_{ij}^{-2} \mid}{\frac{1}{2}\left(\sum\sum r_{ij}^{-2} + \sum\sum s_{ij}^{-2}\right)} \tag{3.8}$$

For P structures, all $P\ (P-1)/2$ possible pairs of structures are first compared, then clustered using a hierarchical clustering method. Clustering is performed by iteratively grouping P structures into larger and larger groups. The similarity between groups is defined to be the average of the similarities between the members of one group compared to the members of another group. At the beginning of the clustering process, the two most similar structures are merged into a single group. Then the next most similar pair of structures or groups of structures are merged. This process continues until all structures are finally merged into a single group at the top of the tree. The final tree-like structure shows the inter-relatedness of the structures within the set.

3. RESULTS

Our first question was whether the distance geometry method for generating compact conformations led to local biases in the bond angles. We found that it did not. While the studies of Gregoret and Cohen [10] and Hunt et al. [13] focus mainly on

chains with intrinsic peptide bond angles, taken from the distribution observed in the PDB, our interest here was to explore a slightly different question of whether polymers without local bond biases could be induced to have secondary structures. Intrinsic peptide bond propensities must surely help in the formation of secondary structures, but we are also interested to know whether other types of polymers might be induced to form secondary structures. Gregoret and Cohen [10] constructed conformations by adding C_α positions to a growing chain. The position of each new C_α atom was selected by picking virtual bond angles (α) and virtual torsion angles (τ) which reproduced the α/τ distribution derived from a database of real protein structures. If, after a given number of attempts, a new C_α position could not be assigned, the previously placed C_α atom was reassigned. Figure 3.4A shows the distribution of α / τ angles reproduced by the chain generation method of Gregoret and Cohen. Their bond conformations are concentrated in specific regions of α / τ space. In contrast, figure 3.4B shows that the α / τ distributions obtained from our constrained poly-alanine chains are uniform, reflecting the absence of local interactions in the chain generation procedure.

Because our poly-alanine chains are based on an all-heavy-atom representation, our ϕ/ψ plots are consistent with ϕ/ψ (Ramachandran) plots of hard sphere models, and include chirality, in contrast to earlier lattice [6, 7] and off-lattice [10] models, for which the chain representations are too simple to include chirality.

ϕ/ψ dihedral angle distributions

In proteins, ϕ/ψ angles are clustered into two primary regions: an α_R region and a β region. These areas are associated with α-helix and β-sheet respectively. In addition, the α_L region is somewhat populated in real proteins. Figure 3.5A shows a ϕ/ψ map derived from proteins in the PDB. The α_R and β regions are prominent on the left hand side of the figure. The less populated cluster on the upper right hand side of the figure represents

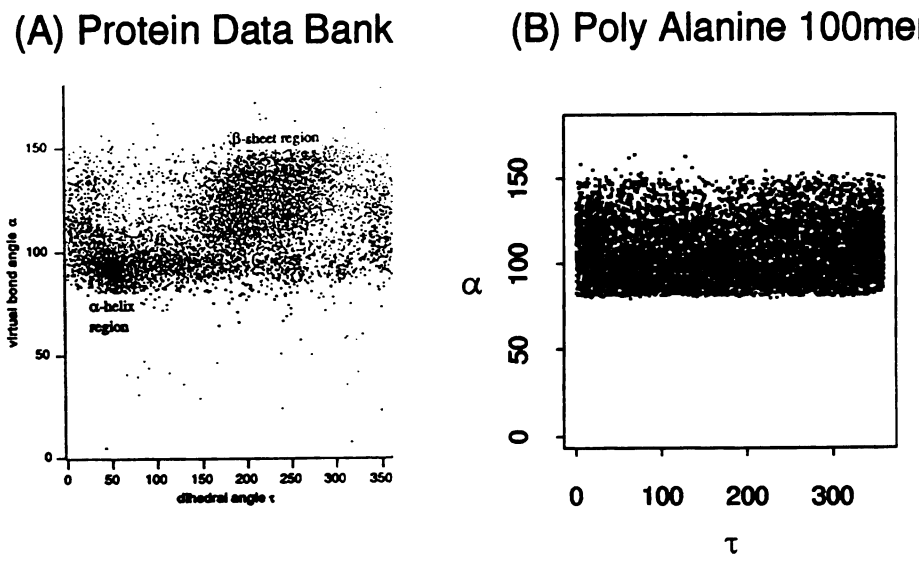


Figure 3.4: Distributions of virtual bond angle α versus virtual torsion angle τ. (A) Distribution derived from real protein structures. Reproduced from [10]. (B) Distribution derived from random, compact 100mer poly-alanine conformations.

the α_L region of the map.

Since the observed distribution of dihedral angles in real proteins is generally consistent with simple steric avoidance [25], it is not unexpected that the dihedral angle distributions of the poly-alanine chains generated in this work are similar to the dihedral angle distribution observed in real proteins. For example, there is a tilted oval near the center of the ϕ/ψ plot (i.e., $(\phi, \psi) \approx (0, 0)$) which is not populated (see figure 3.5B). This region represents steric clashes between the oxygen atom of residue i with either (1) the carbonyl carbon of atom $i + 1$ or (2) with the amide hydrogen of residue $i + 1$. The blank region centered at $\phi = 0$ and extending from $\psi = -180$ to $+180$ is due to contacts between the oxygen atoms of residues i and $i + 1$. There is a forbidden region for all ψ centered

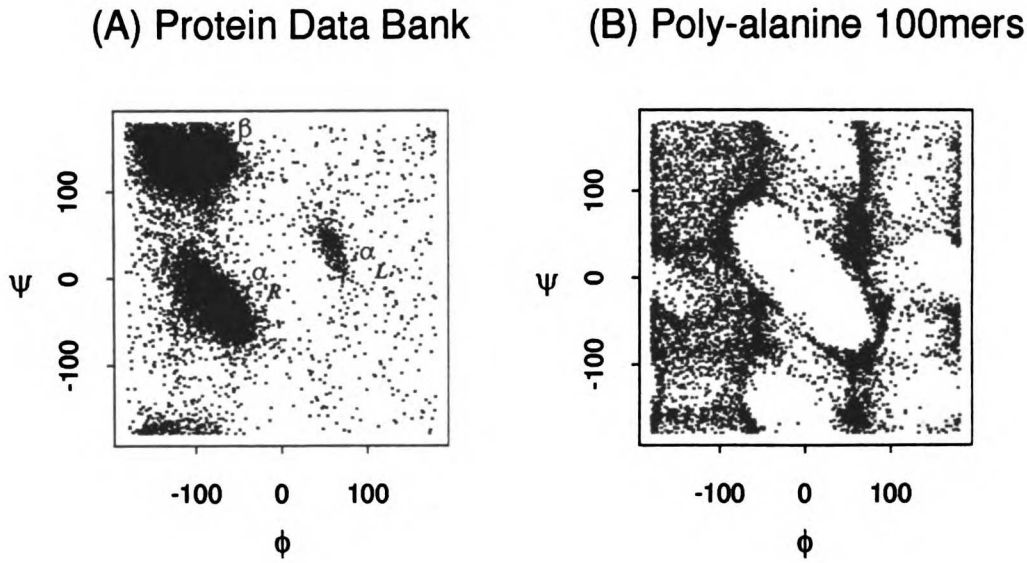


Figure 3.5: (A) Distribution of ϕ/ψ angles derived from crystal structures in to the Brookhaven Protein Data Bank. (B) ϕ/ψ distribution of 100 random, near maximally compact 100mer poly-alanine chains. The ϕ/ψ distributions of all other sets of poly-alanine chains (both compact and open) are nearly identical.

at $\phi \approx 120$ which is due to interactions of the peptide backbone with side chain atoms. Comparison of figures 3.5A and 3.5B shows that most poly-alanine ϕ/ψ angles fall loosely within the β, α_R, and α_L regions of the ϕ/ψ map.

There are also differences, however, between the ϕ/ψ distribution of the poly-alanine chains and the distribution from real proteins. First, the distribution of ϕ and ψ angles is more diffuse and spread out than in real proteins. Second, whereas proteins have a small α_L region corresponding to a left handed helix, these poly-alanines have a long continuous strip of populated dihedral angles defined by $(\phi, \psi) \approx$ $(60 - 80,\ -180 - 180)$. Some of the clustering on the right hand side of the ϕ/ψ map is due to the nature of the error functions used in the distance geometry procedure. This is

because the basin of attraction leading into the allowed region is large relative to the size of that region [15]. Nonetheless, the detailed distribution of dihedral angles observed in real proteins cannot be explained completely on the basis of excluded volume effects. Therefore, although the ϕ/ψ angle combinations observed in the poly-alanine chains do not involve steric violations, some of the ϕ/ψ angle combinations are not energetically favorable conformations. This result is consistent with molecular dynamics simulations of model alanyl dipeptides which show that, in solution, the regions on the left hand side of the ϕ/ψ map (in particular, the α_R and β regions) are strongly favored relative to regions on the right hand side of the map [26].

We found that the dihedral angle distributions in poly-alanine are independent of chain length and compactness. Thus chain compactness appears not to influence local conformational preferences at the level of single pairs of dihedral angles.

Secondary structure in poly-alanine chains

How much secondary structure is there in the confined poly-alanine chains? We used three different criteria to identify helices and sheets. According to **DSSP**, the maximally compact chains do not contain regular secondary structures in the form of α-helices or β-sheets. This result is not unexpected since DSSP relies on the identification of hydrogen bonds to locate secondary structure. Since no energetics were used to generate these poly-alanine chain conformations, there are few residue-to-residue orientations which can be identified as being hydrogen bonded. Applying **Define** to the maximally compact chains shows that the chains have significant strand content. About 14% of residues in all the poly-alanine chains (compact and non-compact) are in extended strand conformations. By the **TC** criterion, the compact chains have significant anti-parallel sheet content. Both Define and TC methods agree that there is very little α-helix content. Figure 3.6 summarizes the absolute amount of secondary structure in

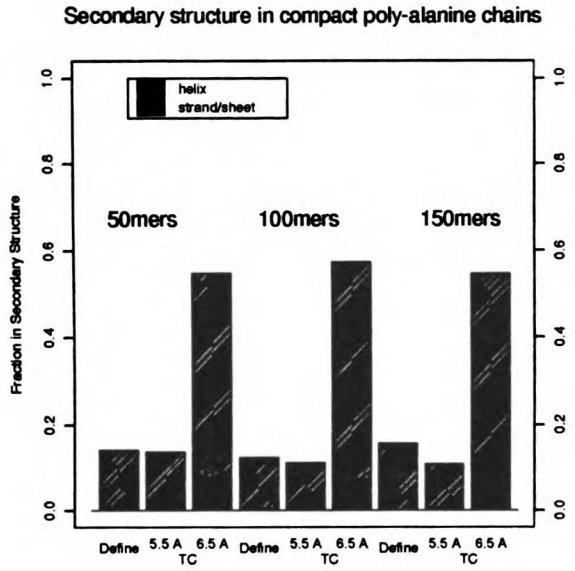


Figure 3.6: The absolute amount of secondary structure in compact poly-alanine 100mers. Define identifies the amount of extended strand and helix. TC identifies sheet and helix. The result of using both a 5.5A and 6.5A cutoff with the TC definition of secondary structure is shown.

poly-alanine 50, 100, and 150mers determined by a variety of criteria.

By all criteria, the amount of secondary structure increases with chain compactness. Since there is no *a priori* correct distance to use as a cutoff for defining a contact for the TC criterion, we systematically varied: (1) the cutoff distances defining a contact and (2) the bond angle constraint required for defining an extended strand conformation. Reducing the cutoff distance defines helices more stringently. Consistent with the earlier lattice studies [6, 7], figures 3.7 and 3.8 show that the amount of helix and sheet defined by these various criteria increases with compactness. The figures also show the effect of altering the stringency of the secondary structure definitions. By varying the cutoff parameter from 5.5 to 6.5-A, and the bond angle of strands from 100 to 120 degrees, the

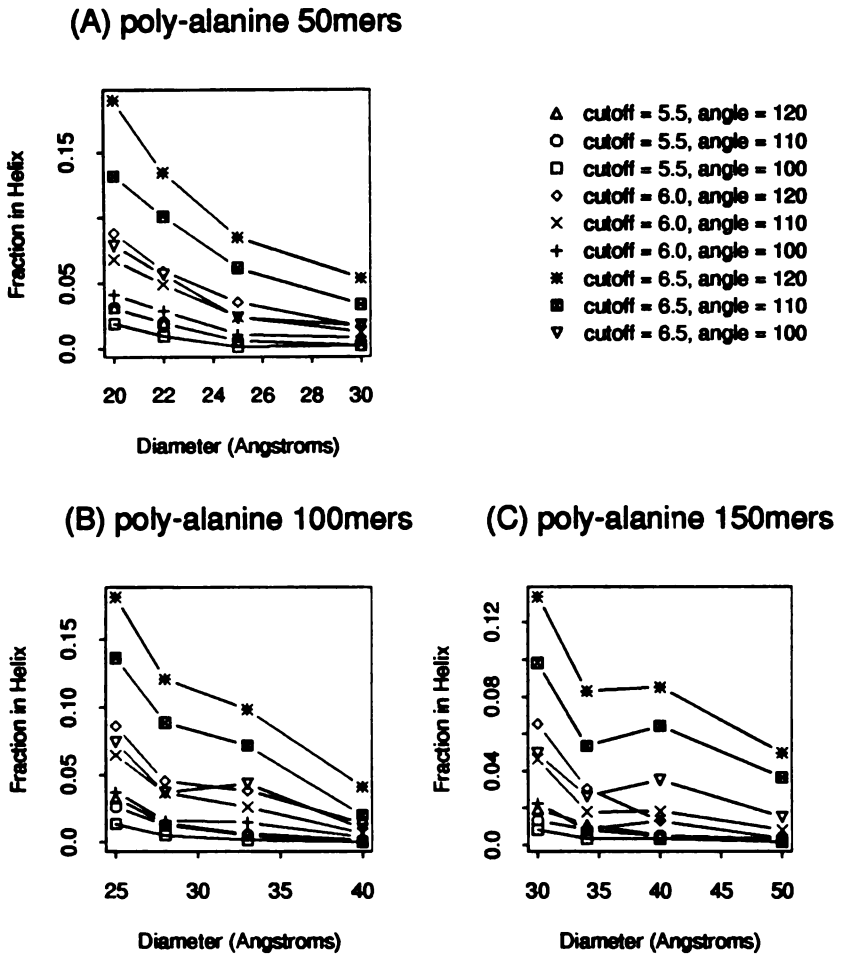


Figure 3.7: Amount of helix as a function of compactness as determined by the TC definition. Cutoffs defining a contact were varied from 5.5 - 6.5-A. The bond angle required for residues to be assigned to a sheet conformation was varied from 100 degrees to 120 degrees. (A) Poly-alanine 50mers (B) Poly-alanine 100mers (C) Poly-alanine 150mers

amount of observed secondary structure can vary over a large range: from about 1% to 20% for helix and from 3% to 50% for sheet. For all criteria, the helix and sheet content increases with compactness.

While the amount of *structure* is strongly dependent on the defining criterion, the amount of *stabilization free energy* is not. Figure 3.9 shows that there is a free energy stabilizing secondary structures that comes from compactness. The loss in free energy, $-\Delta G_{compactness}$, is the logarithm of the ratio of the fraction of secondary structure residues in the compact state to that of the open state.

$$-\Delta G_{compactness}/kT = \ln\left[\frac{(fraction\ in\ 2^{o}structure)_{compact}}{(fraction\ in\ 2^{o}structure)_{open}}\right] \qquad (3.9)$$

Figure 3.9 shows the amount of secondary structure as a function of compactness plotted on a logarithmic scale. The y-axis corresponds to free energy lost in units of kT. Surprisingly, despite the wide range in the absolute numbers of residues in secondary structure by different criteria, the change in free energy with compactness appears to be largely independent of the criteria used.

More informative than the total amount of secondary structure is its distribution. Figures 3.10A-I show the distributions of compact (blue) and open (pink) conformations as a function of the number of residues in secondary structure. Each panel represents a different set of criteria used with the TC definition of secondary structure. From the upper left to the lower right represents decreasing stringency of criterion. In all cases, it is clear that the absolute amount of secondary structure is higher in the compact conformations than in the open conformations. In addition, for poly-alanine chains in compact ensembles, there are more conformations with large numbers of residues in secondary structure than with small numbers. The reverse is true for ensembles of open chain conformations. Almost no open chains have large amounts of secondary structure.

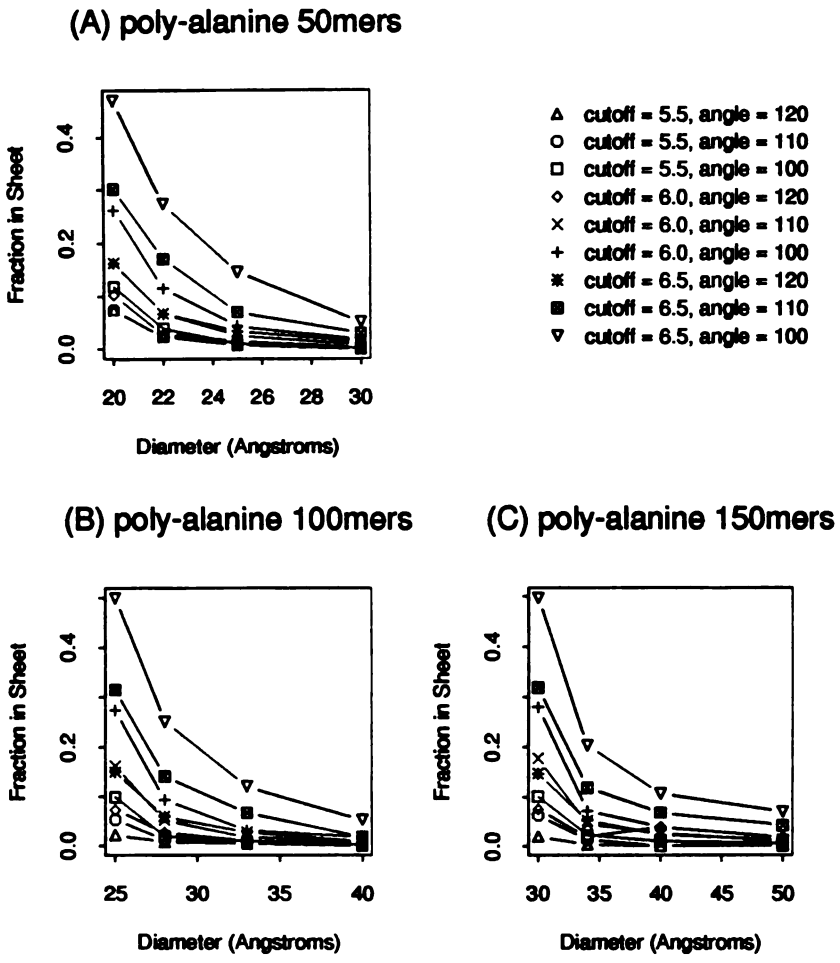


Figure 3.8: Amount of sheet as a function of compactness as determined by the TC definition. Cutoffs defining a contact were varied from 5.5 - 6.5-A. The bond angle required for residues to be assigned to a sheet conformation was varied from 100 degrees to 120 degrees. (A) Poly-alanine 50mers (B) Poly-alanine 100mers (C) Poly-alanine 150mers

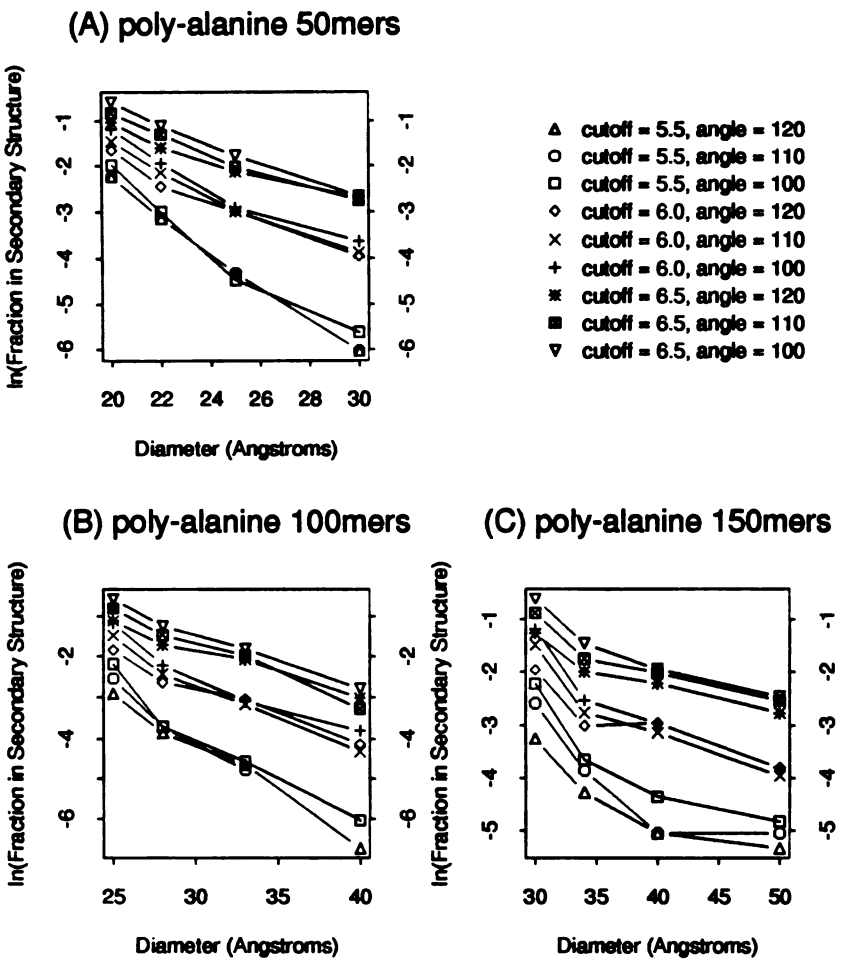
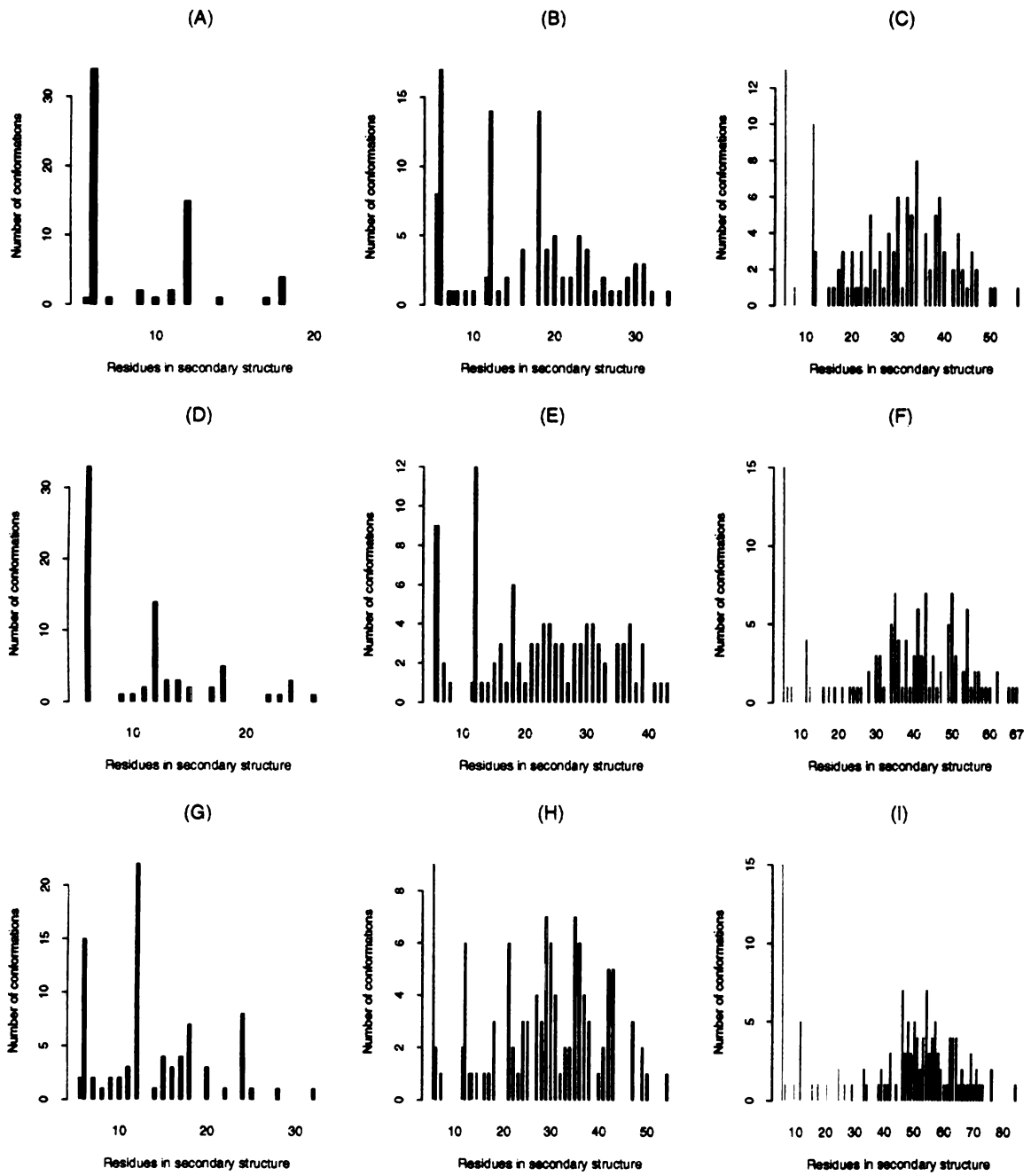


Figure 3.9: Amount of total secondary structure (helix + sheet) as a function of compactness as in figures 5 & 6, except y-axis is plotted on a logarithmic scale. (A) Poly-alanine 50mers (B) Poly-alanine 100mers (C) Poly-alanine 150mers

Figure 3.10 (next page): Histograms showing the number of poly-alanine 100mer conformations as a function the number of residues in secondary structure. The blue bars correspond to compact poly-alanine chains (25A sphere). The pink bars correspond to open poly-alanine chains (40A sphere). Each panel represents a different criteria set used with the TC definition of secondary structure. The criteria used are as follows: (A) contact cutoff = 5.5A, bond angle cutoff = 120 degrees, (B) contact cutoff = 5.5A, bond angle cutoff = 110 degrees, (C) contact cutoff = 5.5A, bond angle cutoff = 100 degrees, (D) contact cutoff = 6.0A, bond angle cutoff = 120 degrees, (E) contact cutoff = 6.0A, bond angle cutoff = 110 degrees, (F) contact cutoff = 6.0A, bond angle cutoff = 100 degrees, (G) contact cutoff = 6.5A, bond angle cutoff = 120 degrees, (H) contact cutoff = 6.5A, bond angle cutoff = 110 degrees, (I) contact cutoff = 6.5A, bond angle cutoff = 100 degrees.

From these data, we can estimate the free energy of stabilization of secondary structures by compactness. We define a structural unit as 6 residues since this is the minimal requirement of the TC criterion for defining a helix or sheet. The calculation is performed by dividing the fraction of compact conformations with N to N+5 residues in secondary structure by the fraction of open conformations with N to N+5 residues in secondary structure. The logarithm of this ratio yields an enhancement factor in units of kT. Note that when there are no conformations with N to N+5 residues in secondary structure, an enhancement factor is undefinable. Hence our statistics are limited since the sets of compact poly-alanine chains represent only a small sampling of all possible compact states. Nonetheless, the enhancement factors for poly-alanine 50mers as a function of the number of secondary structural units are shown in figure 3.11: each *additional* 6-mer structural unit of secondary structure is stabilized (favored) by compactness by about 2 kT. This estimate is largely independent of the identifying



criteria used.

Energy minimization

So far we have described all-heavy-atom poly-alanine chains that are only under the influence of the compactness constraint imposed by a constraining radius. The only constraints are that (1) no two atoms may occupy the same space (i.e., excluded volume) and (2) the entire chain conformation must be configured within the bounds imposed by a constraining radius. No other biases or energies have been included. The results described above show that strict criteria do not identify very much structure in these compact poly-alanine chains. The helices and sheets in these constrained poly-alanines do not look very protein-like. That is, compactness is not a force that specifically drives the formation of α-helices and β-sheets, but rather it stabilizes broad classes of conformations that include helices and sheets as subsets. Compactness acts to favor certain topological repeats, such as (i, i+3) contacts, but a considerable range of bond angles and geometries are consistent with this. The stabilization of secondary structures afforded by compactness can be viewed in the way that diffusion can be viewed as a driving force. It is not a specific pair interaction; it is a global property of an ensemble. This driving force is not structurally specific, like hydrogen bonds or other pair interactions are. We and others [13, 27] believe that both types of interaction -- the stabilization afforded by compactness, and the structural specificity afforded by hydrogen bonding and local propensities -- are required to lock in structures as specific as the α-helices and β-sheets observed in real proteins. Compactness appears to give stability, but not conformational specificity, to secondary structures in globular proteins.

Our results show that compactness increases the amount of ordered structure in compact polymers, but how far are the conformations from energy minima of more realistic secondary structures? The distribution of ϕ/ψ angles in the compact poly-alanine chains deviate substantially from the distribution of dihedral angles observed in real proteins. The degree to which a chain can move is severely limited since the

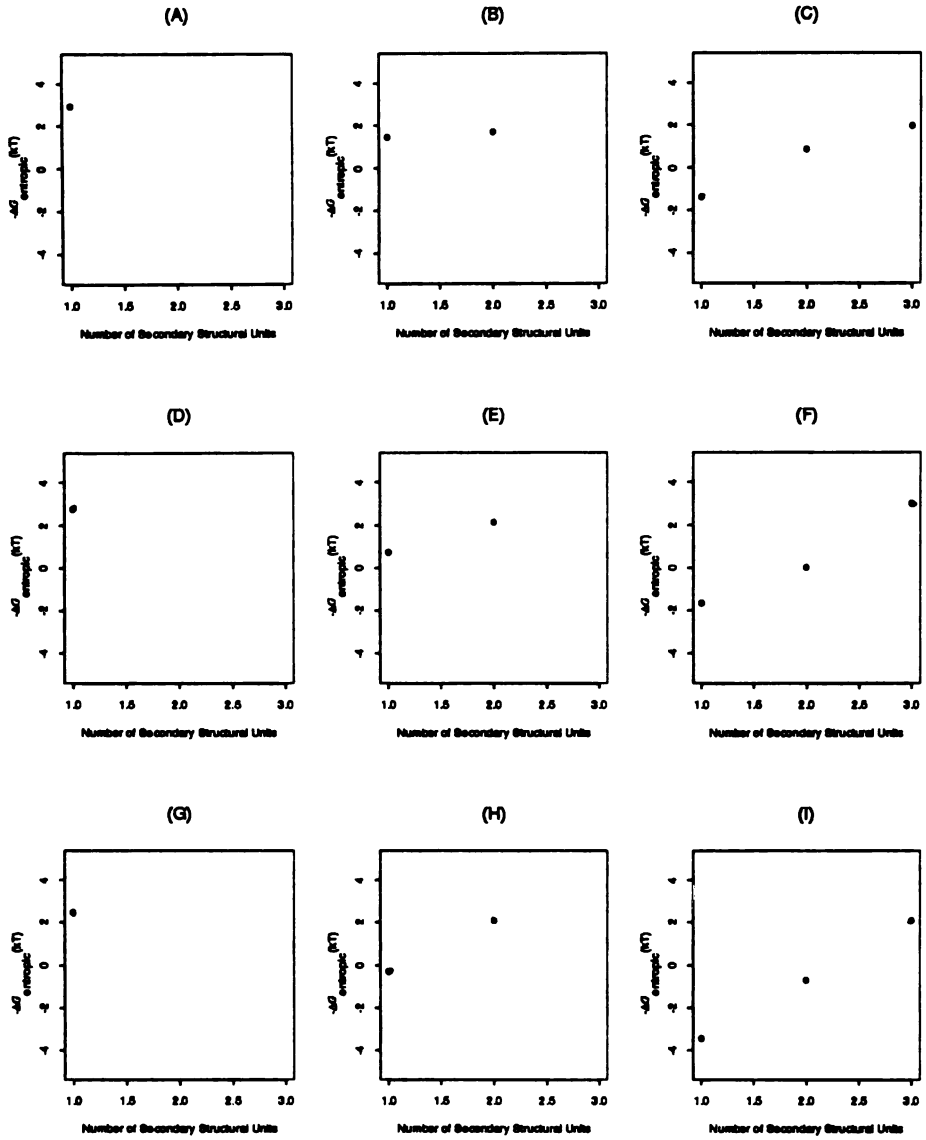


Figure 3.11: Enhancement factors for poly-alanine 50mer conformations transferred from an open ensemble to a compact one as a function of 6 residue secondary structural units (i.e., "1" implies 6 residues in secondary structure, "2" implies 7 - 12 residues, etc). Each panel represents a different criteria set which is given in the legend to figure 3.10.

excluded volume effect is very strong for the compact chains. We now ask whether a small perturbation of random compact poly-alanine conformations by the AMBER force field is sufficient to induce the poly-alanine structures to look more like proteins.

The poly-alanine chains that were generated by distance geometry were subsequently energy minimized using the AMBER potential. In addition to using the unmodified AMBER potential, the force field was modified in several ways to see how specific restraints would alter the ϕ/ψ map of the minimized poly-alanine chains. We tested four types of energy minimization: (1) the unmodified AMBER potential, (2) increasing the strength of the hydrogen bond, (3) adding a torsion angle force toward ϕ/ψ minima observed in alanine dipeptide simulations, and (4) adding constraints to favor the formation of α-helix-like hydrogen bonds. The results of these minimizations are presented below.

(1) Figures 3.12 & B show the ϕ / ψ maps for sets of 10 compact poly-alanine chains before and after 10,000 cycles of conjugate gradient minimization using the unmodified AMBER potential. The minimization procedure makes the ϕ/ψ distribution more protein-like in several ways. The ϕ/ψ angles in the α_R region becomes slightly more concentrated. The strip of torsion angles between $\phi \approx 60 - 80$ becomes less continuous. In general, the overall distribution of torsion angles is less diffuse and more focused into distinct minima.

Nevertheless differences remain between the energy minimized ϕ/ψ distribution shown in figure 3.12B and the real protein distribution. Since any energy minimization strategy seeks the *nearest* energy minimum, the refinement found only the closest ϕ/ψ combinations which would lower the overall energy of the system. For example, the map shows an increased concentration of (ϕ,ψ) angles near $(-70,60)$ and $(65,-60)$ which correspond to the formation of hydrogen bonds between O_{i-1} - HN_{i+1}. These regions

have been identified as energy minima in alanine dipeptide model studies in vacuum and in solution [28-30]. In addition, the concentration of dihedral angles in the α_L region ($\phi,\psi \approx 55,45$) is increased.

(2) We increased the strength of the hydrogen bond by factors of 1.50, 2.00 and 3.00. In each case, the ϕ/ψ distributions are more diffuse (for example, see figure 3.12C). There is a significant increase in the population of the β-sheet region of the ϕ/ψ map. Some ϕ/ψ angles, however, populate regions of the map which are not sampled by real proteins and several of the distances between adjacent C_α atoms become closer than the 3.8 A expected for trans-peptide bonds. Hence, increasing the strength of the hydrogen bond beyond its physical value can result in rather severe distortion of the standard peptide bond geometry.

(3) Next, we introduced a torsion angle forcing potential. Molecular dynamics simulations of alanine dipeptides have mapped out ϕ and ψ angles which correspond to local energy minima [26]. They indicate that two primary free energy minima exist at (ϕ, ψ) ≈ (-110, 120) and (- 120, -40) which correspond to the β region and α_R region respectively. Note that the "helix" ϕ minimum is offset by 50 degrees (to -120) from the ideal value for an α-helix. In the present work, we added forcing potentials in the form of an extra harmonic energy term to shift the dihedral angles ϕ and ψ to these values. The magnitude of the forcing potential ranges from 1.0 to 5.0 kcal/ rad^2 for the β region and from 0.4 to 2.0 kcal / rad^2 for the α_R region.

Figure 12D shows an example of energy minimizing poly-alanine chains subject to the torsion angle forcing term. There was a noticeable shift of torsion angles toward the α_R region of the ϕ/ψ map. Even though the force constants favoring the β region were larger than for the α_R region, there was little enhancement of ϕ/ψ angles in the extended β region.

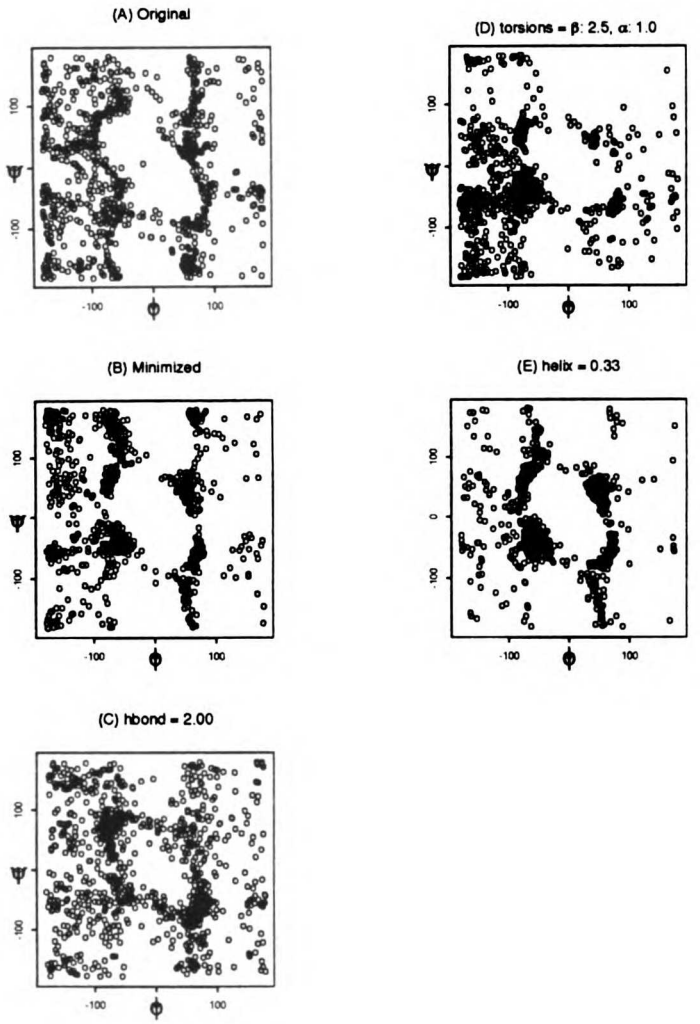


Figure 3.12: ϕ/ψ distributions of 10 near maximally compact poly-alanine 100mers. Results for 50mers and 150mers were essentially identical. (A) un-minimized chains, (B) minimization, (C) minimization with hydrogen bond strength increased by a factor of 2.0, (D) minimization in which torsion angle forcing terms were added to shift (ϕ/ψ) angles to (-110, 120) and to (-120, -40) with force constants set to 2.5 and 1.0 kcal / rad^2 respectively, (E) minimization in which a helix hydrogen bond term was added to force the formation of i to i+4 hydrogen bonds with k_H (see text) set to 0.33 kcal / A^2.

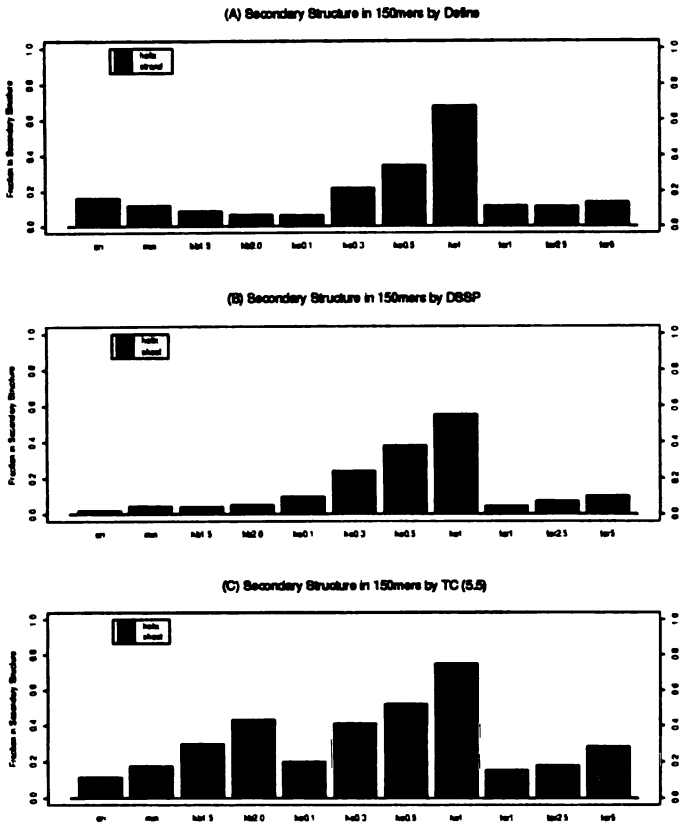


Figure 3.13: Histogram showing absolute amounts of secondary structure in poly-alanine 150mers after various minimization strategies. (A) Secondary structure determined by **Define**. (B) Secondary structure determined by **DSSP**. (C) Secondary structure determined by **TC** with 5.5A cutoff. ori: original conformation, min: AMBER minimization, hb1.5: hydrogen bond strength scaled by 1.5, hb2.0: hydrogen bonds scaled by 2.0, he0.1: helix term with force constant, k_H, set to 0.1 kcal / mol A^2, he0.3: k_H = 0.33, he0.5: k_H = 0.5, he1; k_H = 1.0, tor1: torsion angles terms for β and α regions set to 1.0 and 0.4 kcal / rad^2, tor2.5: torsion terms with force constants equal to 2.5 and 1.0, tor5: torsion terms with force constants equal to 5.0 and 2.0.

(4) Finally, a term was added to force conformations toward the formation of hydrogen bonds between O_i and HN_{i+4}. The extra energy term was a flat bottomed, skewed biharmonic function as implemented within Discover version 2.9 by Biosym Technologies. The form of the term is given by:

$$E_{constrain}(r) = \begin{cases} E_{L,max} + f_{\max}\,(r_{L,max} - r) & r < r_{L,max} \\ k_L\,(r - r_L)^2 & r_{L,max} < r < r_L \\ 0 & r_L < r < r_U \\ k_U\,(r - r_U)^2 & r_U < r < r_{U,max} \\ E_{U,max} + f_{\max}\,(r - r_{U,max}) & r_{U,max} < r \end{cases} \qquad (3.10)$$

Here, the value of r_L was 1.7 A. k_L was fixed at 5.0 kcal / mol A^2 to prevent the atoms from getting too close to one another. r_U was taken to be 2.5 A. Values of k_U ranged from 0.10 to 2 kcal / mol A^2. $r_{L,max}$ and $r_{U,max}$ were selected such that:

$$f_{\max} = 10.0\ kcal\ mol^{-1}A^{-2} = 2k_L\,(r_L - r_{L,max}) = 2k_U\,(r_{U,max} - r_U). \qquad (3.11)$$

$E_{L,max}$ and $E_{U,max}$ were given by:

$$E_{L,max} = k_L\,(r_{L,max} - r_L)^2 \text{ and } E_{U,max} = k_U\,(r_{U,max} - r_U)^2 \qquad (3.12)$$

Figure 12E shows that there is an increase in the population of dihedral angles in the α_R and α_L regions even for very small force constants. The results indicate that both the α_R and α_L regions of the ϕ/ψ map are consistent with the formation of helical hydrogen bonds. The observation of more ϕ/ψ angles in the α_L region of the poly-alanine chains relative to real proteins is due to the fact that the starting (unminimized) poly-alanine conformations have ϕ/ψ angles which populate sterically allowable, but energetically unfavorable, regions on the right hand side of the ϕ/ψ map. When subjected to the force field, the ϕ/ψ angles on the right hand side of the map move toward the closest minimum

(i.e., α_L).

Secondary structure in energy minimized poly-alanine chains

Figure 3.13 shows how introducing the energy perturbations affects the amount of secondary structure in the confined poly-alanine chains. Minimization alone using the unmodified AMBER potential slightly enhances the amount of secondary structure in the poly-alanine chains. According to Define, the strand content is lower in the minimized structures. Strengthening local hydrogen bonds reduces the chain extension. On the other hand, according to both DSSP and TC, energy minimization increases the sheet content. The TC criterion also detected an increase in helix content. Increasing the strength of the hydrogen bond slightly increases the helix content according to both DSSP and TC. The TC criterion also detects a large increase in sheet content. Part of the increased sheet content was due to severe distortions as the poly-alanines become very compact from the strong hydrogen bonds.

By all secondary structure criteria, the amount of α-helix increases after adding either the torsion angle term or the α-helical hydrogen bond force, even when the force constants are small (e.g. 0.1 kcal / mol A^2). As discussed above, when the helical hydrogen bond force is added, many ϕ/ψ angles populate the α_L region of the ϕ/ψ map. This means that left handed helices can form and may be identified as α-helix by secondary structure detection methods which are not sensitive to handedness such as Define and TC. DSSP, however, is sensitive to handedness and parallels the results obtained from Define and TC. Thus sterically constrained compact poly-alanine conformations require only small perturbations to reach α-helices. (We show below that these perturbations are small.)

It is much more difficult, however, to force the formation of β-sheets. Simple energy minimization may not introduce sufficient perturbation to induce sheets. Sheets

require coordination of different strands to come together, whereas helices appear to require only local readjustments that occur readily with small energetic perturbations.

Structural similarity of energy minimized chains

Here we show that energy minimization does not cause large perturbations of the poly-alanine chain conformations. We compared pairwise all conformers of five poly-alanine chains which were energy minimized using the strategies above. For each chain, there are 11 conformational variations: the original plus the result of 10 different energy minimization runs. Structural similarity was evaluated using CONGENEAL [24]. The pairwise comparison data was then used to construct a relatedness tree using hierarchical clustering. Figure 3.14 shows that after minimization, each poly-alanine chain falls within its own "family" of structures. Even when the forcing potentials are very large, the perturbed structures still cluster with their original poly-alanine parent. This indicates that after energy minimization, each poly-alanine chain retains enough of its original chain fold to be classified within its proper family. Also by molecular graphics we observed that the placements of turns in these structures are relatively unchanged. Even with a large helical force, for example, the chains never straighten out into a long helix.

When open chains were energy minimized using the strategies described above, the resulting conformations were only distantly related to the original starting conformation. That is, energy perturbations on open, relatively unconstrained chains led to large changes in the overall fold of the original structure.

Comparison with other studies

Despite the differences in methodology and model, this study is in general agreement with the studies of Gregoret and Cohen, Hao et al., Socci et al., and Hunt et al., in several main conclusions. First, it shows that packing is not structurally specific:

there is considerable conformational diversity. By strict criteria, only a small fraction of the stabilized secondary structures are recognizable α- helices and β-sheets. Compactness cannot account for why α-helices are more prevalent than 3_{10} helices, but it can account for why helices in general are more stable in compact conformations than in open conformations. Second, we find that only small energetic perturbations from the confined poly-alanines are required to give good α-helical conformations, although larger perturbations appear to be required to get good β-sheets.

Third, consistent with the earlier lattice studies of Chan and Dill [5-7], most of these studies show that compactness enhances secondary structure, and hence it provides a driving force. However, Kolinski and Skolnick [9] appear to disagree with this view. They say: "Thus... the secondary structure seen in the folded state is predominantly the result of short range interactions, or conformational propensities, which are more or less in accord with packing requirements in the dense globular state". They make two arguments. First, they note that the distribution of distances between residues i and i+3 in poly-valines look very similar for both compact and random coil ensembles when only the hydrophobic interaction is used. In addition, their distributions also resembled the distribution at low temperatures when only local (i.e., $|i - j| < 7$) interactions were included. Observing no differences in these distributions, they concluded that secondary structure does not come from compactness.

Second, their poly-valine chains collapsed into compact β- sheet-like globular states when local conformational preferences, hydrogen bonds, and nonlocal interactions were included. When local conformational preferences were left out, the poly-valine chains collapsed into compact states which were highly helical. When only the nonlocal interaction is used, the poly-valine chains adopted compact, "disordered" conformations. Since they used a strict criterion of secondary structure based on hydrogen bonding patterns, it is not surprising that they did not see any (hydrogen-bond-

based) secondary structure when only non-local interactions were used. This is consistent with our present results. That Kolinski and Skolnick could generate more secondary structures in compact states using more complex potentials does not prove that compactness is not important. In our view, compactness in their study already effectively eliminates the large number of non-compact conformations. The compact conformations which remain have a larger proportion of residues which can participate in secondary structure. In order to test the hypothesis that compactness induces secondary structure within the context of in their model, Kolinski and Skolnick might have asked how much local and hydrogen bond driving forces are necessary to get native-like secondary structures in the presence and absence of compactness. We believe compactness will reduce the driving force necessary to achieve native-like structures.

Finally, we disagree with a recent suggestion by Karplus and Shakhnovich that the hypothesis of compactness enhancement of secondary structure [6, 7] contradicts Flory's Theorem [31]. Flory's Theorem postulates that the spatial distribution of monomers of an individual polymer chain in a dense multiple-chain polymer melt would resemble that of a random flight [32, 33]. The theorem does not address the conformations of single compact chains. In addition, the Flory theorem addresses only distributions of monomer pairs, and not the multiple monomer correlations in secondary structures. Hence the current results are not inconsistent with the Flory theorem.

4. Conclusions

We have modelled proteins as all-heavy-atom poly-alanine chains of lengths 50, 100, and 150 monomers. They have been configured randomly, subject only to: (1) confinement to different radii of gyration by distance geometry and (2) steric constraints. No local propensities, energies, or other biases have been explicitly included.

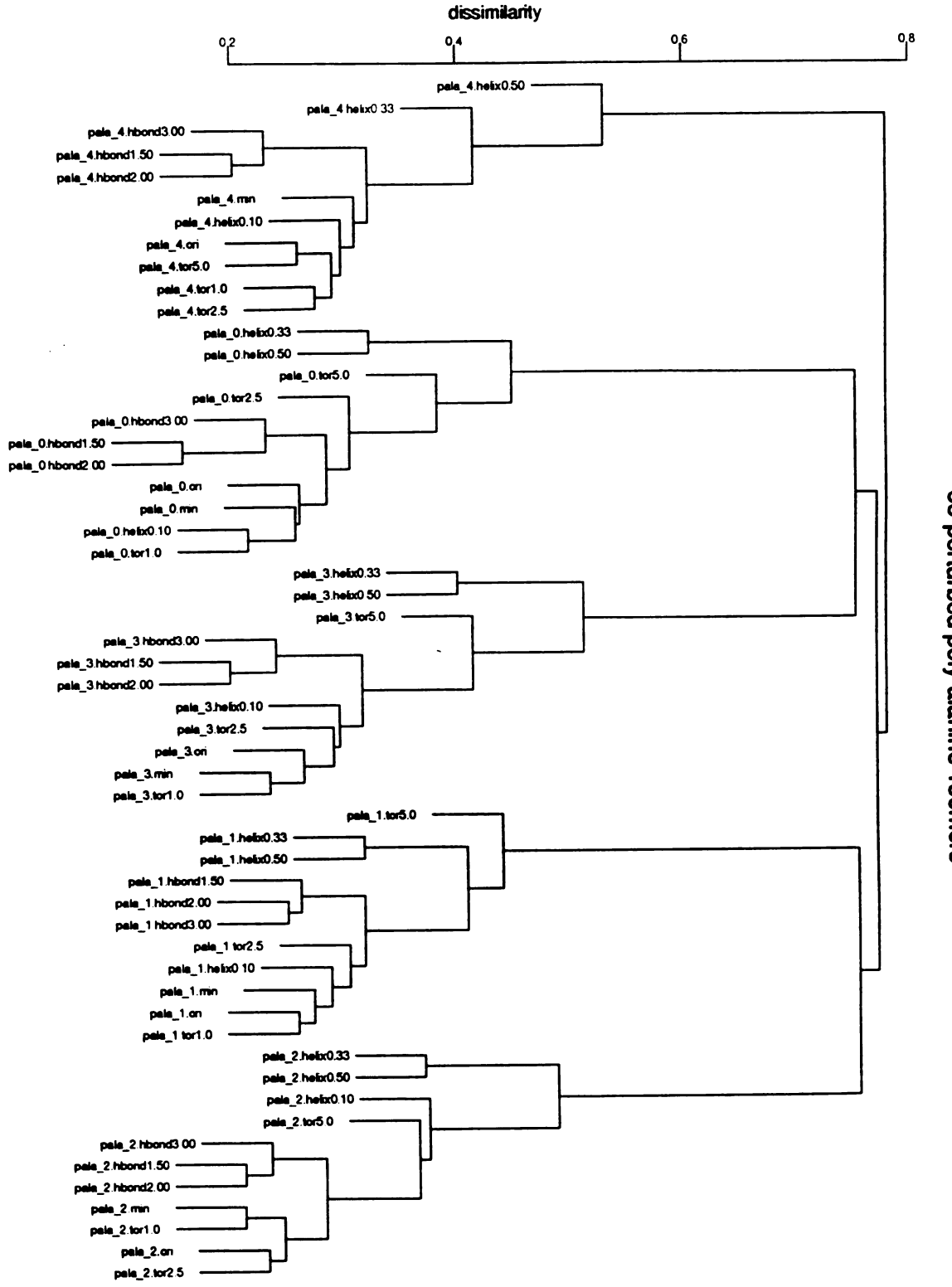


Figure 3.14: Tree showing interrelatedness of energy minimized poly-alanine 150mers. Each poly-alanine chain clusters with its "parent" structure even when forcing potentials are large.

We find that compactness enhances secondary structure by several different criteria of identifying helices and sheets. The total amount of secondary structure observed depends strongly on the criteria used and can vary over a range from nearly zero to 50% for the same chain conformation. Our results agree with other studies that have used strict criteria in that the secondary structures induced from compactness alone are neither protein-like nor as narrow a class of structures as α-helices and β-sheets [9-13]. Our analysis shows, however, that despite wide variation in the absolute number of residues in secondary structure, the stabilization free energy of secondary structures provided by compactness is essentially *independent* of the criterion used. Calculation of the stabilization free energy realized from compactness is estimated to be on the order of 2 kT per secondary structural unit.

Several energy minimization strategies using the AMBER potential are used to determine how far the maximally compact poly-alanine chains are from realistic energy minima. We find that small energetic perturbations to nearby local minima increased the number of dihedral angles in regions of ϕ/ψ space which are commonly observed in real proteins and nudged helices to become α-helices.

Compactness appears to be a structurally nonspecific entropic force that lowers the overall conformational free energy for a large class of helix-like and strand-like structures, among which are the α-helices and β-strands that are specific to peptide backbones. We believe that other polymers could also be driven by compactness to adopt helical and sheet structures. But the microscopic geometric details would be dictated by their preferred backbone conformations. Many crystal structures of synthetic polymers indeed adopt helical or planar zig-zag conformations [34]. The stabilization that results from compactness is due to the vast reduction in the number of conformations of the chain that are accessible in compact states, due to excluded volume. This reduction is a process in which a large fraction of the remaining conformations contain helix-like

and strand-like elements which are capable of filling space more densely than non-repeating conformations can.

Acknowledgements

We thank Sarina Bromberg and Nathan Hunt helpful discussions, Hunt et al. and Socci et al. for making their manuscripts available prior to publication, and the NIH for financial support (Grant numbers GM-34993 for the UCSF group and GM-38221 for T. H.).

References

[1] M. Levitt and J. Greer, "Automatic Identification of Secondary Structure in Globular Proteins," *Journal of Molecular Biology*, 114(2):181-239, 1977.

[2] W. Kabsch and C. Sander, "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features," *Biopolymers*, 22:2577 - 2637, 1983.

[3] K. A. Dill, "Dominant Forces in Protein Folding," *Biochemistry*, 29(31):7133 - 7155, 1990.

[4] T. R. Sosnick, L. Mayne, R. Hiller, and S. W. Englander, "The Barriers in Protein Folding," *Nature: Structural Biology*, 1(3):149 - 156, 1994.

[5] H. S. Chan and K. A. Dill, "Compact Polymers," *Macromolecules*, 22:4559 - 4573, 1989.

[6] H. S. Chan and K. A. Dill, "Origins of Structure in Globular Proteins," *Proceedings of the National Academy of Science, USA*, 87:6388 - 6392, 1990.

[7] H. S. Chan and K. A. Dill, "The Effects of Internal Constraints on the Configurations of Chain Molecules," *Journal of Chemical Physics*, 92(5):3118 - 3135, 1990.

[8] E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng, in: *Crystallographic Databases - Information Content, Software Systems, Scientific Applications*, F. H. Allen, G.Bergerhoff, and R. Seivers, ed., p. 107 - 132, Data Commission of the Int'l Union of Crystallography, Bonn/Cambridge/Chester, 1987.

[9] A. Kolinski and J. Skolnick, "Discretized Model of Proteins I. Monte Carlo Study of Cooperativity in Homopolymers," *Journal of Chemical Physics*, 97(12):9412 - 9426, 1992.

[10] L. M. Gregoret and F. E. Cohen, "Protein Folding: Effect of Packing Density on Chain Conformation," *Journal of Molecular Biology*, 219:109 - 122, 1991.

[11] M. H. Hao, S. Rackovsky, A. Liwo, M. R. Pincus, and H. A. Scheraga, "Effects of Compact Volume and Chain Stiffness on the Conformations of Native Proteins," *Proceedings of the National Academy of Science, USA*, 89:6614 - 6618, 1992.

[12] N. D. Socci, W. S. Bialek, and J. N. Onuchic, "Properties and Origins of Protein Secondary Structure," *Physical Review E.*, 49:3440 - 3443, 1994.

[13] N. G. Hunt, L. M. Gregoret, and F. E. Cohen, "The Origins of Protein Secondary Structure: Effects of Packing Density and Hydrogen Bonding Studied by a Fast Conformational Search," *in press, Journal of Molecular Biology*, 1994.

[14] T. F. Havel, "An Evaluation of Computational Strategies for use in the Determination of Protein Structure from Distance Constraints Obtained by Nuclear Magnetic Resonance," *Progress in Biophysics and Molecular Biology*, 56(1):43 - 78, 1991.

[15] T. F. Havel, "The Sampling Properties of some Distance Geometry Algorithms Applied to Unconstrained Computed Conformations," *Biopolymers*, 29(12-13):1565 - 1585, 1990.

[16] H. S. Chan and K. A. Dill, "Sequence Space Soup of Proteins and Copolymers," *Journal of Chemical Physics*, 95(5):3775 - 3787, 1991.

[17] F. M. Richards, "The Interpretation of Protein Structures: Total volume, Group volume Distributions and Packing Density." *Journal of Molecular Biology*, 82(1):1 - 14, 1974.

[18] C. Chothia, "Structural Invariants in Protein Folding," *Nature*, 254:304 - 308, 1975.

[19] V. N. Maiorov and G. M. Crippen, "Contact Potential that Recognises the Correct Folding of Globular Proteins," *Journal of Molecular Biology*, 227:876 - 888, 1992.

[20] F. M. Richards and C. E. Kundrot, "Identification of Structural Motifs from Protein Coordinate Data: Secondary Structure and First-level Supersecondary Structure," *Proteins*, 3(2):71 - 84, 1988.

[21] S. B. Dev, "Quantitative Prediction of Protein Secondary Structure - Where is the Lacuna?" *Journal of Biological Physics*, 15:57 - 61, 1987.

[22] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, Jr.., and P. Weiner, "A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins," *Journal of the American Chemical Society*, 106:765 - 784, 1984.

[23] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case, "An All Atom Forcefield for Simulations of Proteins and Nucleic Acids," *Journal of Computational Chemistry*, 7:230 - 252, 1986.

[24] D. P. Yee and K. A. Dill, "Families and the Structural Relatedness among Globular Proteins," *Protein Science*, 2:884 - 899, 1993.

[25] G. N. Ramachandran and V. Sasisekharan, "Conformation of Polypeptides and Proteins," *Advances in Protein Chemistry*, 23:283 - 483, 1968.

[26] A. G. Anderson and J. Hermans, "Microfolding: Conformational Probability Map for the Alanine Dipeptide in Water From Molecular Dynamics Simulations," *Proteins*, 3:262 - 265, 1988.

[27] B. Honig, K. Sharp, and A. Yang, "Macroscopic Models of Aqueous Solutions: Biological and Chemical Applications," *Journal of Physical Chemistry*, 97(6):1101 - 1109, 1993.

[28] B. M. Pettitt and M. Karplus, "Conformational Free Energy of Hydration for the Alanine Dipeptide: Thermodynamic Analysis," *Journal of Chemical Physics*, 92:3994 - 3997, 1988.

[29] T. Head-Gordon, M. Head-Gordon, M. J. Frisch, C. L. Brooks, III., and J. A. Pople, "Theoretical Study of Blocked Glycine and Alanine Peptide Analogues," *Journal of the American Chemical Society*, 113:5989 - 5997, 1991.

[30] D. J. Tobias and C. L. Brooks, III., "Theoretical Study of Blocked Glycine and Alanine Peptide Analogues," *Journal of Physical Chemistry*, 96:3864 - 3870, 1992.

[31] M. Karplus and E. Shakhnovich, "Protein Folding: Theoretical Studies of Thermodynamics and Dynamics," in: *Protein Folding*, T. E. Creighton, ed., p. 127 -195, Freeman, New York, 1992.

[32] P. J. Flory, "The Configuration of Real Polymer Chains," *Journal of Chemical Physics*, 17:303 - 310, 1949.

[33] P. G. de Gennes, *Scaling Concepts in Polymer Physics*, p. 54 - 61, Cornell University Press, Ithaca, 1979.

[34] H. Tadokoro, *Structure of Crystalline Polymers*, Wiley, New York, 1979.

Chapter 4

Mapping Protein Structures onto Lattices

Introduction

The work described in this chapter revolves around the first project I started under the direction of Professor Ken Dill. I will use this chapter as an opportunity to reflect upon problems of the past, and to tie together some insights and thoughts I have which were not incorporated into the previous chapters of this thesis.

The aim of the project was to develop a method to find the optimal lattice representation of a real polymer chain such as a protein because we were interested in determining how well a lattice model could represent a real protein structure. Is there, for example, a reasonable mapping of conformations in protein structure space onto conformations in cubic lattice space? To address this question, it was necessary to gain an understanding of how one could determine the "best" description of a real protein structure using a lower resolution lattice model representation.

The principal advantage of a lattice based model is the ability to rapidly perform exhaustive and systematic explorations of conformational space. By constraining monomer positions to lattice sites, the total number of possible chain configurations in a lattice model is dramatically reduced over a freely rotating chain. Despite the simplification, one assumes that the lattice model retains similar physical properties to real chains. By characterizing the ensemble properties of a lattice model, one can gain insights into the behavior of more complex systems. The notion that compactness stabilizes secondary structures in compact polymers, as described in chapter 3, was originally derived from lattice models [1].

Many groups have used lattice models to explore questions of protein structure and stability [1-11]. An issue that arises is the question of how well a lattice model represents

the structure of a real protein. What properties of real proteins can one expect a lattice model to capture? What properties should one not expect a lattice model to capture? Given a lattice-based protein structure prediction algorithm, how similar should the lattice model be to the structure of the real protein? The answer to that question depends upon what one defines as similar.

In this chapter, I will first describe work involving the development of a procedure for building cubic lattice representations of real protein structures. I will then discuss the problem of comparing models of proteins to real proteins, and conclude by introducing some methods for characterizing ensembles of conformations using bond vector correlations.

Error and Similarity

The first issue one must consider when thinking about the problem of mapping real protein structures onto lattices involves defining similarity. It is clear that a measure of similarity by which different configurations can be distinguished as being better or worse than a target structure is required. With respect to the mapping project, two general criteria are based upon geometric and topological considerations.

A geometric criterion is perhaps the most intuitive. Similarity is defined by evaluating how close the spatial path of a lattice chain matches the path of the real chain. The root-mean-square deviation between residues in a model relative to the α-carbon positions in a protein is a convenient geometric measure of similarity. Covell and Jernigan [11], for example, mapped real protein structures onto lattices using a least-squares procedure. They explored three types of cubic lattices: a simple cubic lattice, a body-centered cubic lattice, and a face-centered cubic lattice.

While a simple cubic lattice has only two bond angles (90^o and 180^o) and four torsion angles (0^o, 90^o, 180^o, 270^o), the face-centered cubic lattice provides four additional bond angles and sixteen additional torsion angles. The additional bond and torsion angles provided by the face-centered cubic lattice match the preferences of real protein rather well. By using a lattice model with a high degree of conformational flexibility, Covell and Jernigan were able to generate lattice mappings of real proteins with root-mean-square deviations of about 1.0-Å.

The ability of a face-centered cubic lattice to represent real protein structures with very small geometric deviations comes at some price. While the distance between adjacent α-carbon positions in real proteins is essentially constant at 3.8-Å, the face-centered cubic lattice model allows the distance between two connected residues to adopt values of 2.68, 3.8, and 4.65-Å. Therefore, the distance between two sequential residues in the lattice model can vary by up to 40% of its ideal length. The face-centered cubic lattice model also has a higher coordination number than a simple cubic lattice model; a high coordination number means that a lattice site may have many close neighbors. While this feature allows one to construct an accurate geometric representation of a protein, the size of conformational space for high coordination number models is too large to explore exhaustively. Finally, while a chain configured on a lattice may accurately reproduce the path of a real protein chain, the physical constraints on the lattice chain may differ considerably from the constraints on a real chain. For example, it is generally accepted that real proteins are close to maximally compact [12]. A lattice model which reproduces the exact chain geometry of a real protein, however, is almost certainly not maximally compact.

While it is possible to construct lattice models of proteins with a high degree of geometric accuracy, one sacrifices the accuracy of the model with respect to important physical properties such as compactness and bond lengths. An essential consideration,

therefore, is whether a simple model captures the properties one is most interested in modeling.

When considering how to map real proteins onto a simple cubic lattice, I opted to use a topological criterion of similarity. This choice seemed obvious because of the characteristics of the model. Since the chain was to be configured on a simple cubic lattice, the bond lengths between residues adjacent in sequence were constrained to be constant. Since real proteins are maximally compact, the lattice representation should also be highly compact. The dominant driving force for protein folding was assumed to be the hydrophobic interaction. Therefore, the model should reproduce as many native hydrophobic-hydrophobic interactions as possible. Since these interactions are relatively short ranged, the mapping problem becomes one of finding a lattice representation which reproduces as many of the spatially close intra-residue interactions present in the native protein as possible. An example of a similarity measure based on a topological criterion is CONGENEAL, described earlier in this dissertation.

Representation

Both real protein structures and lattice models of proteins are represented as weighted distance maps. The weighted distance map representation of structure was discussed in detail in chapter 2. Briefly, a weighted distance map is one triangle of an $N \times N$ matrix in which each matrix element, w_{ij}, represents a weight proportional to the importance of the interaction between residues i and j. In analogy to the distance dependence of non-bonded interactions in proteins (e.g., van der Waals, dipole-dipole, etc), w_{ij} is defined to be the distance between the α-carbon positions of residues i and j raised to the inverse sixth power. That is,

$$w_{ij} = d_{ij}^{-6}. \tag{4.1}$$

The topological representation is based on the assumption that residues which are close together in space are more important than residues which are widely separated in space. By this criterion, the best lattice model representation of a real protein is the one which minimizes the difference between the weighted distance map of the lattice model when compared to the weighted distance map of the real protein.

Model Generation

The simplest way to produce a lattice model of a real protein of length N might be to generate all possible N residue lattice-based conformations and find the one that best matches the real protein. Since the number of configurations grows exponentially with chain length, however, it is not possible to exhaustively compute all configurations for chain lengths approaching that of real proteins. Therefore, to address the problem of generating lattice mappings of real proteins, I developed a method to selectively and systematically explore conformational space. The method can produce a set of very good representations of the target protein, although the representations are not guaranteed to be globally optimal. The method involves the use of a "build-up" procedure in which the real protein is divided into short segments. Lattice representations of each chain segment are first sorted, then the best pieces are assembled together. The processes of selection and assembly are iterated until the entire protein is constructed. I first discuss a sequential build-up procedure, then describe an improved algorithm which is much more flexible and produces better lattice models.

Sequential Build-up Procedure

The first step of the sequential build-up strategy is to divide the protein into short 6 residue segments. Each real protein segment is then compared to an exhaustive set of 6 residue lattice configurations. On a three dimensional cubic lattice, there are exactly 92

unique 6 residue conformations. Lattice representations are grown sequentially from the C-terminus to the N-terminus. At the beginning of the process, the best lattice representations of residues 1 - 6 are combined with the best scoring representations of residues 7 - 12. Of the several thousand possible ways of constructing a 12 residue segment, only the 100 - 200 best chain configurations are saved. Each of the best models are then combined with the best lattice representations of residues 13 - 18, and the best 18 residue segments are saved and further grown into longer chains. The process is iterated until a complete lattice representation of the target protein is constructed.

Although it was possible to construct reasonable lattice models of real proteins, the sequential build-up strategy was limited. For proteins longer than about 60 residues, the C-terminal end of the lattice model would invariably represent the real protein better than the N-terminal end.

Non-sequential Build-up Procedure

Since there was no *a priori* reason to construct a lattice model sequentially, the next version of the build-up procedure was designed to treat a pool of lattice model fragments without requiring the fragments to be combined in any particular order. The program works in four stages: piece generation, pairwise combination, linker addition, and chain extension.

In the first stage, the protein is divided into segments. Minimally, the program can start by dividing the protein into short sequential segments as in the sequential build-up procedure described above. The segments, however, can be of arbitrary length and are not required to consist of sequential residues. For example, a segment could be based on an anti-parallel sheet consisting of multiple strands but lacking the residues between the strands. For most applications, the program first identifies helices and strands using an implementation of the secondary structure identification algorithm of Richards and

Kundrot [13]. Since most secondary structures are relatively short, it is possible to determine a small set of very good lattice representations for each segment. These segments provide the basis for constructing the complete protein.

In the second stage, all the segments are compared pairwise. If there are any overlapping residues between two segments, the program will attempt to merge them by determining all the possible rotations and translations which place the overlapping residues in mutually consistent positions and do not contain any excluded volume violations. If there are no overlapping residues between two segments, the program will determine if the segments contain residues which are known to be adjacent. For example, the program will combine a segment containing residues 1 - 6 with a segment containing residues 7 - 12 by finding all orientations in which residues 6 and 7 are adjacent. It is possible to add arbitrary constraints which allow segments with non-sequential residues to be combined. For example, if there is a disulphide bond between two residues in a protein and one wanted the lattice model to contain a contact between the 2 cysteine residues, one would add the two residues to the list of adjacent residues.

After any two segments have been combined, the program will look for unplaced residues whose positions are fully determined. For example, if residues 5 and 7 are assigned positions, there are only two possible positions for residue 6 on a simple cubic lattice. If one of the two positions is already occupied by another residue, then residue 6 will be assigned the remaining lattice site. If both positions are occupied, the segment is discarded as inconsistent.

When no more segments can be combined by pairwise combination and all fully determined positions are filled in, the segments are extended by a few residues and the best extended segments are saved. Chain extension is performed by adding linkers to each of the best segments where the linkers are the 92 unique lattice 6-mers. For reasons

of computational efficiency, only a small set of the best segments are saved.

Finally, the segments are grown at the ends by adding unassigned residues at both the C-terminal and N-terminal ends of each segment. This step is called chain extension and is necessary to fill in any residues which have not yet been placed by the earlier procedures. If a segment contains assigned positions for each residue in a protein, it is complete and written to a file. The remaining set of segments are iterated through the process again until all segments are either complete or shown to be inconsistent and discarded.

Results

The first test of the build-up procedure was to reconstruct a hypothetical two dimensional protein on a square lattice. The two dimensional protein was based on a map of the United States as shown in figure 4.1. The lattice model was constructed by building up a set of models from the west to the east. Using 8 residue segments, a series of complete models was generated in 6 assembly steps. Figure 1 also shows an example of a lattice representation of the United States. Despite being based on topological interactions, the geometry of the model clearly resembles the geometry of the United States. Note that the density of the lattice model is much lower in the west than in the east due to the larger relative size of the western states. The packing density of the ''real'' United States is much higher than that of the lattice model due to the constant size of a lattice site. This will be a problem when mapping real proteins to a regular lattice since amino acids, like the states, have size variations while the lattice sites do not.

The build-up procedure was used to construct lattice representations of three proteins: crambin (CRN), bovine pancreatic trypsin inhibitor (BPTI), and intestinal calcium binding protein (ICB). As discussed above, the best results were obtained using

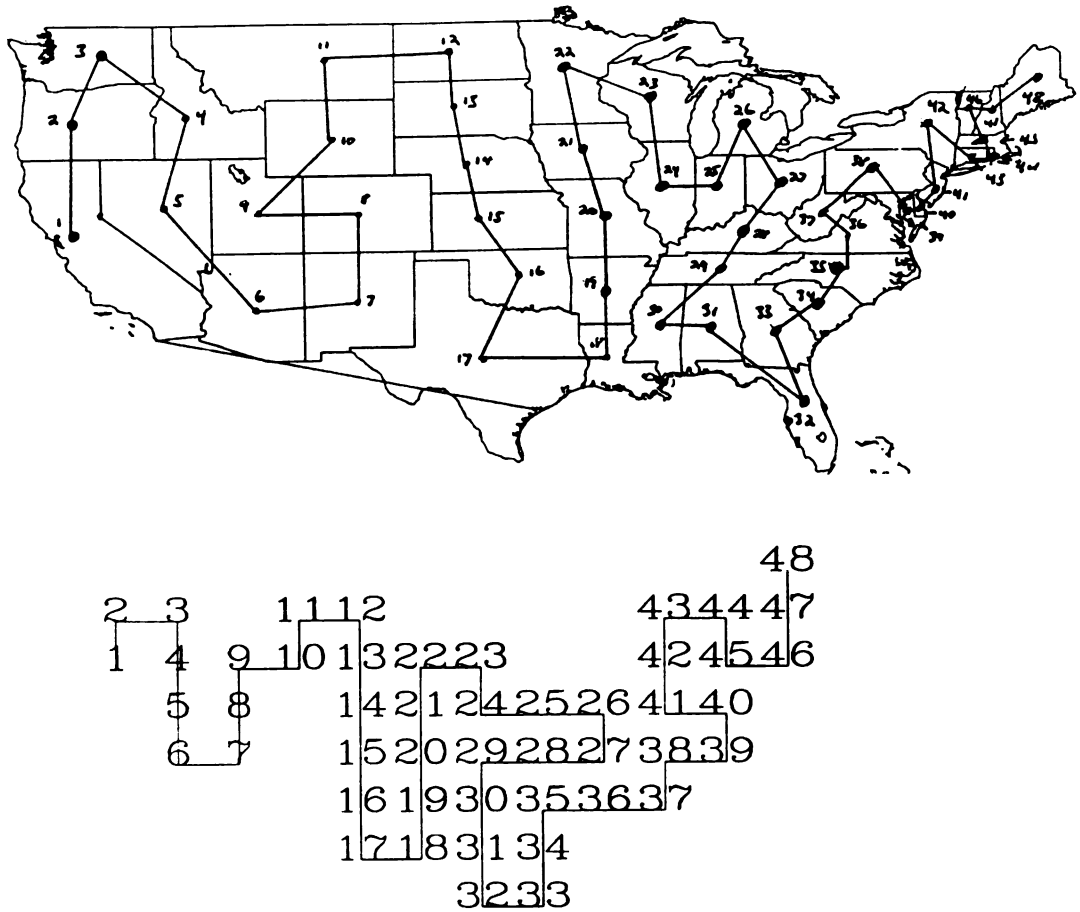


figure 4.1: (above) hypothetical 2 dimensional protein based on map of the United States. (below) Lattice model of protein using sequential build-up procedure.

the second version of the program which was not limited to constructing a lattice representation in a sequential order. In all cases, segments corresponding to secondary structural elements were constructed first. These segments were then combined and

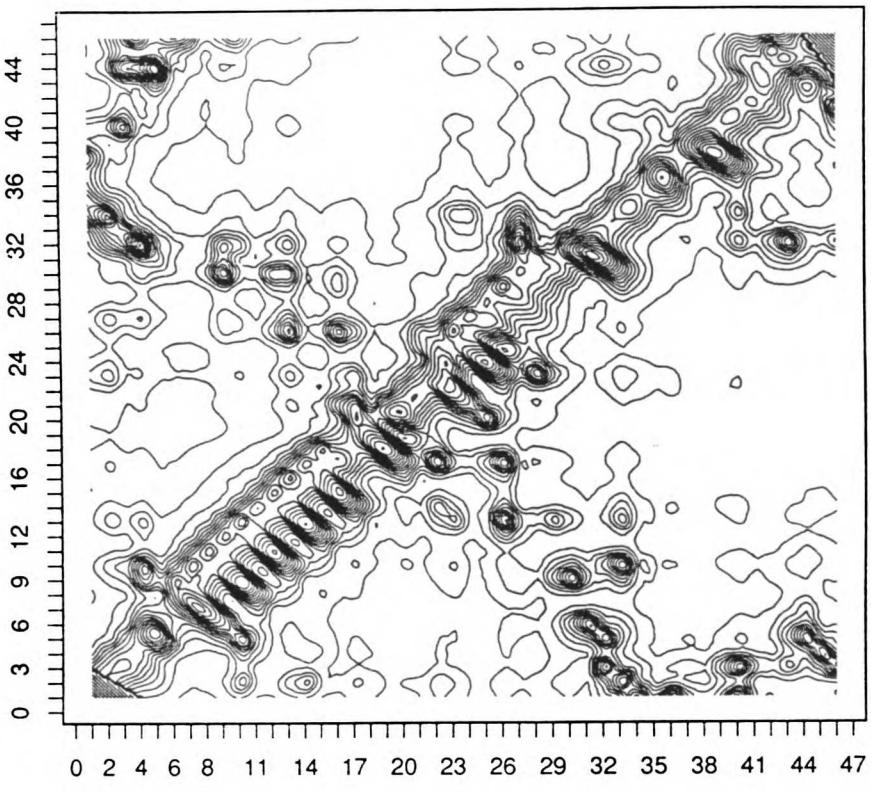


figure 4.2: Weighted distance map of crambin in upper triangle. Weighted distance map of a lattice representation of crambin in lower triangle.

grown to build a representation of the complete protein. Figures 4.2 - 4.3 show weighted distance maps of the three proteins.

It is apparent from the figures that the build-up procedure is somewhat successful in reproducing the intra-residues interactions present in the real protein. The general features are clearly present. Subtle differences in residue-to-residue separations, however, are not captured since the lattice model has less freedom to vary the distance between residues.

The geometries of the lattice representations are similarly reasonable. Figure 4.5 shows a lattice model of crambin adjacent to a α-carbon trace of real crambin. The

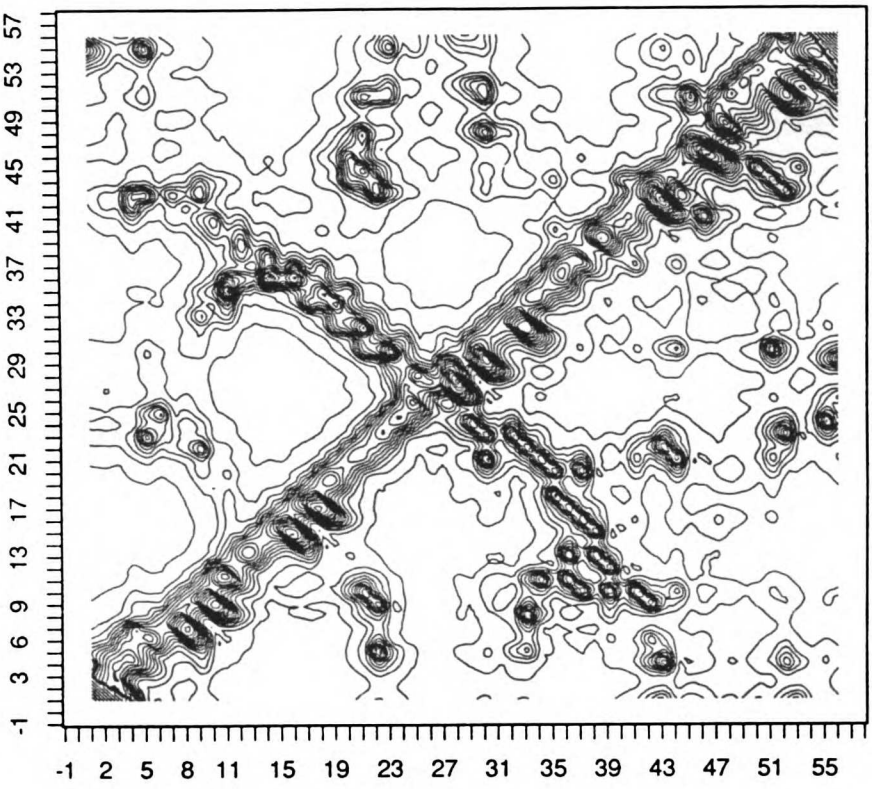


figure 4.3: Weighted distance map of bovine pancreatic trypsin inhibitor (BPTI) in upper triangle. Weighted distance map of a lattice representation of BPTI in lower triangle.

root-mean-square (RMS) deviation of the best crambin model relative to real crambin is about 3.5-Å. Covell and Jernigan [11] could generate simple cubic lattice models of crambin with RMS deviations as low as 2.7-Å. The difference in RMS deviations between Covell and Jernigan and the present work reflects the use of the topological criterion.

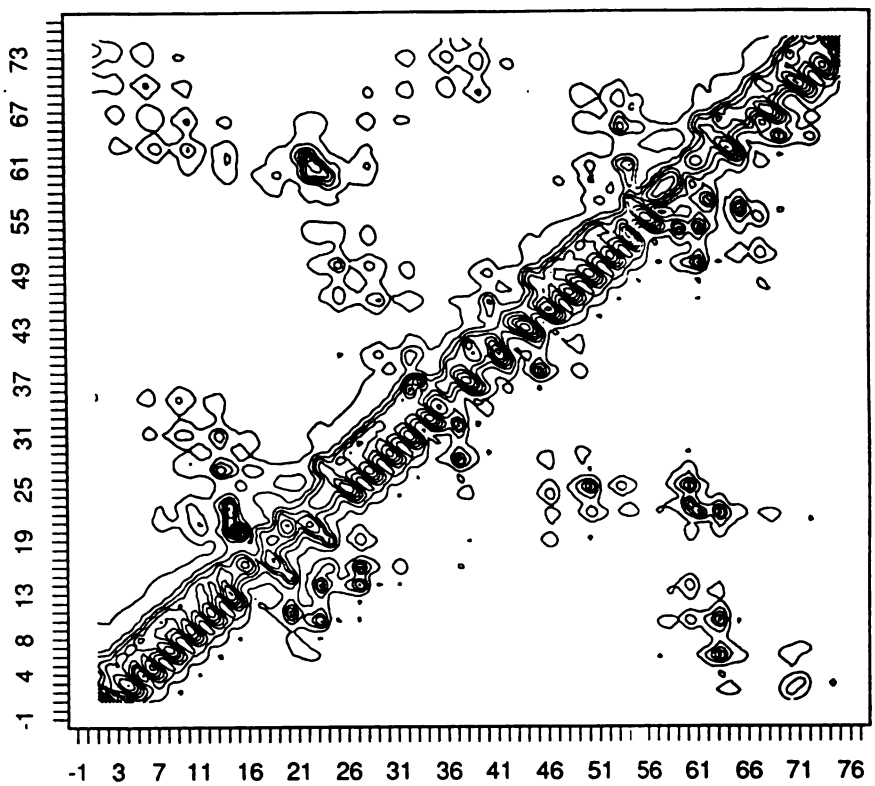


figure 4.4: Weighted distance map of intestinal calcium binding protein (ICB) in upper triangle. Weighted distance map of a lattice representation of ICB in lower triangle.

Discussion

Computational Limitations

While there are positive results for mapping small (under 100 amino acids) proteins onto a cubic lattice using the build-up procedure, generating models for larger proteins was problematic. The primary limitation was computation time. The size of conformational space increases exponentially with increasing chain length. Consequently, the computer time necessary to model large proteins becomes prohibitive. In addition to run time, generating models of large proteins is limited with respect to

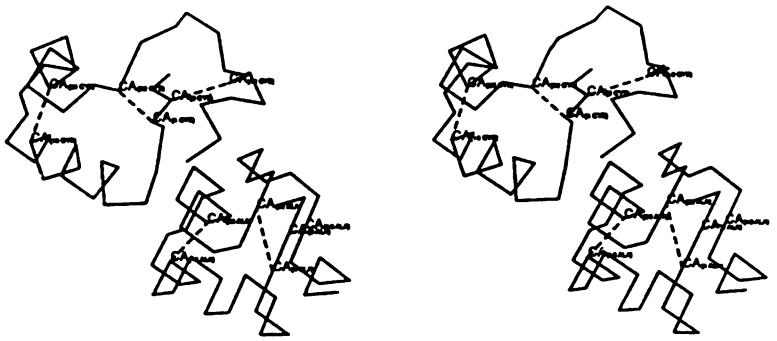


figure 4.5: Stereoview of a lattice representation of crambin adjacent to an α-carbon trace of crambin.

memory and disk space. Large proteins, for example, require the program to manipulate more segments than small proteins. Segments size also becomes much larger for large proteins. In addition, it is necessary to store more intermediate conformations to disk since more rounds of pairwise combination are required to generate a complete lattice model.

These limitations will become less important as computers get faster and the cost of memory and disk space goes down. Nonetheless, the number of possible conformations increases exponentially with chain length, so intelligent strategies must be employed to effectively explore a reasonable amount of conformational space. One strategy I used with some success was to examine the patterns of contacts in each segment for inconsistencies. Figure 3.1 shows that on a cubic lattice, some contacts are prohibited in context of certain patterns of contacts. By building a library of prohibited contact patterns, one can screen a set of lattice segments for the presence of inconsistent contacts. If a segment contains contacts which are known to be inconsistent, then it can be

discarded.

Compactness of Lattice Representations

Many folding studies using lattice models attempt to maximize close interactions between hydrophobic residues due to the dominant role the hydrophobic effect has on protein folding [14]. Since real proteins are nearly maximally compact [12], it is reasonable to expect a model of protein structure to be maximally compact. The decision to use a topological criterion for the construction of lattice representations of real proteins was based on the assumption that by reproducing as many closely contacting residues as possible, the resulting models would be compact. On one level, this assumption proved to be correct. Models built by optimizing topological similarity contain more nearest neighbor contacts than models built by geometric criteria.

The lattice models which reproduce the topology of real proteins, however, are not *maximally* compact. It is possible to determine the number of nearest neighbor contacts in a maximally compact cubic lattice chain [15]. The best lattice models of real proteins obtained by the build-up program contained about 75% of the maximal number of nearest neighbor contacts. There are several reasons why the lattice representations are not more compact. First, maximally compact lattice chains minimize surface area by adopting regular shapes (i.e., cubes). Real proteins, on the other hand, describe much more complex surfaces. Second, cubic lattice chains can pack more efficiently than real protein chains. Lattice chains can achieve a packing density of 1.0. Real proteins, however, pack no closer than close packed spheres with a packing density of about 0.76.

Lattice Chain Physics

Real proteins have restricted bond and torsion angles which are easily seen by examining the distribution of ϕ/ψ angles occupied by real proteins. By restricting a chain to lattice sites, the chains have severely restricted bond angles and torsion angles. In

effect, the lattice model imposes an infinitely strong potential in which there are only a few allowed bond and torsion angles. Polypeptide and lattice chains are similar in the sense that the conformational freedom of a chain is restricted by effective potentials. In the case of real proteins, these potentials originate from steric interactions and local energy minima. For lattice chains, the potentials originate from the requirement that the chain configure itself on lattice sites. Differences in the potentials between real and lattice chains make the problem of mapping real proteins onto lattices a difficult one. For example, a major difference between lattice chains and real polypeptide chains is chirality. Polypeptides consist of L-amino acids which are chiral. The chiral nature of proteins manifests itself in many ways. ϕ/ψ plots derived from real proteins are preferentially occupied on one side. α-helices are right-handed. β-sheets frequently have a left-handed twist. Lattice chains, on the other hand are achiral. There is no preference to form a right-handed over a left-handed helix, for example. Figure 4.5 shows that the lattice representation of an α-helix in crambin does not take the form of a right handed helix. The lattice chain simply places residues in a way that best mimics the pattern of contacting residues. Since the lattice chain is achiral, there is no constraint requiring it to adopt a handed conformation.

Protein Spectra

A useful way to distinguish the physical properties of ensembles of polymers is by examining correlations between residues in the chain. Many studies have used correlation functions to study conformational properties of polymers (for example, [16, 17]). Here, I will show plots of interresidue distance correlations and bond projection correlations for real proteins, maximally compact random poly-alanine chains, and compact chains configured on a simple cubic lattice. I call these plots protein spectra, and view them as an informative way to visualize certain properties of polymer

ensembles.

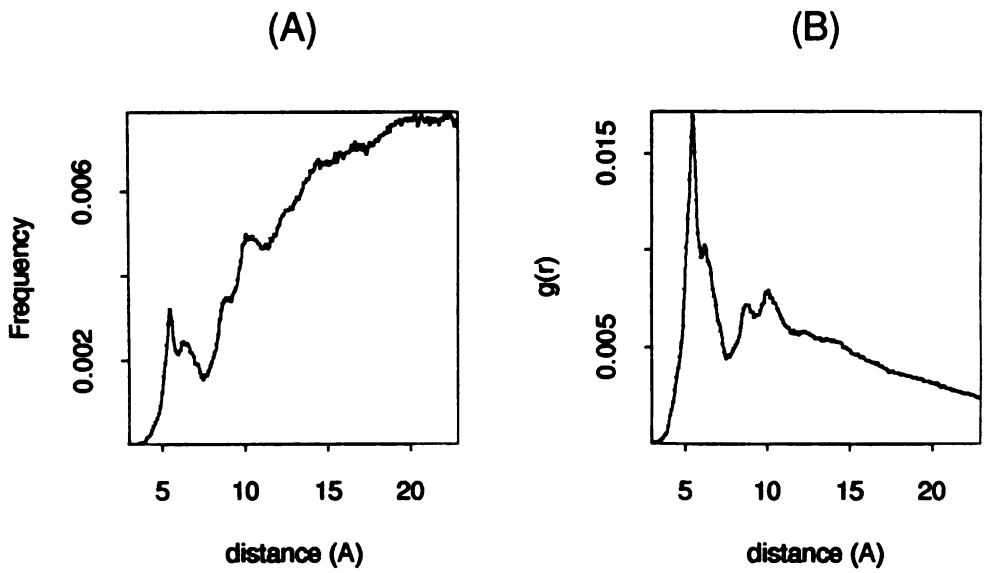


figure 4.6: (A) Histogram of distances between non-adjacent α-carbon atoms in set of protein structures derived from the PDB. (B) Distance correlation function of non-adjacent α-carbon atoms derived from proteins used in (A).

A distribution of separation distances between non-adjacent α-carbon residues in proteins is shown in figure 4.6A. The histogram shows a peak at about 5.4-Å which indicates the closest distance two non-adjacent residues come to one another. There is a minimum in the distribution at around 7.5-Å indicating that relatively few pairs of residues are separated by this distance. Other features of this distribution are obscured because of the increasing number of residue pairs separated by relatively large distances. One way to clarify this distribution is to plot a inter-residue distance correlation function, $g_{\alpha\alpha}(r)$. This function is given by

$$g_{\alpha\alpha}(r) = \frac{N(r)}{4\pi r^2 \Delta r}, \qquad (4.2)$$

where $N(r)$ is the number of residues which are a separated by a distance between r to $r + \Delta r$. Figure 4.6B shows the distance correlation function for the same real protein set used to generate figure 4.6A. This distance correlation plot normalizes the distribution to take into account the increasingly larger volume in which residues can exist at increasing distances. The normalization factor of $4\pi r^2 \Delta r$ sharpens the peaks and yields a characteristic spectra.

Much more detailed spectra can be obtained if one generates distance correlation functions for residues which are separated by a specific number of residues, m, along the chain. Figure 4.7 shows a series of distance correlation functions generated for residues separated by m residues where m is varied from 2 to 21. When m equals 1, the the distance correlation function monitors the variation of distances between adjacent residues in a chain. There is essentially no variation and the separation distance is equal to 3.8-Å. Many of the other distance correlation functions contain very sharp peaks. When m is equal to 2 or 3, the correlation functions show that there is little allowed variation. Most residues separated by 2 or 3 residues along the chain are limited to a very small range of separation distances. Note that distinct peaks are evident for correlation functions up to $m = 16$. Figure 4.8 shows the distance correlation functions for an ideal α-helix. Comparison of figures 4.7 and 4.8 suggests that these peaks represent residues involved in α-helices.

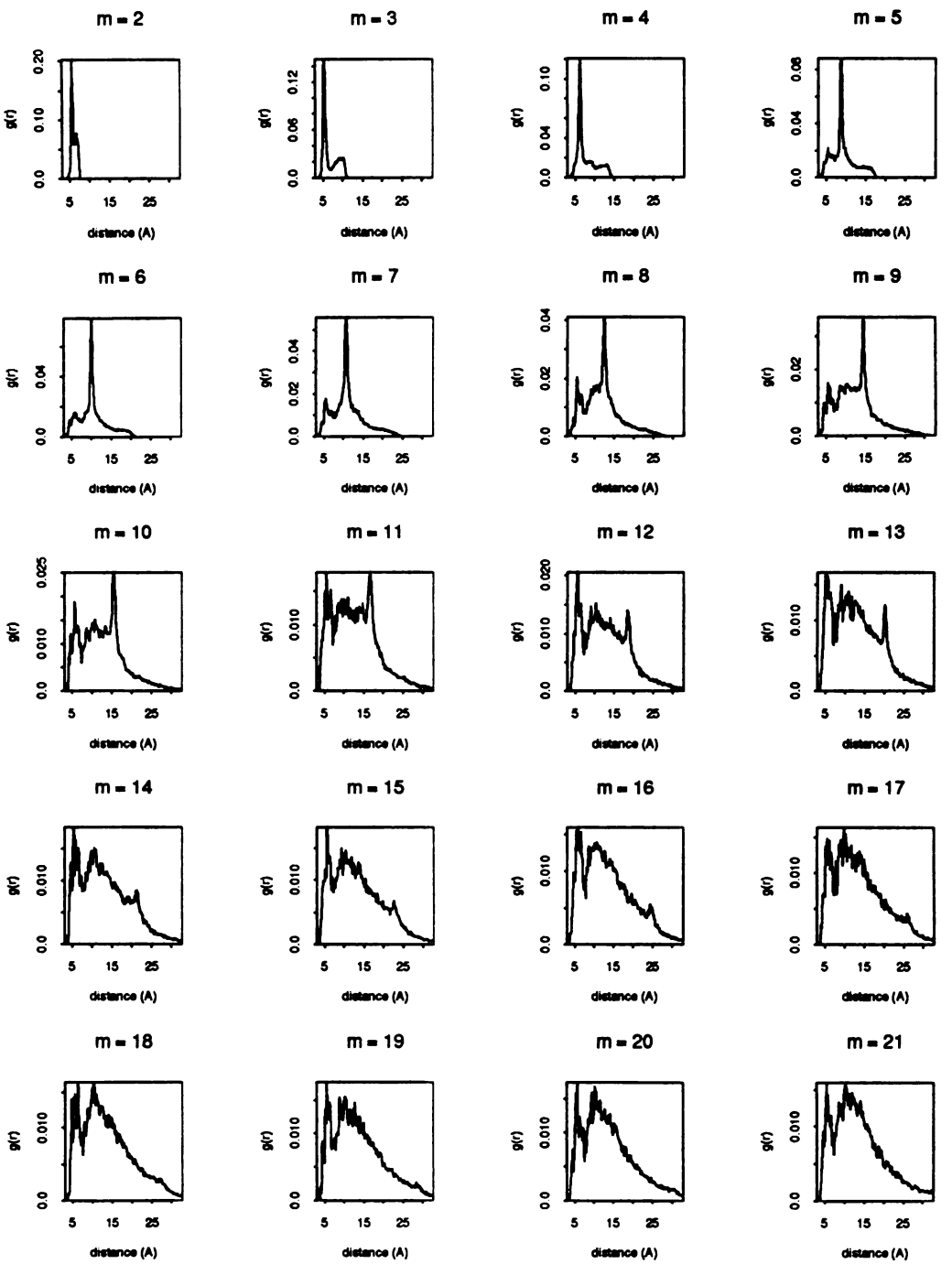


figure 4.7: Distance correlation functions of PDB proteins. Each plot represents the correlation function for residues which are separated by the specified number of residues.

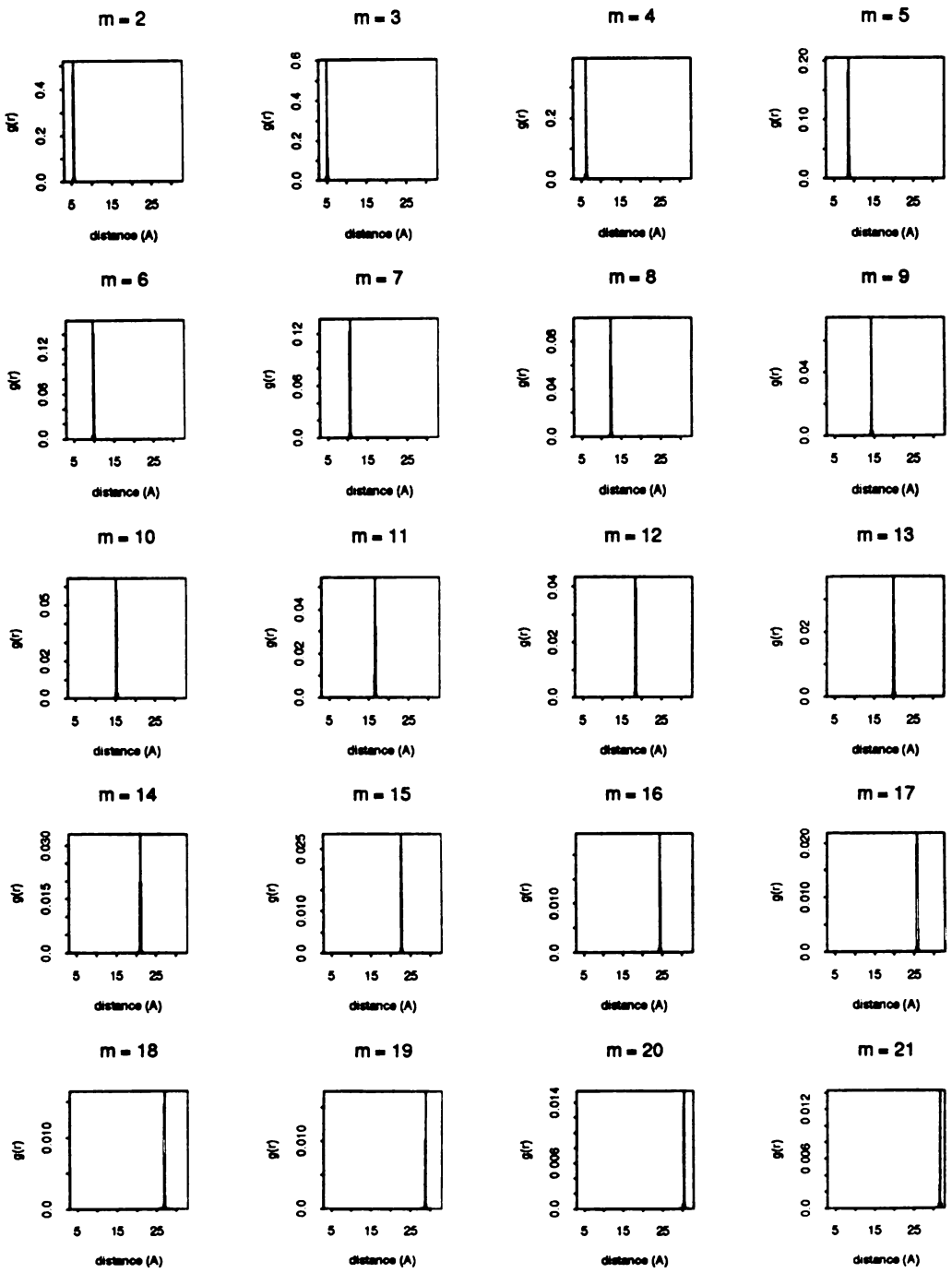


figure 4.8: Distance correlation functions of of an ideal α-helix. Each plot represents the correlation function for residues which are separated by the specified number of residues.

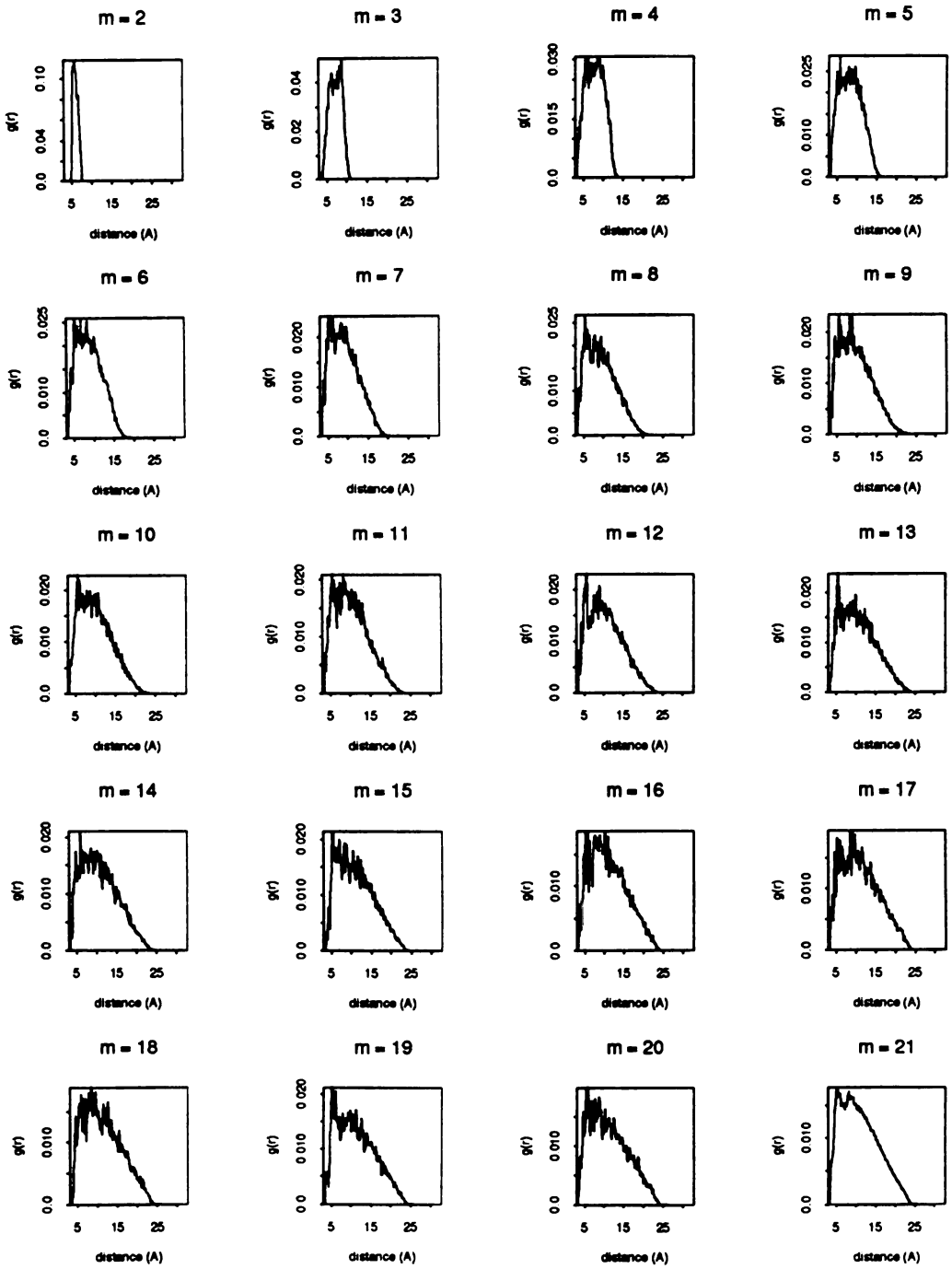


figure 4.9: Distance correlation functions of maximally compact poly-alanine 100mers. Each plot represents the correlation function for residues which are separated by the specified number of residues.

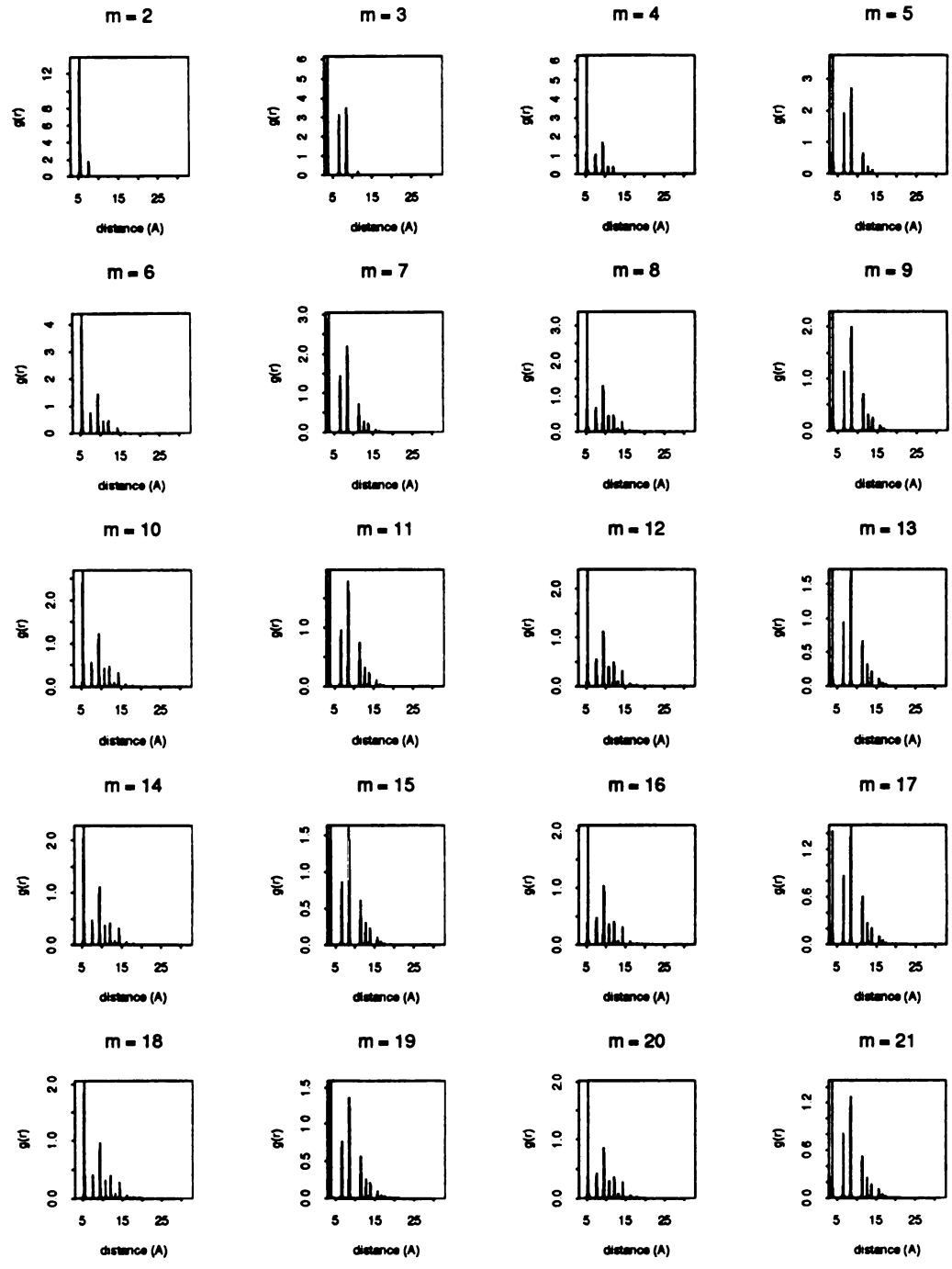


figure 4.10: Distance correlation functions of compact cubic lattice chains. Each plot represents the correlation function for residues which are separated by the specified number of residues.

One can use specific patterns of correlations to search conformations for the presence of α-helices. If a conformation consists of associated residues which correspond to the α-helix peaks for $m = 2,3, \cdots, n$, then one has identified a helix of length n. Stringency of the definition can be altered by varying the range of values defining a helix peak for each of the correlation plots. A helix detector based on these ideas has the advantage that the stringency of the criterion can be determined by constructing plots such a the one shown in figure 4.7 so that the range of values defining a helix is based on the variation observed in real protein structures.

When m is equal to 3 and 4, there are also broader peaks at larger separation distances. These peaks correspond to residues involved in strands. As m gets larger, two trends are apparent. First, the distributions at larger distances become broader. This indicates that there is a lot of variability in the distance between two residues which are widely separated in sequence. Second, a peak at about 5.4-Å becomes apparent. This peak corresponds to near neighbor spatial interactions between residues which are separated in sequence but close in space and represent the interactions which give rise to the large peak in figure 4.6B.

Figure 4.9 shows the distance correlation functions for maximally compact poly-alanine chains as described in chapter 3. The general shape of the distributions for $m > 11$ is very similar to the distribution obtained for real proteins. There are also peaks at about 5.4-Å for large m. In contrast to the real protein correlation functions, however, there is an obvious absence of sharp peaks. This indicates that the poly-alanine chains do not have residue-to-residue correlations which are as specific as those in real proteins. As discussed in chapter 3, these specific interactions are due to primarily to hydrogen bond interactions which give secondary structures their exquisite specificity. Figure 4.10 shows the distance correlation functions for compact simple cubic lattice models. Many differences between the lattice model and real proteins are obvious. The lattice model

has much less variation with respect to inter-residue separations; just a few distances are occupied in each plot. The nearest neighbor interactions for odd m are at 3.8-Å as opposed to 5.4-Å in real proteins. For even m, residues can not be configured to exist on adjacent lattice sites so the closest distance is $\sqrt{2} \times 3.8 \approx 5.4$ Å. In spite of these differences, the lattice model correlation functions approximate the distributions defined by real proteins for large m.

It is also possible to define bond projection correlation functions. They are defined as the dot product of two bond vectors separated by m residues in sequence. That is

$$projection(m) = b_i \cdot b_{i+m} = \tag{4.3}$$

$$(x_{i+1} - x_i)(x_{i+m+1} - x_i+m) + (y_{i+1} - y_i)(y_{i+m+1} - y_i+m) + (z_{i+1} - z_i)(z_{i+m+1} - z_i+m).$$

Figures 4.11, 4.12, and 4.13 show the projection correlation functions for real proteins, compact poly-alanine chains, and cubic lattice chains respectively. These figures mirror the observations made in the distance correlation functions. Sharp peaks in the real protein correlations are due to the presence of well defined secondary structures. The poly-alanine chains show much more variability and at high m the projection correlations for real proteins and random poly-alanine chains look very similar. Cubic lattice chains, in contrast, clearly show that there are only three ways bonds can be oriented relative to one another.

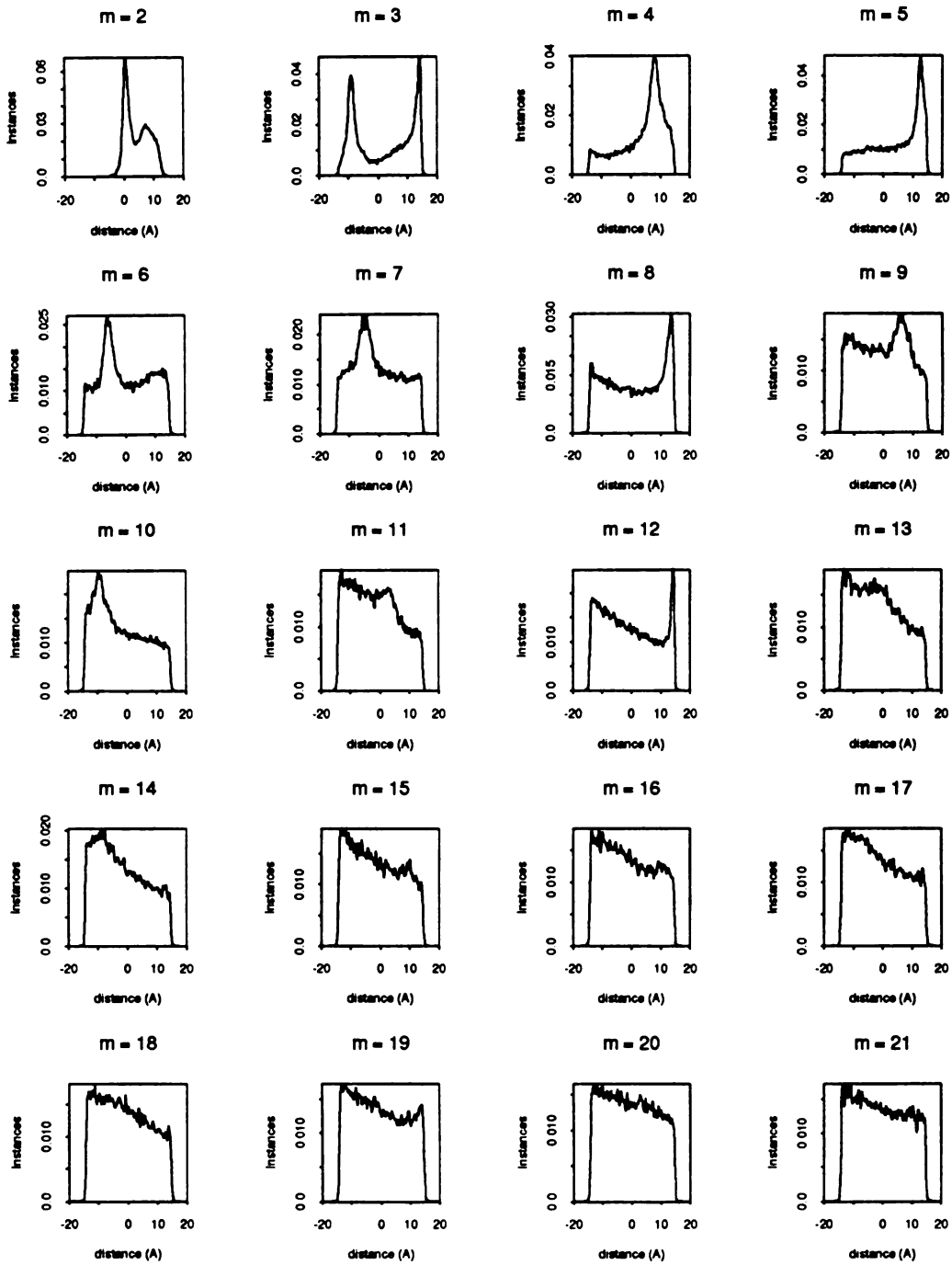


figure 4.11: Bond projection correlation functions of PDB proteins. Each plot represents the correlation function for residues which are separated by the specified number of residues.

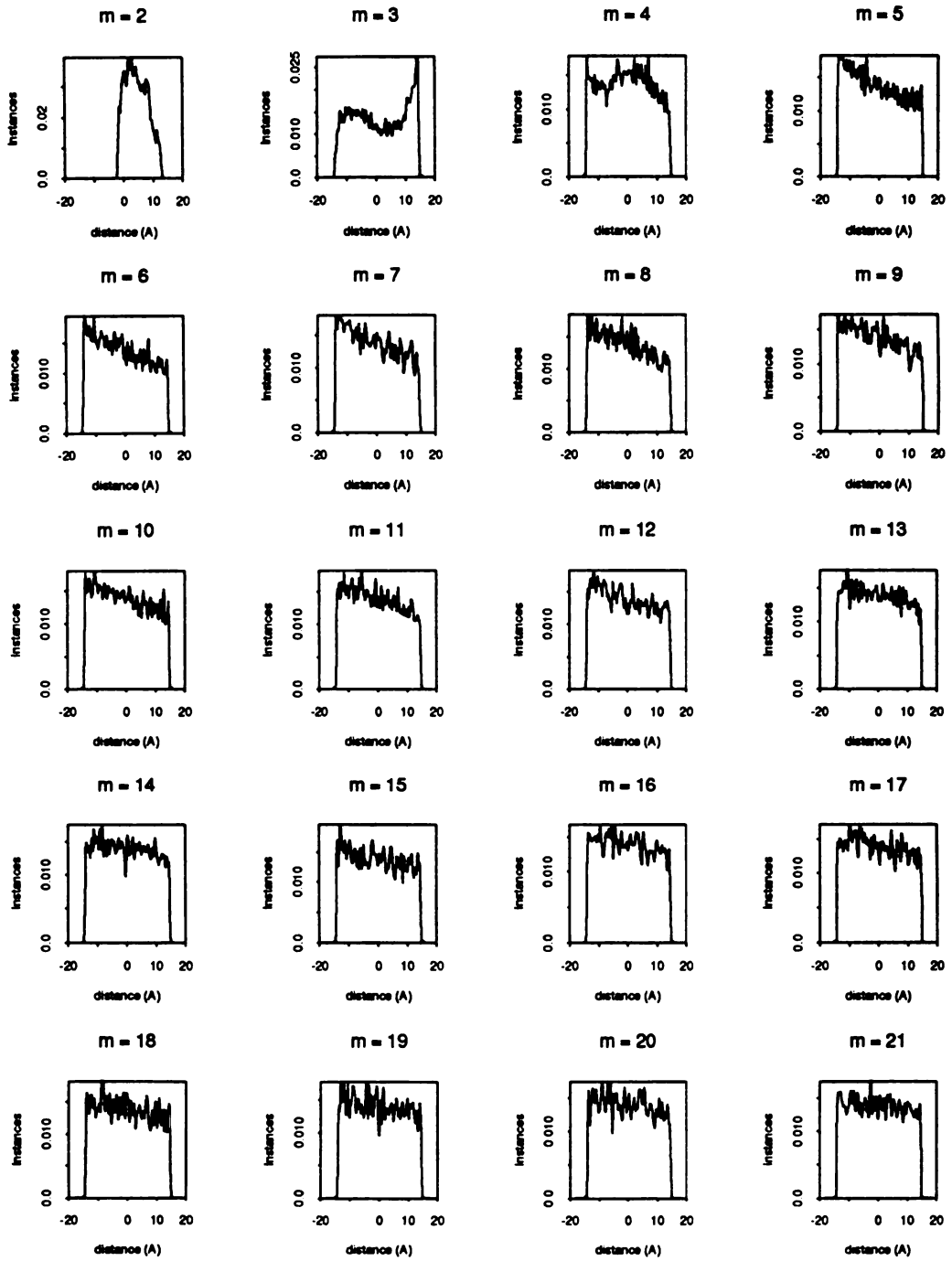


figure 4.12: Bond projection correlation functions of maximally compact poly-alanine 100mers. Each plot represents the correlation function for residues which are separated by the specified number of residues.

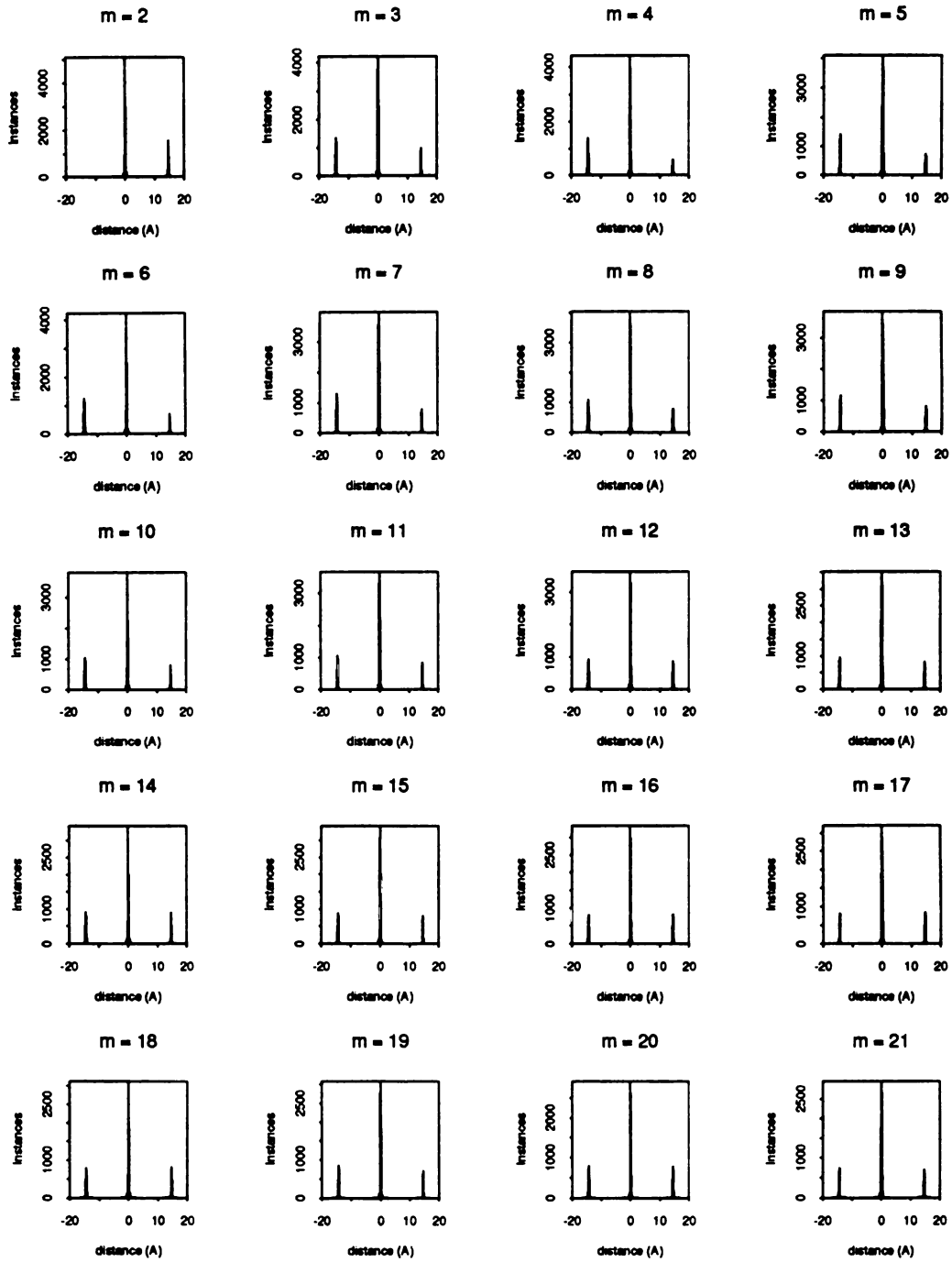


figure 4.13: Bond projection correlation functions of compact cubic lattice chains. Each plot represents the correlation function for residues which are separated by the specified number of residues.

Conclusions

The problem of finding a lattice representation is is difficult one. One must decide which property of a polypeptide chain is of the greatest relevance to the questions being asked. For some purposes, this may be geometric equivalence, as in the the work of Covell and Jernigan. For other purposes, a topological criterion may be more appropriate.

Using a build-up procedure, it is possible to generate lattice models which reproduce the topological interactions present in real proteins. The build-up procedure attempts to build a lattice model which maximizes the number of closely interacting residues in a real protein. The decision to use a topological criterion is due to the nature of the lattice model. The lattice model is based on the assumption that the native state of a protein is maximally compact and the dominant driving force for protein folding comes from the hydrophobic interaction. Ideally, the best lattice representation will be maximally compact. Despite the use of topological criterion, however, lattice representations of real proteins are not maximally compact. This is because real proteins describe complex volumes which do not minimize surface area whereas a maximally compact cubic lattice chain has minimal surface area. Consequently, a lattice-based protein folding simulation, based on the sequence of a real protein, may find configurations which are more compact and of lower energy than the conformation generated by the build-up procedure.

The observation that a lattice representation of a real protein is not the lowest energy lattice configuration suggests that there may not exist a one-to-one mapping between real protein structures and simple models of protein structures. One must be aware of the differences in physical constraints between model systems and real polypeptides. There are, for example, large differences between the rotational degrees of freedom allowed polypeptide chains relative to lattice chains. These differences are

apparent by examining distance and projection correlation functions of ensembles of real and lattice chains.

The distance and projection correlation functions, which I refer to as protein spectra, are very useful way to characterize the properties of ensembles of polymers. Using them, it is possible to highlight the similarities and differences between lattice chains, poly-alanine chains, and real proteins. With respect to real chains, it is possible to identify helices by virtue of the sharply defined peaks in both the distance and projection correlation functions.

References

[1] H. S. Chan and K. A. Dill, "Origins of Structure in Globular Proteins," *Proceedings of the National Academy of Science, USA*, 87:6388 - 6392, 1990.

[2] H. S. Chan and K. A. Dill, "Compact Polymers," *Macromolecules*, 22:4559 - 4573, 1989.

[3] H. S. Chan and K. A. Dill, "Sequence Space Soup of Proteins and Copolymers," *Journal of Chemical Physics*, 95(5):3775 - 3787, 1991.

[4] K. F. Lau and K. A. Dill, "Theory for Protein Mutability and Biogenesis," *Proceedings of the National Academy of Science, USA*, 87:638 - 642, 1990.

[5] P. D. Thomas and K. A. Dill, "Local and Nonlocal Interactions in Globular Proteins and Mechanisms of Alcohol Denaturation," *Protein Science*, 2:2050 - 2065, 1993.

[6] E. E. Lattman, K. M. Fiebig, and K. A. Dill, "Modeling Compact Denatured States of Proteins," *Biochemistry*, 33:6158 - 6166, 1994.

[7] S. Bromberg and K. A. Dill, "Side-chain Entropy and Packing in Proteins," *Protein Science*, 3:997 - 1009, 1994.

[8] A. Kolinski and J. Skolnick, "Discretized Model of Proteins I. Monte Carlo Study of Cooperativity in Homopolymers," *Journal of Chemical Physics*, 97(12):9412 - 9426, 1992.

[9] A. Godzik, J. Skolnick, and A. Kolinski, "Simulations of the Folding Pathway of Triose Phosphate Isomerase-type alpha/beta Barrel Proteins," *Proceedings of the National Academy of Science, USA*, 89:2629 - 2633, 1992.

[10] A. Kolinski and J. Skolnick, "Monte Carlo Simulations of Protein Folding. I. Lattice Model and Interaction Scheme," *Proteins*, 18:338 - 352, 1994.

[11] D. G. Covell and R. L. Jernigan, "Conformations of Folded Proteins in Restricted Spaces," *Biochemistry*, 29:3287 - 3294, 1990.

[12] F. M. Richards, "The Interpretation of Protein Structures: Total volume, Group volume Distributions and Packing Density." *Journal of Molecular Biology*, 82(1):1 - 14, 1974.

[13] F. M. Richards and C. E. Kundrot, "Identification of Structural Motifs from Protein Coordinate Data: Secondary Structure and First-level Supersecondary Structure," *Proteins*, 3(2):71 - 84, 1988.

[14] K. A. Dill, "Dominant Forces in Protein Folding," *Biochemistry*, 29(31):7133 - 7155, 1990.

[15] H. S. Chan and K. A. Dill, "The Effects of Internal Constraints on the Configurations of Chain Molecules," *Journal of Chemical Physics*, 92(5):3118 - 3135, 1990..

[16] M. H. Hao, S. Rackovsky, A. Liwo, M. R. Pincus, and H. A. Scheraga, "Effects of Compact Volume and Chain Stiffness on the Conformations of Native Proteins," *Proceedings of the National Academy of Science, USA*, 89:6614 - 6618, 1992.

[17] N. D. Socci, W. S. Bialek, and J. N. Onuchic, "Properties and Origins of Protein Secondary Structure," *Physical Review E.*, 49:3440 - 3443, 1994.

Appendix A

References

for dataset of 158

protein structures

References for 158 protein dataset

451c Y. Matsuura, T. Takano, and R. E. Dickerson (1982). Structure of cytochrome C_{551} from P. *aeruginosa* refined at 1.6 Angstroms resolution and comparison of the two redox forms. *J. Mol. biol. 156,* 389

155c R. Timkovich and R. E. Dickerson (1976). The structure of *Paracoccus denitrificans* cytochrome C_{550}. *J. Biol. Chem. 251,* 4033

256b F. Lederer, A. Glatigny, P. H. Bethge, H. D. Bellamy, and F. S. Mathews (1981). Improvement of the 2.5 Angstroms resolution model of cytochrome B_{562} by redetermining the primary structure and using molecular graphics. *J. Mol. Biol. 148,* 427

1aat V. N. Malaskevich, V. M. Kochkina, IU. M. Torchinskii, and E. G. Arutiunian (1982). Oxoglutarate-induced conformational changes in cytosolic aspartate aminotransferase. *Dokl. Akad. Nauk SSSR 267* 1257

1abp G. L. Gilliland and F. A. Quiocho (1981). Structure of the L-arabinose-binding protein from *Escherichia coli* at 2.4 Angstroms resolution. *J. Mol. Biol. 146,* 341

2abx R. A. Love and R. M. Stroud (1986). The crystal structure of α-bungarotoxin at 2.5 Angstroms resolution. Relation to solution structure and binding to acetylcholine receptor. *Protein Eng. 1,* 37

2act E. N. Baker, E. J. Dodson (1980). Crystallographic refinement of the structure of actinidin at 1.7 Angstroms resolution by fast Fourier least-squares methods. *Acta Crystallogr. A36,* 559

1acx V. Z. Pletnev, A. P. Kuzin, and L. V. Malinina (1982). Actinoxanthin structure at the atomic level. *Bioorg. Khim. 8,* 1637

6adh_a H. Eklund, J.-P. Samama, L. Wallen, C.-I. Branden, A. Akeson, and T. A. Jones (1981). Structure of triclinic ternary complex of horse liver alcohol dehydrogenase at 2.9 Angstroms resolution. *J. Mol. Biol. 146,* 561

3adk D. Dreusicke, P. A. Karplus, and G. E. Schulz (1988). Refined structure of porcine cytosolic adenylate kinase at 2.1 Angstroms resolution. *J. Mol. Biol. 199,* 359

2ait A. D. Kline, W. Braun, and K. Wuthrich (1988). Determination of the complete three-dimensional structure of the α-amylase inhibitor tendamistat in aqueous solution by nuclear magnetic resonance and distance geometry. *J. Mol. Biol. 204* 675

1alc K. R. Acharya, D. I. Stuart, N. P. C. Walker, M. Lewis, and D. C. Phillips (1989). Refined structure of baboon α-lactalbumin at 1.7 Angstroms resolution. Comparison with c-type lysozyme. *J. Mol. Biol. 208,* 99

2alp M. Fujinaga, L. T. J. Delbaere, G. D. Brayer, and M. N. G. James (1985). Refined structure of α-lytic protease at 1.7 Angstroms resolution. Analysis of hydrogen bonding and solvent structure. *J. Mol. Biol. 184,* 479

4ape T. L. Blundell, J. A. Jenkins, B. T. Sewall, L. H. Pearl, J. B. Cooper, I. J. Tickle, B. Veerapandian, and S. P. Wood (1990). X-ray analyses of aspartic proteinases. The three-dimensional structure at 2.1 Angstroms resolution of endothiapepsin. *J. Mol. Biol. 211,* 919

7api R. Engh, H. Loebermann, M. Schneider, G. Wiegand, R. Huber, and C.-B. Laurell (1989). The s variant of human α_1-antitrypsin, structure and implications for function and metabolism. *Protein Eng. 2,* 407

3app M. N. G. James and A. R. Sreecki (1983). Structure and refinement of penicillopepsin at 1.8 Angstroms resolution. *J. Mol. Biol. 163,* 299

2apr K. Suguna, R. R. Bott, E. A. Padlan, E. Subramanian, S. Sheriff, G. H. Cohen, and D. R. Davies (1987). Structure and refinement at 1.8 Angstroms resolution of the aspartic proteinase from *rhizopus chinensis. J. Mol. Biol. 196,* 877

2atc_c R. B. Honzatko, J. L. Crawford, H. L. Monaco, J. E. Ladner, B. F. P. Edwards, D. R. Evans, S. G. Warren, D. C. Wiley, R. C. Ladner, and W. N. Lipscomb (1982). Crystal and molecular structures of native and CTP-liganded aspartate carbamoyltransferase from *Escherichia coli. J. Mol. Biol. 160,* 219

2atc_r R. B. Honzatko, J. L. Crawford, H. L. Monaco, J. E. Ladner, B. F. P. Edwards, D. R. Evans, S. G. Warren, D. C. Wiley, R. C. Ladner, and W. N. Lipscomb (1982). Crystal and molecular structures of native and CTP-liganded aspartate carbamoyltransferase from *Escherichia coli. J. Mol. Biol. 160,* 219

2aza_a E. N. Baker (1988). Structure of azurin from *Alcaligenes denitrificans.* Refinement at 1.8 Angstroms resolution and comparison of the two crystallographically independent molecules. *J. Mol. Biol. 203, 1071*

3b5c F.S.Mathews, P. Argos and M. Levine (1972). The structure of cytochrome b_5 at 2.0 Angstroms resolution. *Cold Spring Harbor Symp. Quant. Biol. 36,* 387

1bds P. C. Driscoll, A. M. Gronenborn, L. Beress, and G. M. Clore (1989). Determination of the three-dimensional solution structure of the antihypertensive and antiviral protein BDS-I from the sea anemone *Anemonia sulcata.* A study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing. *Biochemistry 28,* 2188

3blm O. Herzberg and J. Moult (1987). Bacterial resistance to β-lactam antibiotics. Crystal structure of β-lactamase from *Staphylococcus aureus* PC1 at 2.5 Angstroms resolution. *Science 236,* 694

1bp2 B. W. Dijkstra, K. H. Kalk, W. G. J. Hol, and J. Drenth (1981). Structure of bovine pancreatic phospholipase A2 at 1.7 Angstroms resolution. *J. Mol. Biol. 147,* 97

3c2c G. E. Bhatia (1981). Refinement of the crystal structure of oxidized *Rhodospirillum rubrum* cytochrome c2. *Thesis, University of California, San Diego*

2ca2 A. E. Eriksson, P. M. Kylsten, T. A. Jones, and A. Liljas (1988). Crystallographic studies of inhibitor binding sites in human carbonic anhydrase II. A pentacoordinated binding of the SCN^- ion to the zinc at high pH. *Proteins. Struct., Funct. Genet. 4,* 283

8cat_a I. Fita and M. G. Rossmann (1985). The NADPH binding site on beef liver catalase. *Proc. Nat. Acad. Sci. USA 82,* 1604

1cbp J. M. Guss, E. A. Merritt, R. P. Phizackerly, B. Hedman, M. Murata, K. O. Hodgson, and H. C. Freeman (1988). Phase determination by multiple-wavelength x-ray diffraction. Crystal structure of a basic "blue" copper protein from cucumbers. *Science 241,* 806

1cc5 D. C. Carter, K. A. Melis, S. E. O'Donnell, B. K. Burgess, W. F. Furey, Junior, B.-C. Wang, and C. D. Stout (1985). Crystal structure of *Azotobacter* cytochrome c_5 at 2.5 Angstroms resolution. *J. Mol. Biol. 184,* 279

1ccr H. Ochi, Y. Hata, N. Tanaka, M. Kakudo, T. Sakurai, S. Aihara, and Y. Morita (1983). Structure of rice ferricytochrome c at 2.0 Angstroms

resolution. *J. Mol. Biol. 166,* 407

2ccy_a B. C. Finzel, P. C. Weber, K. D. Hardman, and F. R. Salemme (1985). Structure of ferricytochrome c' from *Rhodospirillum molischianum* at 1.67 Angstroms resolution. *J. Mol. Biol. 186,* 627

2cdv Y. Higuchi, M. Kusunoki, Y. Matsuura, N. Yasuoka, and M. Kakudo (1984). Refined structure of cytochrome c_3 at 1.8 Angstroms resolution. *J. Mol. Biol. 172, 109*

2ci2 C. A. McPhalen and M. N. G. James (1987). Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds. *Biochemistry 26,* 261

3cln Y. S. Babu, C. E. Bugg, and W. J. Cook (1988). Structure of calmodulin refined at 2.2 Angstroms resolution. *J. Mol. Biol. 204,* 191

1cms G. L. Gilliland, E. L. Winborne, J. Nachman, and A. Wlodawer (1990). The three-dimensional structure of recombinant bovine chymosin at 2.3 Angstroms resolution. *Proteins: Struct., Funct., Genet. 8,* 82

2cna G. N. Reeke, Jr, J. W. Becker, and G. M. Edelman (1975). The covalent and three-dimensional structure of concanavalin a, IV. Atomic coordinates, hydrogen bonding, and quaternary structure. *J. Biol. Chem. 250,* 1525

5cpa D. C. Rees, M. Lewis, and W. N. Lipscomb (1983). Refined crystal structure of carboxypeptidase A at 1.54 Angstroms resolution. *J. Mol. Biol. 168,* 367

2cpp T. L. Poulos, B. C. Finzel, and A. J. Howard (1987). High-resolution crystal structure of cytochrome P450CAM. *J. Mol. Biol. 195,* 687

5cpv A. L. Swain, R. H. Kretsinger, and E. L. Amma (1989). Restrained least squares refinement of native calcium and cadmium-substituted carp parvalbumin using x-ray crystallographic data at 1.6-Angstroms resolution. *J. Biol. Chem. 264,* 16620

1crn M. M. Teeter (1984). Water structure of a hydrophobic protein at atomic resolution. Pentagon rings of water molecules in crystals of crambin. *Proc. Nat. Acad. Sci. USA 81,* 6014

1cro_o Y. Takeda, J. G. Kim, C. G. Caday, E. Steers, Jr, D. H. Ohlendorf, W. F. Anderson, and B. W. Matthews (1986). Different interactions used by cro repressor in specific and nonspecific DNA binding. *J. Biol. Chem. 261,* 8608

2cro A. Mondragon, C. Wolberger, and S. C. Harrison (1989). Structure of phage 434 cro protein at 2.35 Angstroms resolution. *J. Mol. Biol. 205,* 179

1cse W. Bode, E. Papamokos, and D. Musil (1987). The high-resolution x-ray crystal structure of the complex formed between subtilisin carlsberg and eglin c, an elastase inhibitor from the leech *Hirudo medicinalis*. Structural analysis, subtilisin structure and interface geometry. *Eur. J. Biochem. 166,* 673

1ctf M. Leijonmarck and A. Liljas (1987). Structure of the C-terminal domain of the ribosomal protein L7/L12 from *Escherichia coli* at 1.7 Angstroms. *J. Mol. Biol. 195,* 555

1ctx M. D. Walkinshaw, W. Saenger, and A. Maelicke (1980). Three-dimensional structure of the "long" neurotoxin from cobra venom. *Proc. Nat. Acad. Sci. USA 77,* 2400

5cyt T. Takano (1984) Refinement of myoglobin and cytochrome c *Methods and applications in crystallographic computing*, p.262 (S. R. Hall and T.

Ashida, Eds). Oxford University Press, Oxford, England

2cyp B. C. Finzel, T. L. Poulos, and J. Kraut (1984). Crystal structure of yeast cytochrome c peroxidase refined at 1.7 Angstroms resolution. *J. Biol. Chem. 259,* 13027

3dfr J. T. Bolin, S. J. Filman, D. A. Matthews, R. C. Hamlin, and J. Kraut (1982). Crystal structures of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase refined at 1.7 Angstroms resolution. I. General features and binding of methotrexate. *J. Biol. Chem. 257,* 13650

5ebx P. W. R. Corfield, T.-J. Lee, and B. W. Low (1989). The crystal structure of erabutoxin a at 2.0 Angstroms resolution. *J. Biol. Chem. 264,* 9239

1ecd W. Steigemann and E. Weber (1979). Structure of erythrocruorin in different ligand states refined at 1.4 Angstroms resolution. *J. Mol. Biol. 127,* 309

1efm F. Jurnak (1985). Structure of the GDP domain of EFTU and location of the amino acids homologous to *ras* oncogene proteins. *Science 230,* 32

2enl L. Lebioda, B. Stec, and J. M. Brewer (1989). The structure of yeast enolase at 2.25 Angstroms resolution. An 8-fold β+α-barrel with a novel $\beta\beta\alpha\alpha(\beta\alpha)_6$ topology. *J. Biol. Chem. 264,* 3685

2est D. L. Hughes, L. C. Sieker, J. Bieth, and J.-L. Dimicoli (1982). Crystallographic study of the binding of a trifluoroacetyl dipeptide anilide inhibitor with elastase. *J. Mol. Biol. 162,* 645

1etu T. F. M. la Cour, J. Nyborg, S. Thirup, and B. F. C. Clark (1985). Structural details of the binding of guanosine diphosphate to elongation factor tu from *E. coli* as studied by x-ray crystallography. *EMBO J. 4* , 2385

2fb4_h

2fb4_l M. Marquart, J. Deisenhofer, R. Huber, and W. Palm (1980). Crystallographic refinement and atomic models of the intact immunoglobulin molecule KOL and its antigen-binding fragment at 3.0 Angstroms and 1.9 Angstroms resolution. *J. Mol. Biol. 141,* 369

1fc1 J. Deisenhofer (1981). Crystallographic refinement and atomic models of a human Fc fragment and its complex with fragment b of protein a from *Staphylococcus aureus* at 2.9 and 2.8 Angstroms resolution. *Biochemistry 20,* 2361

4fd1 C. D. Stout (1989). Refinement of the 7 Fe ferredoxin from *Azotobacter* at 1.9 Angstroms resolution. *J. Mol. Biol. 205,* 545

3fxn W. W. Smith, R. M. Burnett, G. D. Darling, and M. L. Ludwig (1977). Structure of the semiquinone form of flavodoxin from *Clostridium MP*. Extension of 1.8 Angstroms resolution and some comparisons with the oxidized state. *J. Mol. Biol. 117,* 195

3gap_c

3gap_o I. T. Weber and T. A. Steitz (1987). Structure of a complex of catabolite gene activator protein and cyclic AMP refined at 2.5 Angstroms resolution. *J. Mol. Biol. 198,* 311

2gbp N. K. Vyas, M. N. Vyas, and F. A. Quiocho (1988). Sugar and signal-transducer binding sites of the *Escherichia coli* galactose chemoreceptor protein. *Science 242,* 1290

1gcr L. Summers, G. Wistow, M. Narebor, D. Moss, P. Lindley, C. Slingsby, T. Blundell, H. Bartunik, and K. Bartels (1984). X-ray studies of the lens specific proteins. The crystallins *Pept. Protein Rev. 3,* 147

1gd1_o T. Skarzynski, P. C. E. Moody, and A. J. Wonacott (1987). Structure of holo-glyceraldehyde-3-phosphate dehydrogenase from *Bacillus stearothermophilus* at 1.8 Angstroms resolution. *J. Mol. Biol. 193,* 171

2gls_a M. M. Yamashita, R. J. Almassy, C. A. Janson, D. Cascio, and D. Eisenberg (1989). Refined atomic model of glutamine synthetase at 3.5 Angstroms resolution. *J. Biol. Chem. 264,* 17681

1gp1_a O. Epp, R. Ladenstein, and A. Wendel (1983). The refined structure of the selenoenzyme glutathione peroxidase at 0.2-nm resolution. *Eur. J. Biochem. 133,* 51

3grs P. A. Karplus and G. E. Schulz (1987). Refined structure of glutathione reductase at 1.54 Angstroms resolution. *J. Mol. Biol. 195,* 701

1hho_a B. Shaanan (1983). Structure of human oxyhaemoglobin at 2.1 Angstroms resolution. *J. Mol. Biol. 171,* 31

1hip C. W. Carter Jr,, J. Kraut, S. T. Freer, N.-H. Xuong, R. A. Alden, and R. G. Bartsch (1974). Two Angstrom crystal structure of oxidized chromatium high potential iron protein. *J. Biol. Chem. 249,* 4212

1hkg T. A. Steitz, M. Shoham, and W. S. Bennett, Jr. (1981). Structural dynamics of yeast hexokinase during catalysis. *Philos. Trans. R. Soc. London B293,* 43

2hla_h

2hla_m T. P. J. Garrett, M. A. Saper, P. J. Bjorkman, J. L. Strominger, and D. C. Wiley (1989). Specificity pockets for the side chains of peptide antigens in HLA-Aw68. *Nature 342,* 692

2hmg_1

2hmg_2 W. I. Weis, A. T. Bruenger, J. J. Skehel, and D. C. Wiley (1990). Refinement of the influenza virus hemagglutinin by simulated annealing. *J. Mol. Biol. 212,* 737

1hmq_a M. A. Holmes and R. E. Stemkamp (1991). The structures of met and azidomet hemerythrin at 1.66 Angstroms resolution. *J. Mol. Biol. 220,* 723

1hoe J. W. Pflugrath, G. Wiegand, R. Huber, and L. Vertesy (1986). Crystal structure determination, refinement and the molecular model of the α-amylase inhibitor Hoe-467a. *J. Mol. Biol. 189,* 383

3hvp A. Wlodawer, M. Miller, M. Jaskolski, B. K. Snthyanarayana, E. Baldwin, I. T. Weber, L. M. Selk, L. Clawson, J. Schneider, and S. B. H. Kent (1989). Conserved folding in retroviral proteases. Crystal structure of a synthetic HIV-1 protease. *Science 245,* 616

2i1b J. P. Priestle, H.-P. Schaer, and M. G. Gruetter (1989). Crystallographic refinement of interleukin-1β at 2.0 Angstroms resolution. *Proc. Nat. Acad. Sci. USA 86,* 9667

3icb D. M. W. Szebenyi and K. Moffat (1986). The refined structure of vitamin D-dependent calcium-binding protein from bovine intestine. Molecular details, ion binding, and implications for the structure of other calcium-binding proteins. *J. Biol. Chem. 261,* 8761

4ins_a E. N. Baker, T. L. Blundell, J. F. Cutfield, S. M. Cutfield, E. J. Dodson, G. G. Dodson, D. M. Crowfoot Hodgkin, R. E. Hubbard, N. W. Isaacs, C. D. Reynolds, K. Sakabe, N. Sakabe, and N. M. Vijayan (1988). The structure of 2ZN pig insulin crystals at 1.5 Angstroms resolution. *Philos. Trans. R. Soc. London B319,* 369

1kga I. M. Mavridis, M. H. Hatada, A. Tulinsky, and L. Lebioda (1982). Structure of 2-keto-3-deoxy-6-phosphogluconate aldolase at 2.8 Angstroms resolution. *J. Mol. Biol. 162,* 419

2lbp J. S. Sack, S. D. Trakhanov, I. H. Tsigannik, and f. A. Quiocho (1989). Structure of the L-leucine-binding protein refined at 2.4 Angstroms resolution and comparison with the leu/ile/val-binding protein structure. *J. Mol. Biol. 206,* 193

3ldh J. L. White, M. L. Hackert, M. Buehner, M. J. Adams, G. C. Ford, P. J. Lentz, Jr., I. E. Smiley, S. J. Steindel, and M.G. Rossmann (1976). A comparison of the structures of apo dogfish M_4 lactate dehydrogenase and its ternary complexes. *J. Mol. Biol. 102,* 759

2lh4 E. G. Arutyunyan, I. P. Kuranova, B. K. Vainshtein, and W. Steigemann (1980). X-ray structural investigation of leghemoglobin. VI. Structure of acetate-ferrileghemoglobin at a resolution of 2.0 Angstroms (Russian) *Kristallografiya 25,* 80

2liv J. J. Sack, M. A. Saper, and F. A. Quiocho (1989). Periplasmic binding protein structure and function. Refined x-ray structures of the leucine/isoleucine/valine-binding protein and its complex with leucine. *J. Mol. Biol. 206,* 171

1lrd S. R. Jordan and C. O. Pabo (1988). Structure of the lambda complex at 2.5 Angstroms resolution. Details of the repressor-operator interactions. *Science 242,* 893

1lyz R. Diamond (1974). Real-space refinement of the structure of hen egg-white lysozyme. *J. Mol. Biol. 82,* 371

1lz1 P. J. Artymiuk and C. C. F. Blake (1981). Refinement of human lysozyme at 1.5 Angstroms resolution. Analysis of non-bonded and hydrogen-bond interactions. *J. Mol. Biol. 152,* 737

3lzm L. H. Weaver and B. W. Matthews (1987). Structure of bacteriophage T4 lysozyme refined at 1.7 Angstroms resolution. *J. Mol. Biol. 193,* 189

1mbd S. E. V. Phillips and B. P. Schoenborn (1981). Neutron diffraction reveals oxygen-histidine hydrogen bond in oxymyoglobin. *Nature 292,* 81

4mdh J. J. Birktoft, G. Rhodes, and L. J. Banaszak (1989). Refined crystal structure of cytoplasmic malate dehydrogenase at 2.5 Angstroms resolution. *Biochemistry 28,* 6065

2mev_vp1

2mev_vp2

2mev_vp3 S. Krishnaswamy and M. G. Rossmann (1990). Structural refinement and analysis of mengo virus. *J. Mol. Biol. 211,* 803

4mlt T. C. Terwilliger and D. Eisenberg (1982). The structure of melittin. I. Structure determination and partial refinement. *J. Biol. Chem. 257,* 6010

1mon_a C. Ogata, M. Hatada, G. Tomlinson, W.-C. Shin, and S.-H. Kim (1987). Crystal structure of the intensely sweet protein monellin. *Nature 328,* 739

1nxb D. Tsernoglou, G. A. Petsko, and R. A. Hudson (1978). Structure and function of snake venom curarimimetic neurotoxins. *Mol. Pharmacol. 14,* 710

2ovo W. Bode, O. Epp, R. Huber, M. Laskowski Jr, and W. Ardelt (1985). The crystal and molecular structure of the third domain of silver pheasant ovomucoid (OMSVP3). *Eur. J. Biochem. 147,* 387

2pab C. C. F. Blake, M. J. Geisow, S. J. Oatley, B. Rerat, and C. Rerat (1978). Structure of prealbumin, secondary, tertiary and quaternary interactions determined by Fourier refinement at 1.8 Angstroms. *J. Mol. Biol. 121,* 339

9pap I. G. Kamphuis, K. H. Kalk, M. B. A. Swarte, and J. Drenth (1984). Structure of papain refined at 1.65 Angstroms resolution. *J. Mol. Biol. 179,* 233

2paz E. T. Adman, S. Turley, R. Bramson, K. Petratos, D. Banner, D. Tsernoglou, T. Beppu, and H. Watanabe (1989). A 2.0 Angstroms structure of the blue copper protein (cupredoxin) from *Alcaligenes faecalis* s-6. *J. Biol. Chem. 264,* 87

1pcy J. M. Guss and H. C. Freeman (1983). Structure of oxidized poplar plastocyanin at 1.6 Angstroms resolution. *J. Mol. Biol. 169,* 521

4pep A. R. Sielecki, A. A. Fedorov, A. Boodhoo, N. S. Andreeva, and M. N. G. James (1990). The molecular and crystal structures of monoclinic porcine pepsin refined at 1.8 Angstroms resolution. *J. Mol. Biol. 214,* 143

1pfk_c

1pfk_o Y. Shirakihara and P. R. Evans (1988). Crystal structure of the complex of phosphofructokinase from *Escherichia coli* with its reaction products. *J. Mol. Biol. 204,* 973

3pgk T. N. Bryant, P. J. Shaw, N. P. Walker, P. L. Wendell, and H. C. Watson The structure of yeast phosphoglycerate kinase at 0.25 nm resolution. *To be published*

3pgm S. I. Winn, J. Warwicker, and H. C. Watson The structure of yeast phosphoglycerate mutase at 0.28 nm resolution. *To be published*

1phh H. A. Schreuder, J. M. van der Laan, W. G. J. Hol, and J. Drenth (1988). Crystal structure of p-hydroxybenzoate hydroxylase complexed with its reaction product 3,4-dihydroxybenzoate. *J. Mol. Biol. 199,* 637

1phy D. E. McRee, J. A. Tainer, T. E. Meyer, J. van Beeumen, M. A. Cusanovich, and E. D. Getzoff (1989). Crystallographic structure of a photoreceptor protein at 2.4 Angstroms resolution. *Proc. Nat. Acad. Sci. USA 86,* 6533

2pka W. Bode, Z. Chen, K. Bartels, C. Kutzbach, G. Schmidt-Kastner, and H. Bartunik (1983). Refined 2 Angstroms x-ray crystal structure of porcine pancreatic kallikrein a, a specific trypsin-like serine proteinase. Crystallization, structure determination, crystallographic refinement, structure and its comparison with bovine trypsin. *J. Mol. Biol. 164,* 237

2plv_vp1

2plv_vp2

2plv_vp3 D. J. Filman, R. Syed, M. Chow, A. J. Macadam, P. D. Minor, and J. M. Hogle (1989). Structural factors that control conformational transitions and serotype specificity in type 3 poliovirus. *EMBO J. 8,* 1567

1pp2_r S. Brunie, J. Bolin, D. Gewirth, and P. B. Sigler (1985). The refined crystal structure of dimeric phospholipase A_2 at 2.5 Angstroms. Access to a shielded catalytic center. *J. Biol. Chem. 260,* 9742

1ppt T. L. Blundell, J. E. Pitts, I. J. Tickle, S. P. Wood, and C.-W. Wu (1981). X-ray analysis (1.4 Angstroms resolution) of avian pancreatic polypeptide. Small globular protein hormone. *Proc. Nat. Acad. Sci. USA 78,* 4175

1prc_c

1prc_l

1prc_m

1prc_h J. Deisenhofer and H. Michel (1989). The photosynthetic reaction center from the purple bacterium *Rhodopseudomonas viridis*. *Science 245,* 1463

2prk C. Betzel, G. P. Pal, and W. Saenger (1988). Synchrotron x-ray data collection and restrained least-squares refinement of the crystal structure of proteinase K at 1.5 Angstroms resolution. *Acta Crystallogr. B44,* 163

1pte J. A. Kelly, J. R. Knox, P. C. Moews, G. J. Hite, J. B. Bartolone, H. Zhao, B. Joris, J.-M. Frere, and J.-M. Ghuysen (1985). 2.8 Angstroms structure of penicillin-sensitive D-alanyl carboxypeptidase-transpeptidase from *Streptomyces r61* and complexes with beta-lactams. *J. Biol. Chem. 260,* 6449

5pti A. Wlodawer, J. Walter, R. Huber, and L. Sjolin (1984). Structure of bovine pancreatic trypsin inhibitor. Results of joint neutron and x-ray refinement of crystal form II. *J. Mol. Biol. 180,* 301

4ptp J. L. Chambers and R. M. Stroud (1979). The accuracy of refined protein structures, comparison of two independently refined models of bovine trypsin. *Acta Crystallogr. B35* , 1861

1pyp E. G. Arutiunian, S. S. Terzian, A. A. Voronova, I. P. Kuranova, E. A. Smirnova, B. K. Vainstein, W. E. Hohne, and G. Hansen (1981). X-ray diffraction study of inorganic pyrophosphatase from baker's yeast at the 3 Angstroms resolution (Russian). *Dokl. Akad. Nauk SSSR 258,* 1481

1r69 A. Mondragon, S. Subbiah, S. C. Almo, M. Drottar, and S. C. Harrison (1989). Structure of the amino-terminal domain of phage 434 repressor at 2.0 Angstroms resolution. *J. Mol. Biol. 205,* 189

1rbb_a R. L. Williams, S. M. Greene, and A. McPherson (1987). The crystal structure of ribonuclease B at 2.5 Angstroms resolution. *J. Biol. Chem. 262,* 16020

1rei O. Epp, E. E. Lattman, M. Schiffer, R. Huber, and W. Palm (1975). The molecular structure of a dimer composed of the variable portions of the bence-jones protein REI refined at 2.0 Angstroms resolution. *Biochemistry 14,* 4943

1rhd J. H. Ploegman, G. Drent, K. H. Kalk, and W. G. J. Hol (1978). Structure of bovine liver rhodanese. I. Structure determination at 2.5 Angstroms resolution and a comparison of the conformation and sequence of its two domains. *J. Mol. Biol. 123,* 557

2rhe W. Furey, Jr, B. C. Wang, C. S. Yoo, and M. Sax (1983). Structure of a novel bence-jones protein (RHE) fragment at 1.6 Angstroms resolution. *J. Mol. Biol. 167,* 661

4rhv_vp1

4rhv_vp2

4rhv_vp3 E. Arnold and M. G. Rossmann (1988). The use of molecular-replacement phases for the refinement of the human rhinovirus 14 structure. *Acta Crystallogr. A44,* 270

3rn3 B. Howlin, D. S. Moss and G. W. Harris (1989). Segmented anisotropic refinement of bovine ribonuclease A by the application of the rigid-body TLS model. *Acta Crystallogr. A45,* 851

1rns R. J. Fletterick and H. W. Wyckoff (1975). Preliminary refinement of protein coordinates in real space. *Acta Crystallogr. A31,* 698

2rnt J. Koepke, M. Maslowska, U. Heinemann, and W. Saenger (1989). Three-dimensional structure of ribonuclease T_1 complexed with guanylyl-2′,5′-guanosine at 1.8 Angstroms resolution. *J. Mol. Biol. 206,* 475

3rp2 S. J. Remington, R. G. Woodbury, R. A. Reynolds, B. W. Matthews, and H. Neurath (1988). The structure of rat mast cell protease II at 1.9 Angstroms resolution. *Biochemistry 27,* 8097

7rsa A. Wlodawer, L. A. Svensson, L. Sjolin, and G. L. Gilliland (1988). Structure of phosphate-free ribonuclease A refined at 1.26 Angstroms. *Biochemistry 27,* 2705

5rub_a G. Schneider, Y. Lindqvist, and T. Lundqvist (1990). Crystallographic refinement and structure of ribulose-1,5-bisphosphate carboxylase from *Rhodospirillum rubrum* at 1.7 Angstroms resolution. *J. Mol. Biol. 211,* 989

5rxn K. D. Watenpaugh, L. C. Sieker, and L. H. Jensen (1980). Crystallographic refinement of rubredoxin at 1.2 Angstroms resolution. *J. Mol. Biol. 138,* 615

4sbv_a A. M. Silva and M. G. Rossmann (1985). The refinement of southern bean mosaic virus in reciprocal space. *Acta Crystallogr. B41,* 147

2sga J. Moult, F. Sussman, and M. N. G. James (1985). Electron density calculations as an extension of protein structure refinement. *Streptomyces griseus* protease at 1.5 Angstroms resolution. *J. Mol. Biol. 182,* 555

3sgb R. J. Read, M. Fujinaga, A. R. Sielecki, and M. N. G. James (1983). Structure of the complex of *Streptomyces griseus* protease b and the third domain of the turkey ovomucoid inhibitor at 1.8 Angstroms resolution. *Biochemistry 22,* 4420

1sn3 R. J. Almassy, J. C. Fontecilla-Camps, F. L. Suddath, and C. E. Bugg (1983). Structure of variant-3 scorpion neurotoxin from *Centruroides sculpturatus ewing*, refined at 1.8 Angstroms resolution. *J. Mol. Biol. 170,* 497

2sns F. A. Cotton, E. E. Hazen, Jr, and M. J. Legg (1979). Staphylococcal nuclease. Proposed mechanism of action based on structure of enzyme-thymidine 3′,5′-biphosphate-calcium ion complex at 1.5 Angstroms resolution. *Proc. Nat. Acad. Sci. USA 76,* 2551

2sod_o J. A. Tainer, E. D. Getzoff, K. M. Beem, J. S. Richardson, and D. C. Richardson (1982). Determination and analysis of the 2 Angstrom structure of copper, zinc superoxide dismutase . *J. Mol. Biol. 160,* 181

1srx A. Holmgren, B.-O. Soderberg, H. Eklund, and C.-I. Branden (1975). Three-dimensional structure of *Escherichia coli* thioredoxin-S_2 to 2.8 Angstroms resolution. *Proc. Nat. Acad. Sci. USA 72,* 2305

2ssi Y. Satow, Y. Watanabe, and Y. Mitsui (1980). Solvent accessibility and microenvironment in a bacterial protein proteinase inhibitor SSI (*Streptomyces* subtilisin inhibitor). *J. Biochem. (Tokyo) 88,* 1739

2stv T. A. Jones and L. Liljas (1984). Structure of satellite tobacco necrosis virus after crystallographic refinement at 2.5 Angstroms resolution. *J. Mol. Biol. 177,* 735

2taa Y. Matsuura, M. Kusunoki, W. Harada, and M. Kakudo (1984). Structure and possible catalytic residues of taka-amylase A. *J. Biochem. (Tokyo) 95,* 697

2tbv A. J. Olson, G. Bricogne, and S. C. Harrison (1983). Structure of tomato bushy stunt virus. IV. The virus particle at 2.9 Angstroms resolution. *J. Mol. Biol. 171,* 61

1tec P. Gros, M. Fujinaga, B. W. Dijkstra, K. H. Kalk, and W. G. J. Hol (1989). Crystallographic refinement by incorporation of molecular dynamics. Rhe thermostable serine protease thermitase complexed with eglin-c. *Acta Crystallogr. B45,* 488

1thi A. M. de Vos, M. Hatada, H. van der Wel, H. Krabbendam, A. F. Peerdeman, and S.-H. Kim (1985). Three-dimensional structure of thaumatin I, an intensely sweet protein. *Proc. Nat. Acad. Sci. USA 82,* 1406

1tim D. W. Banner, A. C. Bloomer, G. A. Petsko, D. C. Phillips, C. I. Pogson, I. A. Wilson, P. H. Corran, A. J. Furth, J. D. Milman, R. E. Offord, J. D. Priddle, and S. G. Waley (1975). Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 Angstroms resolution using amino acid sequence data. *Nature 255,* 609

2tmv K. Namba, R. Pattanayek, and G. Stubbs (1989). Visualization of protein-nucleic acid interactions in a virus. Refined structure of intact tobacco mosaic virus at 2.9 Angstroms resolution by x-ray fiber diffraction. *J. Mol. Biol. 208,* 307

4tnc K. A. Satyshur, S. T. Rao, D. Pyzalska, W. Drendel, M. Greaser, and M. Sundaralingam (1988). Refined structure of chicken skeletal muscle troponin c in the two-calcium state at 2 Angstroms resolution. *J. Biol .Chem. 263,* 1628

1tnf M. J. Eck and S. R. Sprang (1989). The structure of tumor necrosis factor-α at 2.6 Angstroms resolution. Implications for receptor binding. *J. Biol. Chem. 264,* 17595

1ubq S. Vijay-Kumar, C. E. Bugg, and W. J. Cook (1987). Structure of ubiquitin refined at 1.8 Angstroms resolution. *J. Mol. Biol. 194,* 531

1utg I. Morize, E. Surcouf, M. C. Vaney, Y. Epelboin, M. Buehner, F. Fridlansky, E. Milgrom, and J. P. Mornon (1987). Refinement of the $c222_1$ crystal form of oxidized uteroglobin at 1.34 Angstroms resolution. *J. Mol. Biol. 194,* 725

9wga C. S. Wright (1990). 2.2 Angstroms resolution structure analysis of two refined n-acetylneuraminyllactose-wheat germ agglutinin isolectin complexes. *J. Mol. Biol. 215,* 635

1wrp C. L. Lawson, R.-G. Zhang, R. W. Schevitz, Z. Otwinowski, A. Joachimiak, and P. B. Sigler (1988). Flexibility of the DNA-binding domains of *trp* repressor. *Proteins. Struct., Funct. Genet. 3,* 18

4xia K. Henrick, C. A. Collyer, and D. M. Blow (1989). Structures of D-xylose isomerase from *Arthrobacter* strain B3728 containing the inhibitors xylitol and D-sorbitol at 2.5 Angstroms and 2.3 Angstroms resolution, respectively. *J. Mol. Biol. 208,* 129

2yhx C. M. Anderson, R. E. Stenkamp, and T. A. Steitz (1978). Sequencing a protein by x-ray crystallography. II. Refinement of yeast hexokinase B co-ordinates and sequence at 2.1 Angstroms resolution. *J. Mol. Biol. 123,* 15