INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

H4.SMR/775-16

COLLEGE IN BIOPHYSICS: EXPERIMENTAL AND THEORETICAL ASPECTS OF BIOMOLECULES

26 September - 14 October 1994

Miramare - Trieste, Italy

Structure Prediction Methods in Protein

Nicholas D. Socci
University of California at San Diego
La Jolla, - CA, USA

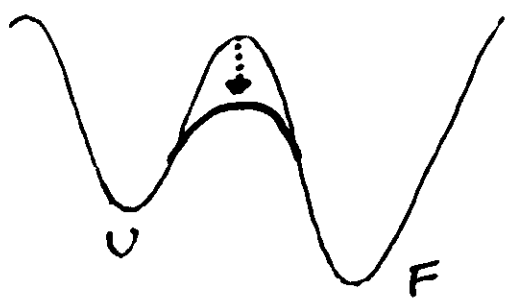
Thermodynamic Hypothesis: seq encodes Struct by determing Free energy Min.

- For some prokins Kinetic factors may also be important.

Although Seq encodes Struct, some prokins need help to fold.

- Disulfide Bonds, normally to strong to break spontaneously. Anfinsen used β-MOH. Cells use Disulfide Isomerase Enzyme with free Cystein on surface
- Prolines cis ⇌ trans also slow (But faster than) compared to other steps: prolyl isomerase.

These two work like any other enzyme: lower free energy barrier

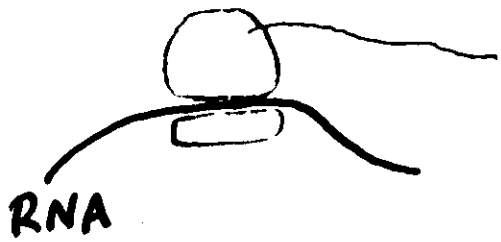


In cells (in vivo) additional problems: Aggregation

Hydrophobic effect drives not only folding but aggregation of several chains

this reduces surface area even more so favored over folding. (Also problem in vitro, for some proteins)

- Cells have high protein concentrations
- Transcription much slower than folding (or aggregat



For a 100 AA protein at the end of transcription 30-40 AA still in ribosome

Cells have Chaperones (Heat shock, related) which bind to unfolde proteins.

DnaK - Binds to End out of Ribosome

GroEL - Binds to Free proteins

- Many Functions - Heat Shock
  - Translocation
  - Assist in formation of oligomers

} Precise mechan not Fully understood

*SOME USE ATP

Interesting Kinetics:

① Proteases: Synthesized in an Inactive form (Zymogen), usually consisting of an extra piece

Pro Region

Varies in size from
10's – 100's of AA

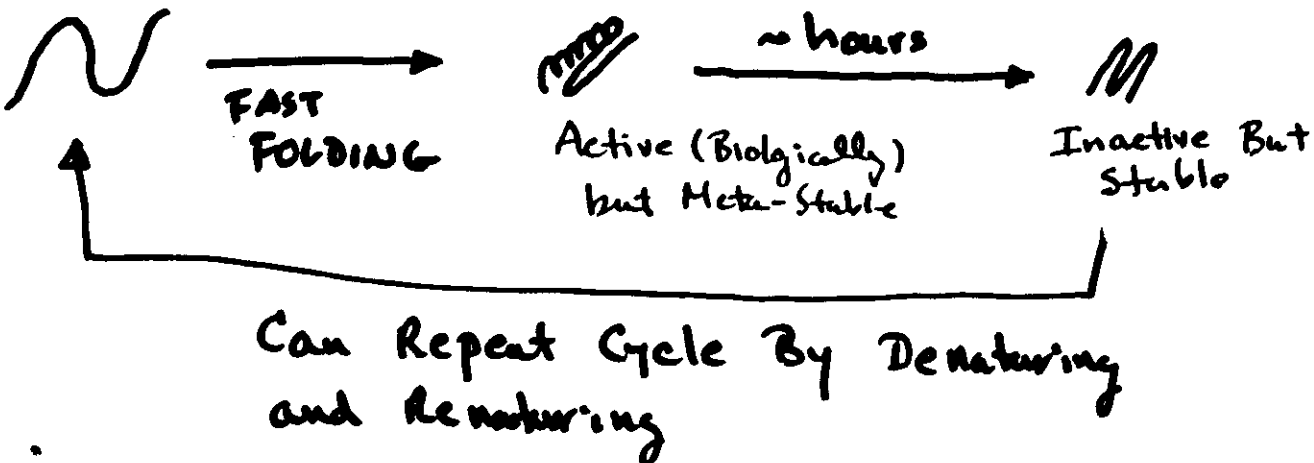
To Activate must cleave or cut off pro region.

However to fold need pro-region

Examples: • Chymotripsin (just cleavage)
• α-Lytic (166 AA pro region)
↓ Lowers barrier

② Serpins - "Kinetic" Switch

Protease Inhibitors



③ Influenza Virus pH barrier (Not Thermo)

What does this mean for the thermodynamic hypothesis?

① Recall Diversity (a.k.a. Annoying Exception)
Always seem to be special cases in biology.

② Are the above special: YES → IMHO

- All seem to have specific FUNCTIONAL Purposes
- Many, Many proteins have the much simpiler behavior (~25 yr of experiments)

Some disaggree: Baker, Biochemistry 33, 7505 (19
feel it may be a more, general feature

③ Restricted Thermo. Hyp. → Global Free Eng Min of accessible states.

But how long are you willing to wait??
(Finite system, everything accessible)

④ No one has shown, "True" (← My definition) metastability
ie NO PATHWAYS (ala Levinthal)

Leave issue of folding for now...

Look at the folded structure of proteins

Not only compact but highly ordered

Structural Hierarchy

- Primary = sequence
- Secondary *
- Tertiary = folded struct
- Quaternary = multi unit systems

* Super Secondary ← Domains

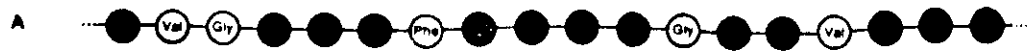
Secondary Structure = Local ordering (conformation) of chain

Regular 2° Struct 50-60%

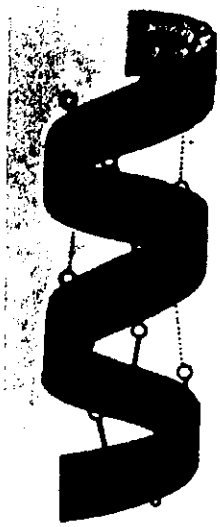
- helix - α most common ($\underset{\text{not obs}}{\pi}$, $\underline{\underline{3.0}}$ rare, triple collagen)
  - chiral
  - right handed [no extended, left handed
- Sheets, β
  - parallel
  - antiparallel
- Turns (3 types) → at surface

- Random Coil

Figure 1-19
Hierarchies of structure.

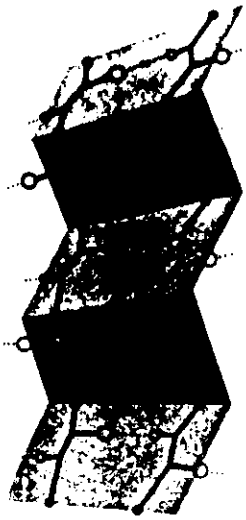


PRIMARY STRUCTURE (Amino acid sequence in the protein chain)



α helix

B



β sheet

SECONDARY STRUCTURE

C

Domains (dark color) in an antibody molecule

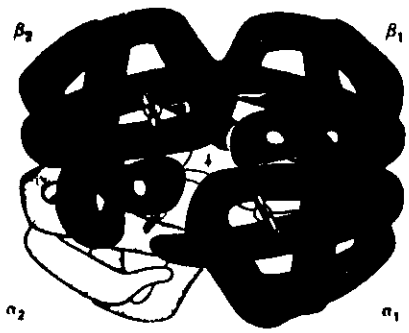
LOCAL FOLDING

D

TERTIARY STRUCTURE

One complete protein chain (β chain of hemoglobin)

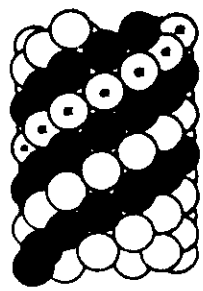
E



QUATERNARY STRUCTURE

The four separate chains of hemoglobin assembled into an oligomeric protein

F



MACROMOLECULAR ASSEMBLY

α (white) and β (color) tubulin molecules in a microtubule.

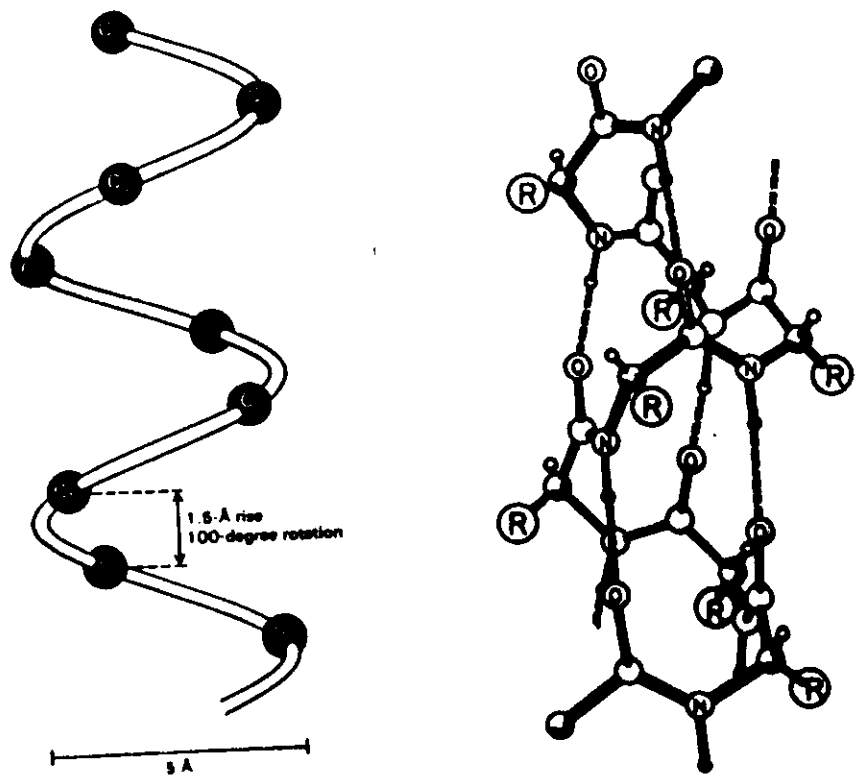
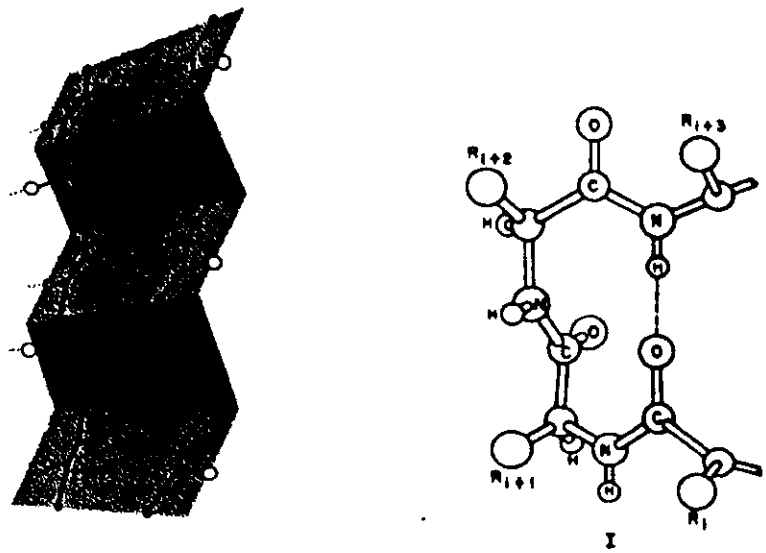


Figure *.4: Two representations of the α-helix. The left picture is a simple structure showing only the α-carbons and the trace of the main chain. The picture on the right is an (almost) all atom picture showing all but the side chains in detail. Note the hydrogen bonding between every fourth residue. Figures from [77, 69].

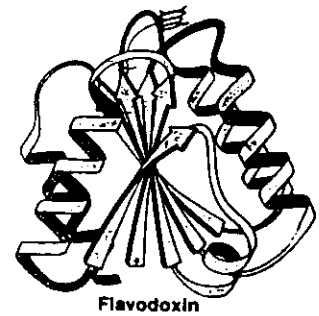


I

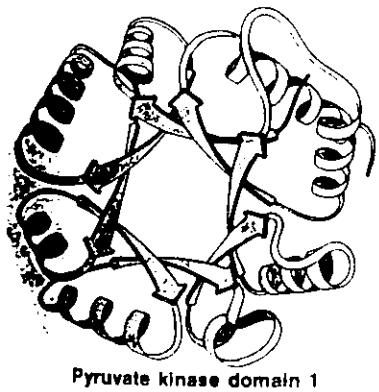
Figure *.5: THe structure on the left is an anti-parallel β-sheet. The amide plans have been shaded in. The figure on the right is a reverse turn.

Triose phosphate isomerase

Lactate dehydrogenase domain 1

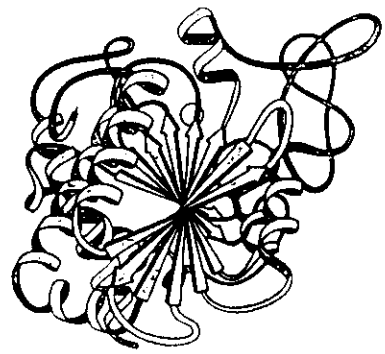


Flavodoxin

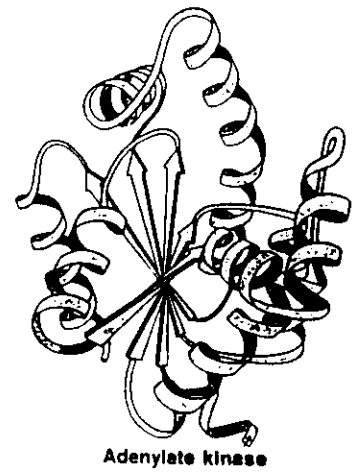


Pyruvate kinase domain 1

(a)

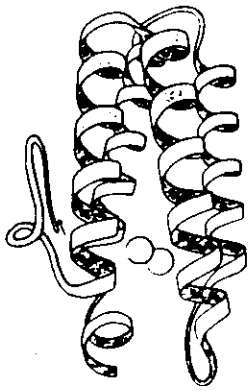


Carboxypeptidase

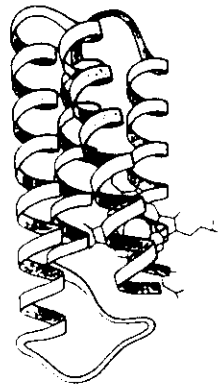


Adenylate kinase

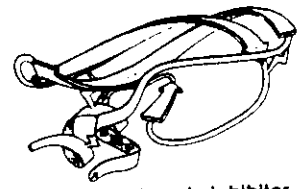
(b)



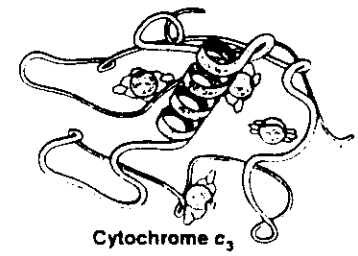
Myohemerythrin



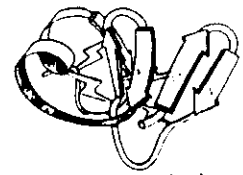
Cytochrome b_{562}



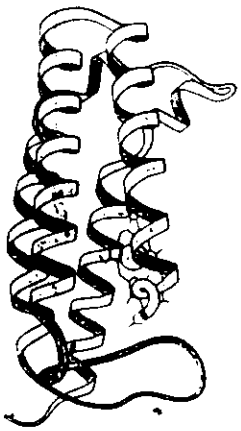
Pancreatic trypsin inhibitor



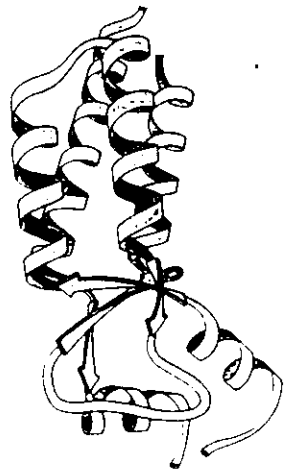
Cytochrome c_3



Wheat germ agglutinin domain 2



Cytochrome c'



Tobacco mosaic virus protein

Cytochrome b_5

Papain domain 1

Examine proteins from a physical viewpoint

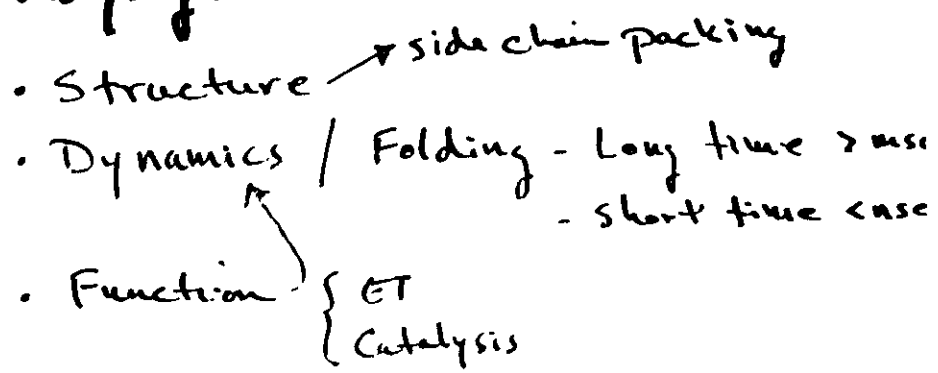
Much of what we know about proteins comes from biochemist. This information tends to be quite detailed and specific.

Let's use ideas from physics and examine various aspects of proteins.

- First use a simplified model for proteins.
- Map real proteins onto this model
- Measuring various physically interesting quantities

Proteins are complex systems: to understand them it is very useful to examine simple protein-like models.

- Remove all but the essential "ingredients"
- What is essential usually depends on what you are interested in studying.



- Probably there is NO simplest Universal (good for all problems) protein-like model

For Folding Really do need to simplify

Eg: All atom + solvent simulations

cf: Wüthrich 2 nanosecs ≅ 3 months supercomputer CPU Time

Perspective

1 hr ~ $1,000 (1.500.000 L) Very generous

~2,000 hrs OR $2,000,000 (3. Billion L)

Not a calculation for mortals

But even if you had money still no good

Proteins take at least ~1 millisec to fold

⇒ 12.5 years at current computer power

Assume that computers double in speed every two years, it will be 20 years until we can fold all atom simulations on Month time scales.

Protein Folding:
A General View

Nicola Socci

penultimate Lecture → Tomorrow WILL NOT Be about Structure Predicti

---

Look at proteins again, but with a different perspective (new eyes)

- Put our glasses on (better metaphor, take them off if you are myopic (nearsighted))

- Coarse Grain: qualitatively (renormalize)

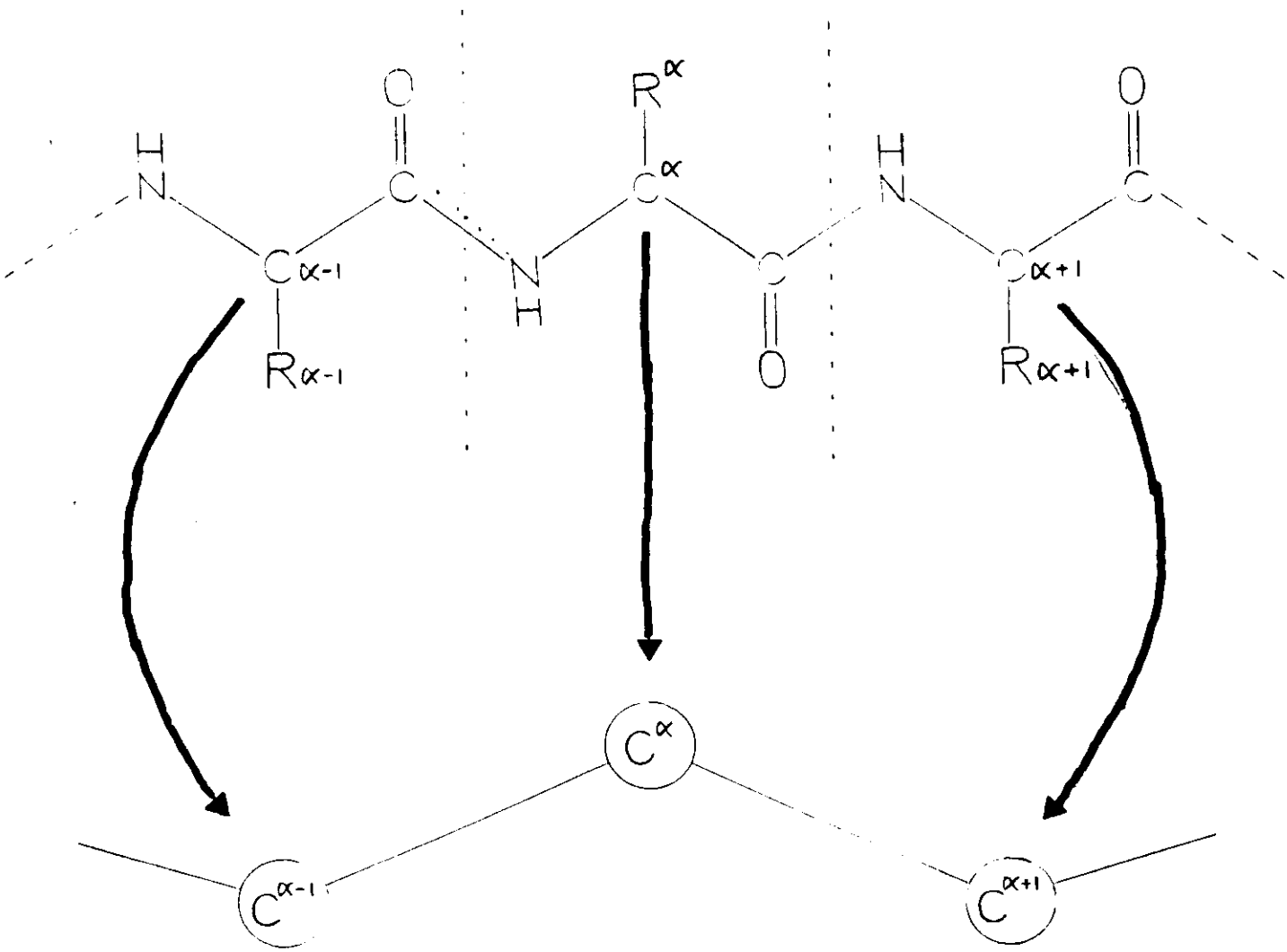
- Fancy way of saying we are going to ignore most of the atomic level detail.

- However; remember not to throw away too much

STUMP THE BAND

Simple models for proteins.

- Reduced description of proteins.



First Step: Use the simple model to understand real proteins.

Use the model to simplify acutal protein structures and then study various statistical properties.

Objective: Look for universal features shared by all* proteins, rather than specific features of individual proteins

- i.e. what are average (typical) protein properties

- Specific Aim: Define 2° structure in a simple generic way

* OLD caveat: "all" means globular, water soluble, crystallizable proteins since we will be using the PDB

eliminate redundant examples

Take the Protein Data Bank of roughly 110 proteins and measure the following quantities average over this ensemble:

1. Radius of gyration. Scaling behavior.

2. Pair correlation function. Potentials

3. Angle distribution functions

4. Angle correlation functions. (Secondary Structure)

Goal: • Understand the universal features of protkins

• Future Work - Need to define Hamiltonian (potential). Only have half of the model. By looking at real protein determine the potential "parameters"

Radius of Gyration & End to End Length

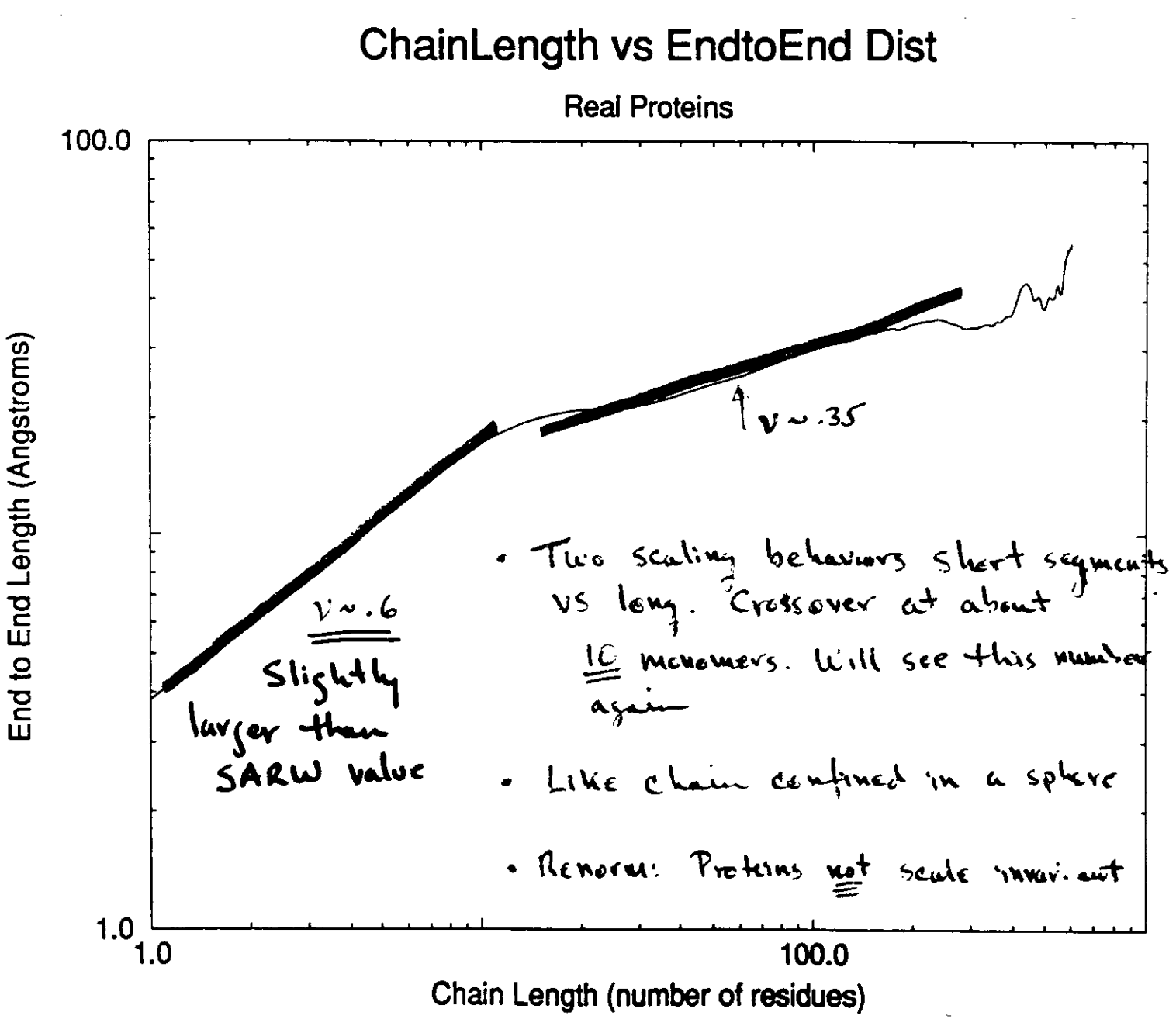
$$R_{Gyr} = \sqrt{\frac{\sum_{i=1}^{N} |\vec{r}_i - \vec{r}_{com}|^2}{N}} \quad , \vec{r}_{com} = \frac{\sum_{i=1}^{N} \vec{r}_i}{N}$$

For sphere  $R_G = R$,  Uniform $R_G = \sqrt{\frac{3}{5}} R$

<u>End to End</u> plot number monomers along chain vs seperation in space

Do it for all subsegments of the chain

ie distance between all 2 monomers
3
4
⋮
N


ChainLength vs EndtoEnd Dist
Real Proteins
100.0
1.0
1.0
100.0
End to End Length (Angstroms)
Chain Length (number of residues)
ν ~ .35
ν ~ .6
Slightly larger than SARW value
• Two scaling behaviors short segments vs long. Crossover at about 10 monomers. Will see this number again
• Like chain confined in a sphere
• Renorm: Proteins not scale invariant

Pair Correlational Function (radial):
(Radial Distribution Function)

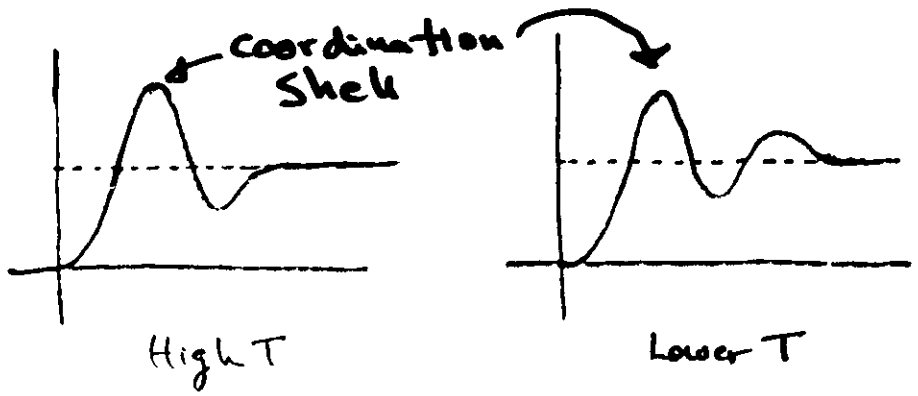
Gives Probability two monomers seperated by a distance R

$\rho(R)$ ..... R

Note Normalized by spherical shell $\frac{1}{4\pi R^2}$ ← Dr. Bohr showed unnormalized ones

For Crystal sharp peaks

For a Liquid



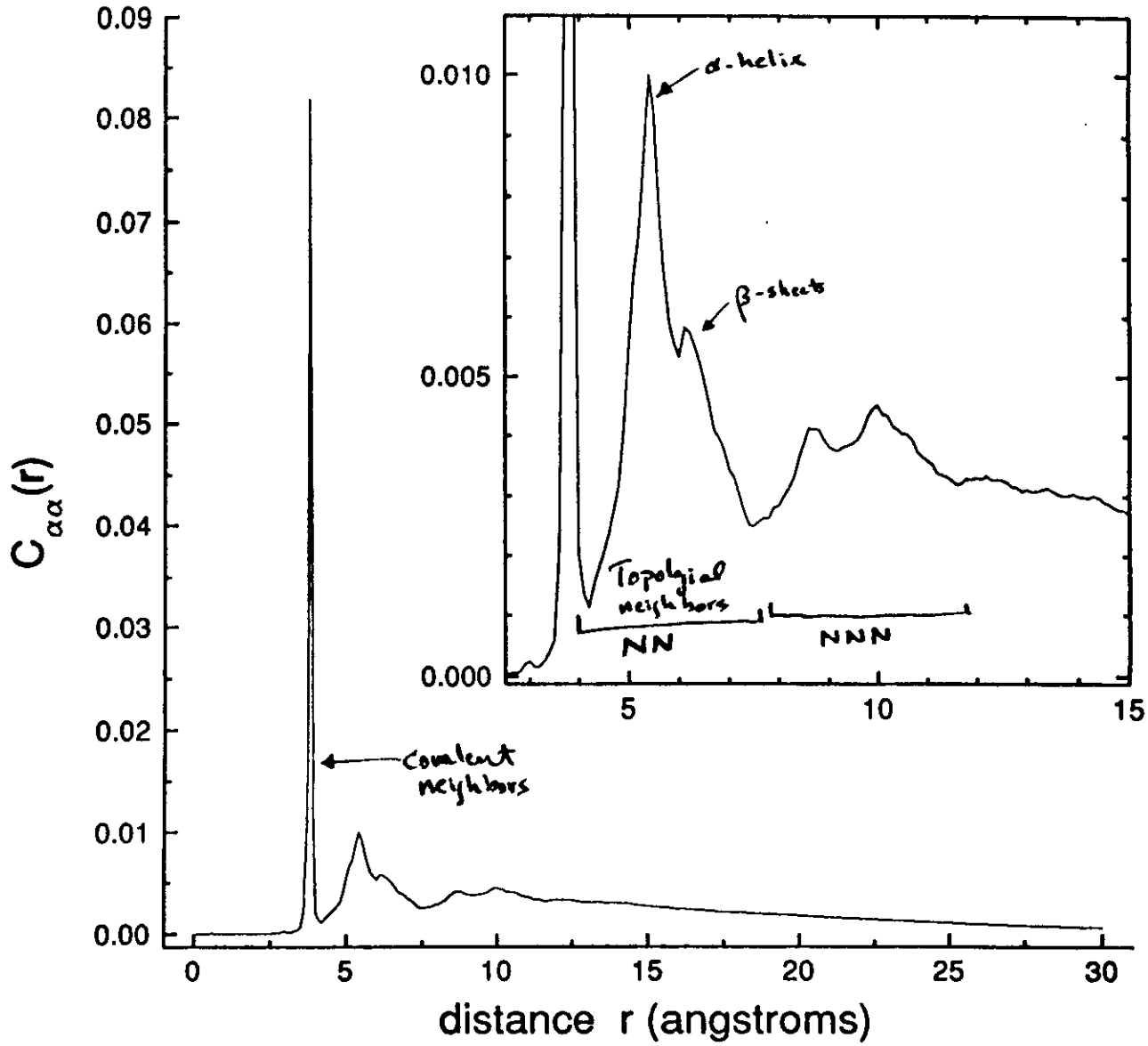
Can Also measure mixed correlation function

ie $\rho_{HP}(R)$ ..... Polar

PP
HH

Hydrophobic

What is the prob that a Polar monomer is R from a hydrophobic one



① Two nearest neighbor length scales (Lattice problem (simple))

② Well defined coordination shell

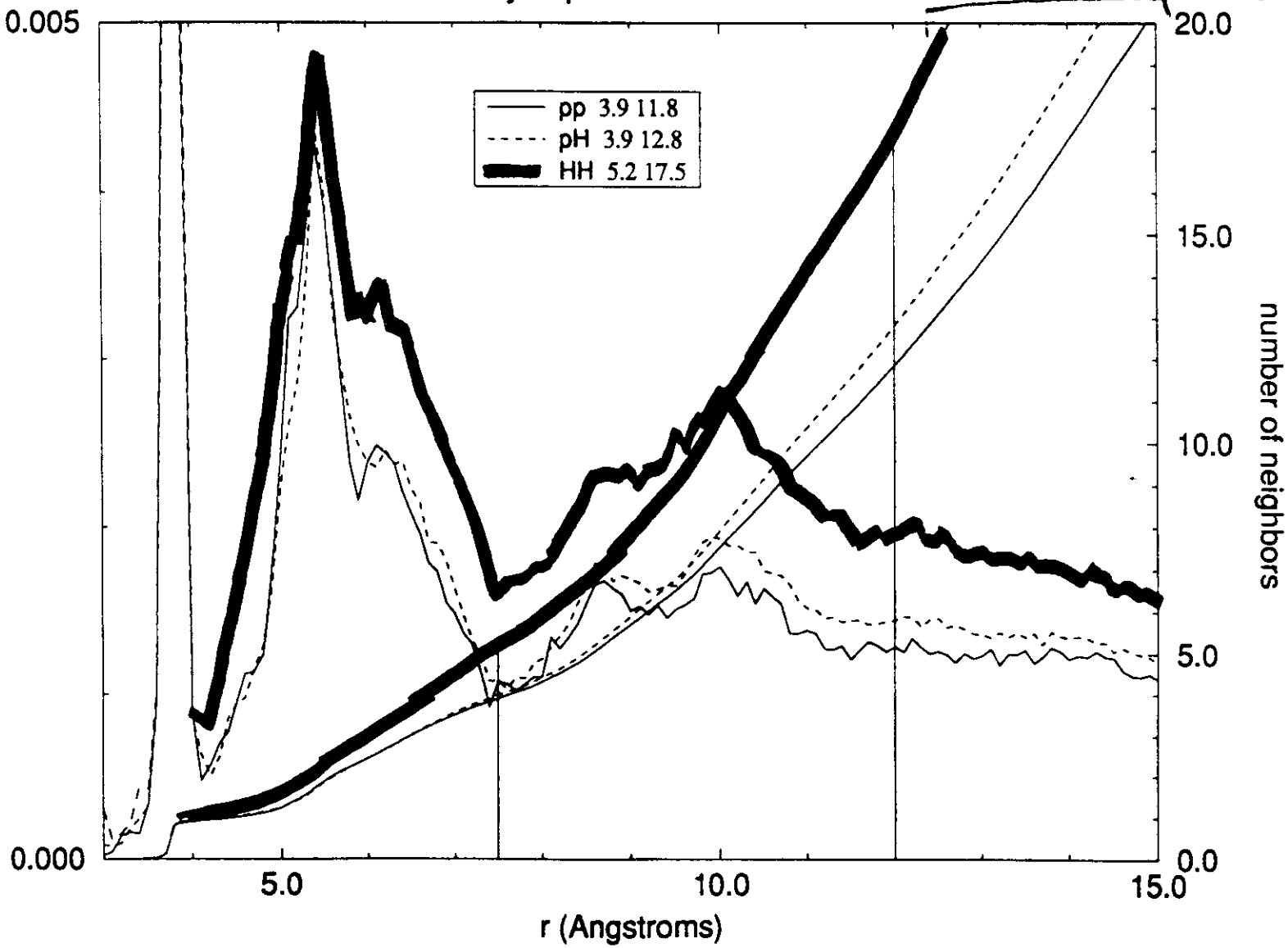
③ Bifurcation of peaks ⇔ 2° Structure

④ Potentials BBGKY or chem log

Real Proteins

Pair Correlation Function

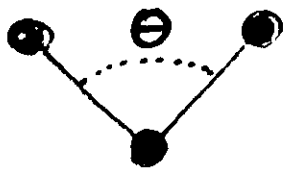
$4\pi \int_0^R \rho_{ij}(r) r^2 dr$ # of monomers in a shell of Radius R



Angle Distribution Function.

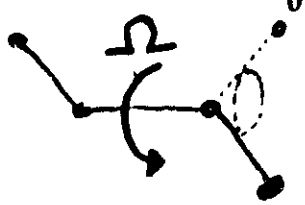
- Simple model also has ~2 angle degrees of freedom per residue just like proteins. But slightly different.
- Examine the distribution of angles to study the local structure of protein chains

① Bond angles: in proteins bond angles fixed here it becomes free



- Excluded Volume
- Bifurcation: 2 peaks (α, β + everything) but very different in size and shape
  α-helix - better defined than β-sheets

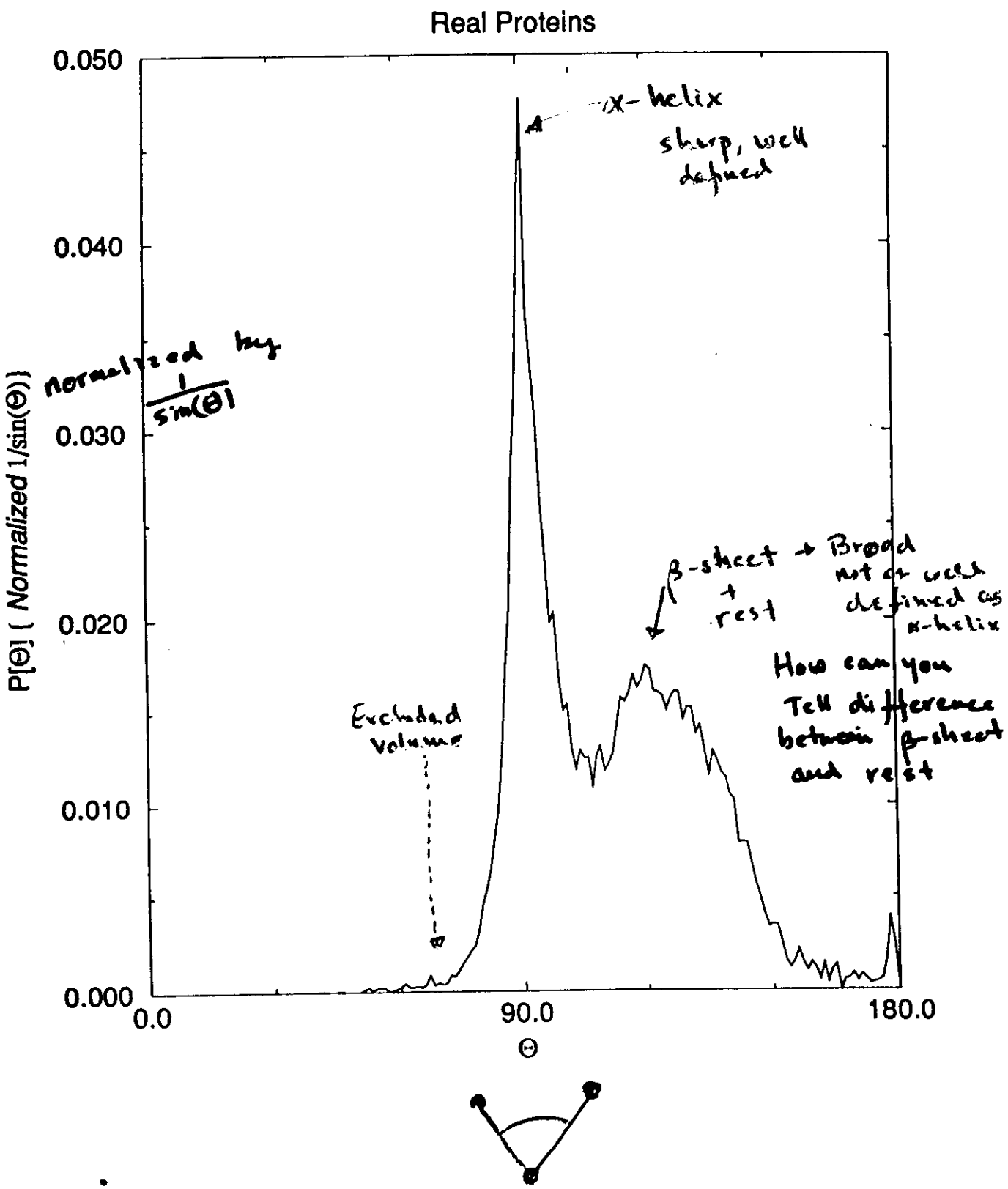
② Torsion (Dihedral Angle): Not new used back in 1972! to define 2° struct.



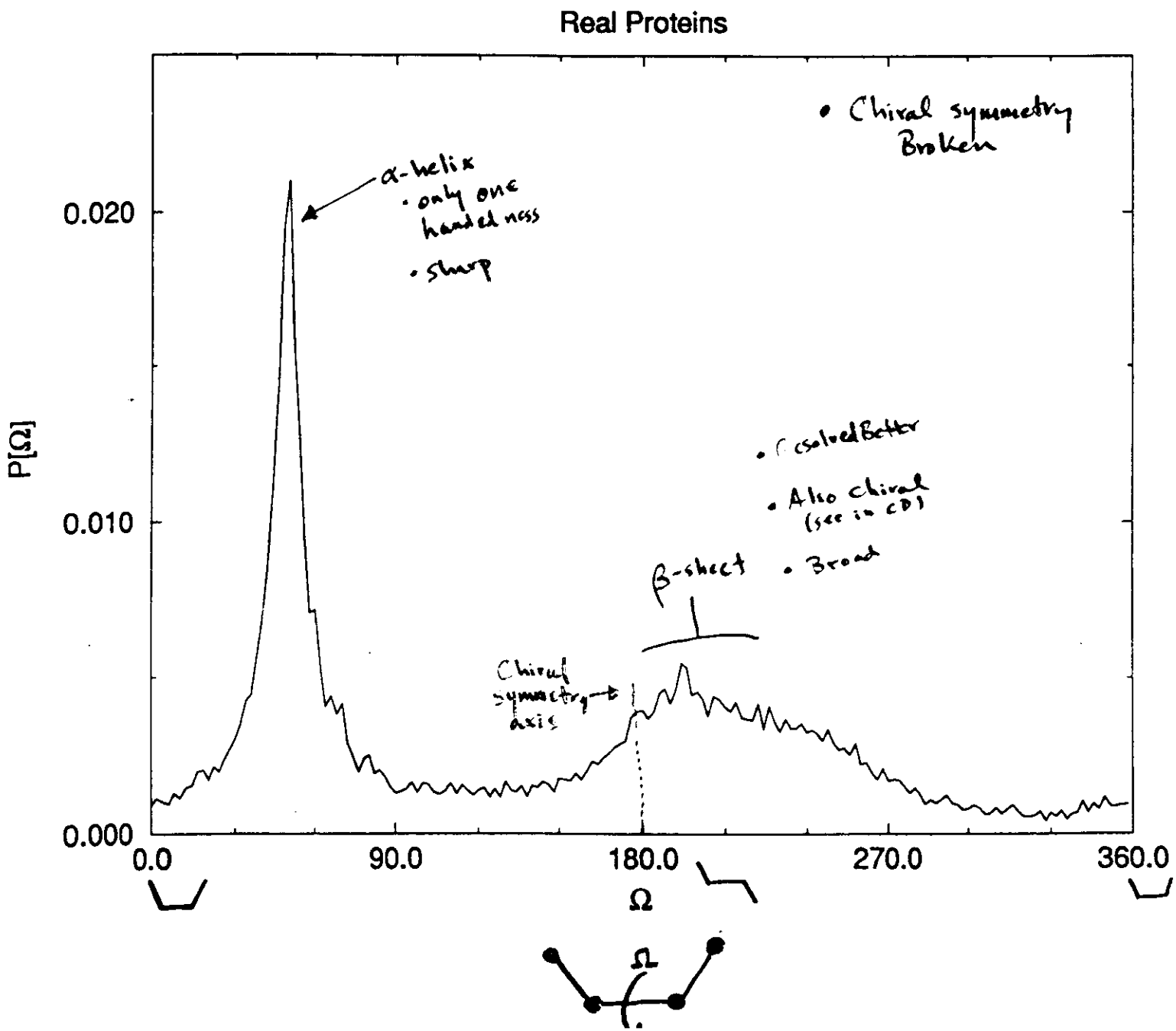
- Bifurcation again. Peaks different
- Chiral symmetry broken ($\phi \to \phi \pm \pi$)

③ Quasi Ramachandran Plot

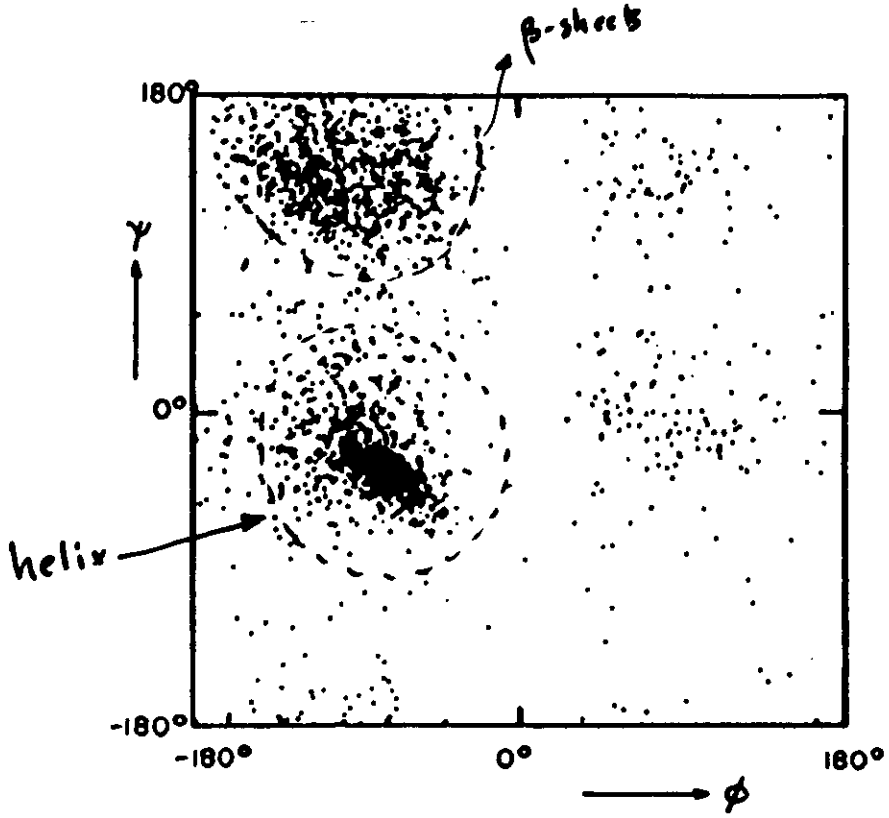
Planar Angle Distribution Plot



Torsion Angle Distribution

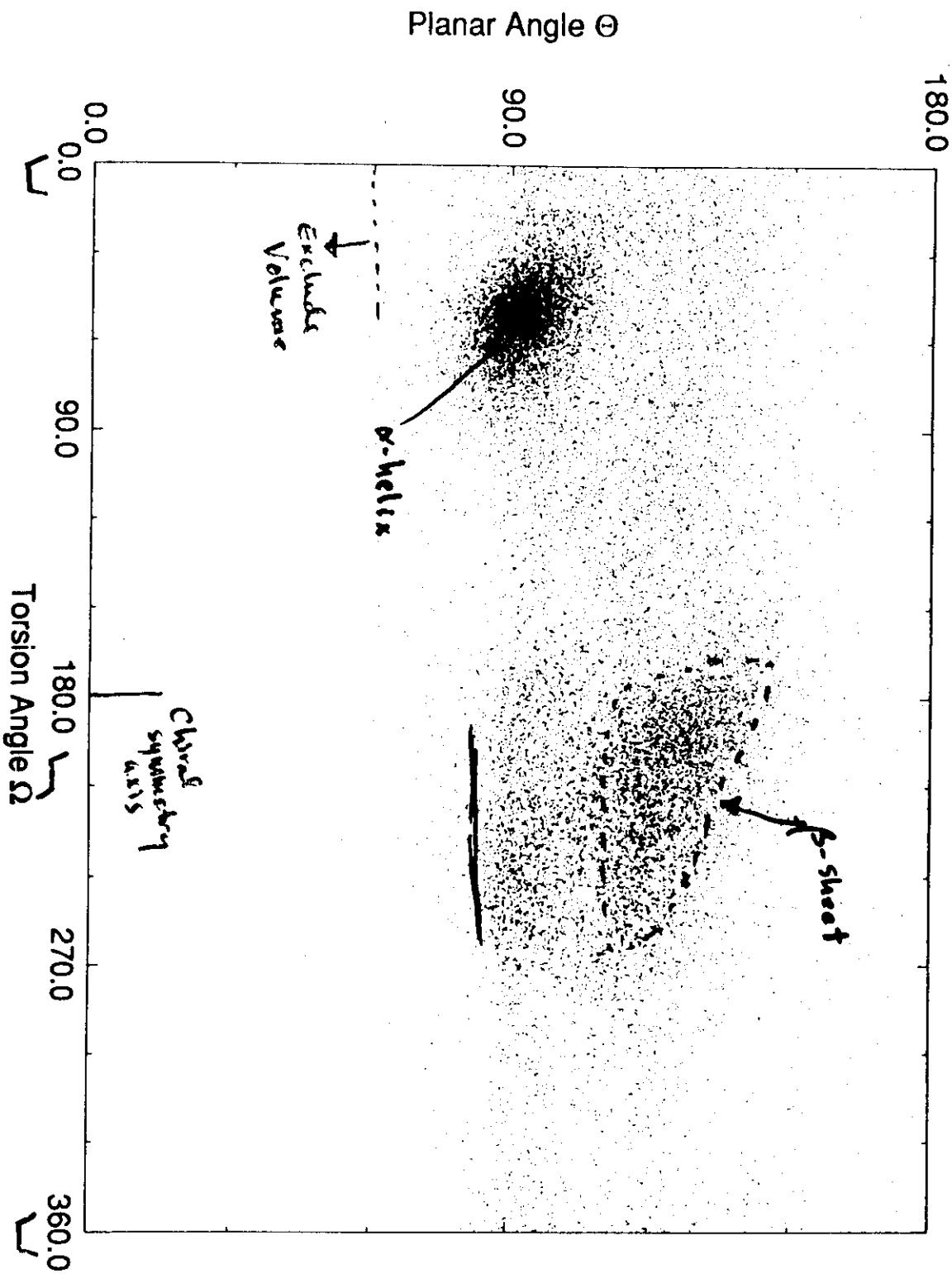


Ramachandran Map



- Note usually map is made by looking a steric interactions in di, tri peptides

- This density was taken from the PDB


Torsion/Planar Rham Plot
Real Proteins
Planar Angle Θ
Torsion Angle Ω
0.0
90.0
180.0
270.0
360.0
0.0
90.0
180.0
α-helix
β-sheet
Excluded Volume
Chiral symmetry axis

correlations ≡
• Green Fun
• Kernels
• Propagator

Chain Correlation Functions

previous corr. functions where through space

now let us look at correlations along the chain

$C(\Delta)$ where Δ is the separation (in number of residues) along the chain

- First Look at Sequence Correlations: just polar vs Hydrophobic

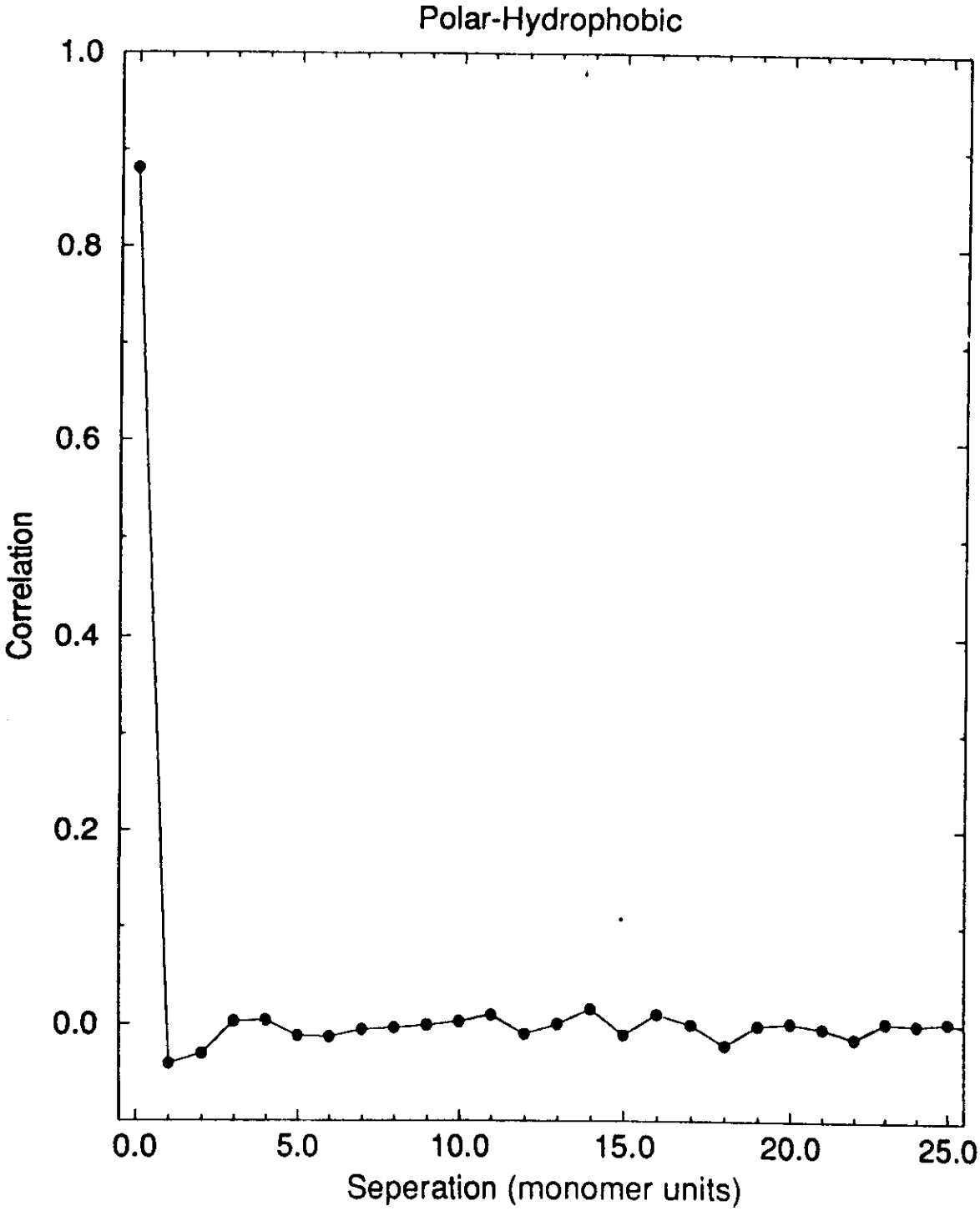
ppHHHpHppHHpHHHpHppp

Δ

- No correlations!
- Same is true if one measures corr of the 20 Letters AVTGIN ....
- This is one difficulty in predicting 2° struct from local sequence (Identical penta-peptid

N.B. Random Sequences of AA not proteins, yet Protein Sequences "appear" random!?!

Chain Correlation

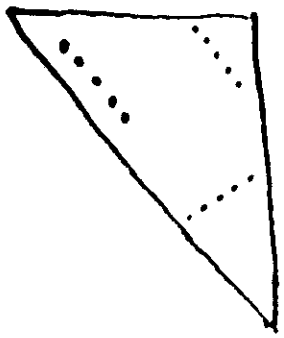


Chain Correlations Cont.

What is secondary structure?

Biology Definition: Keys in on H-Bonds

- Define 2° structure (recognize) by Look for H-bonding patterns. (Kabsch & Sander)



$i \cdots i+4$, $i+1 \cdots i+4+1$ α-helix

$i \cdots i+j$, $i+1 \cdots i+j+1$ parallel β-sheet

$i \cdots i+j$, $i+1 \cdots i+j-1$ anti-parallel β

$i \cdots i+3$ $(i+4)$ turn

- Not Quite so simple: what is H-bond to what not also easy to determine

- Nice definition but can be too specific, nor is it easy to calculate,
- Bigger Problem: our simple model does not have Hydrogens (no H-bonds)

Can we define secondary structure in our simple models??

Yes, we can

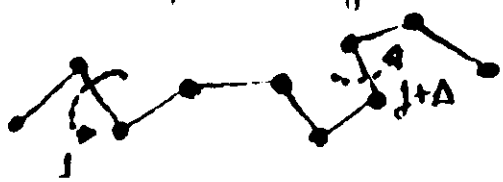
Re Read Bio Def: This time key in on regular.

Regular Arrangement of chain, ie Correlated along the chain.

But what is correlated?

- Not Sequence
- Angles

Measure Angle - Angle correlation Function



Not through space but along chain

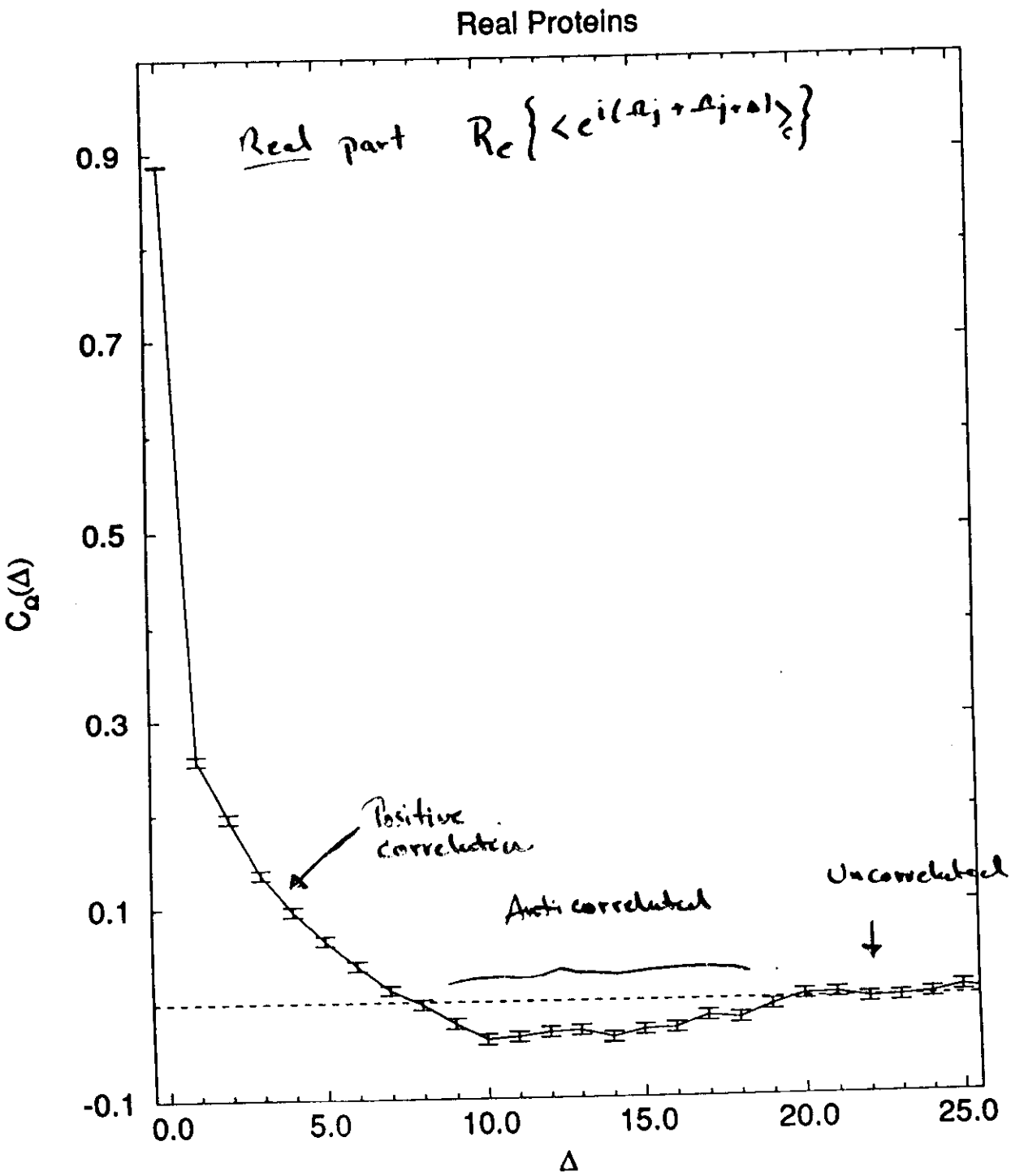
$$\left\langle e^{i(\Omega_j - \Omega_{j+\Delta})} \right\rangle_C$$ ← Connected correlation (ie subtract mean $\left|\langle e^{i\Omega_j}\rangle\right|^2$)

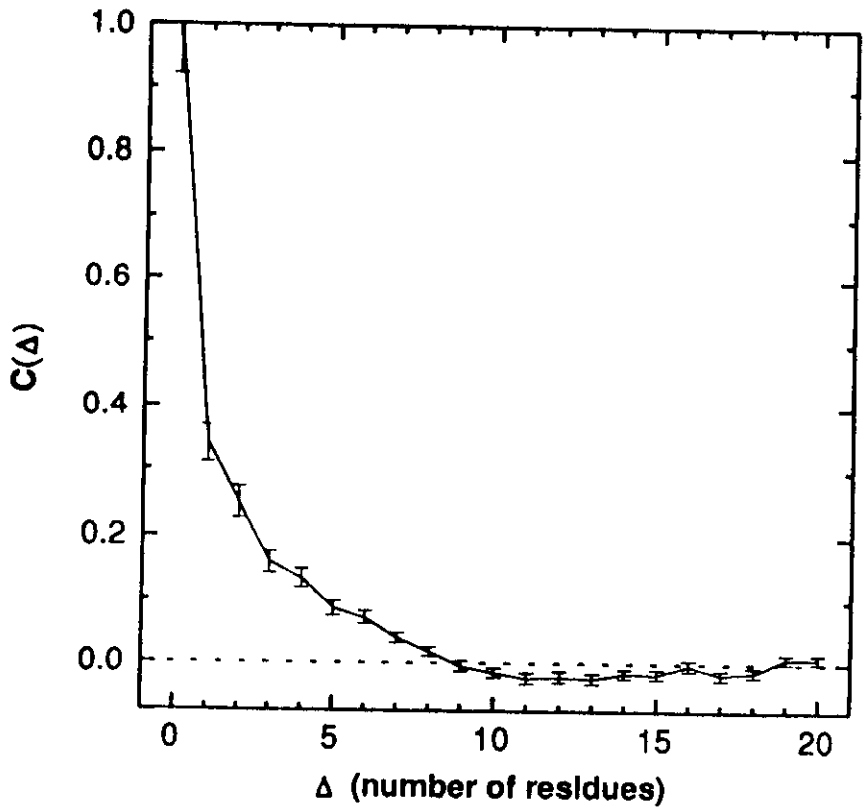
- Long correlations: up to 10 AA

  10 is the "average" size of 2° in proteins

- Anti correlated piece: super secondary struct

Torsion Angle Correlation





Real (θ)

Simple models continued:

- Simplified Potentials
  - Dominant force in protein folding: compactification of structure due to hydrophobic effect.

$$\mathcal{H} = \sum_i \frac{1}{2} k_{\text{cov}} \left(|\vec{r}_i - \vec{r}_{i+1}| - l_{\text{cov}} \right)^2 + \frac{1}{2} \sum_{ij} V(r_{ij})$$

Covalent Term

$l_{cov} = 3.78$ Å

$k_{cov} = 1$

Every thing else

$r_{ij} = |\vec{r}_i - \vec{r}_j|$

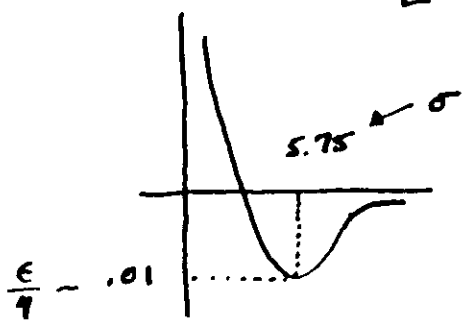
  - Pick $V(r_{ij})$ to compact protein.
  - Start with simple two body interactions.

Goal: Find a protein-like potential for later study in folding and dynamics

3 potentials

① LJ

$$V(R) = \epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$



② Rad Gyr

$$V(\{\vec{r}_i\}) = \epsilon\left\{\sum_{i<j}\left(\frac{\sigma}{r_{ij}}\right)^{12} + \sum_{i=1}^{N}\left|\vec{r}_i - \vec{r}_{com}\right|^2\right\}$$

$\epsilon = .01$ σ pick by look at $\rho_{aa}(R)$

③ Rad Gyr + Chiral Local (chain) Int

$$V_C = \hat{s}_{i-1} \cdot (\hat{s}_i \times \hat{s}_{i+1})$$

"Fold" (compact) the Chain.

Minimize the Energy

Use a simple function minimization routine to minimize the potential energy.

Possible Problem: Model is a homopolymer will have many (e^N) low energy structure that look very different.

Make the follow conjecture

Each one of the many different low energy structs of the homopolymer will correspond to the Native structure of a <u>different</u> seq in a related heteropolymer model

So by generating a set of compact structures we can then average over sequence, and repeat the analysis we did for real proteins

Score Card.

How protein-like are the potentials

- No potential 100% protein-like

Each seams to capture different structural features of real proteins

- No Secondary Structure !

Need Better Potentials: Possibilities

① Angle Potentials (Explicit)

② Combine the two (LJ - Radgyr)

③ Heteropolymer. Maybe Conjecture is Incorrect.

④ Model too simple :
- Side chains
- ....

More Questions (Physics)

- Hetero vs Homopolymer
  - Hetero ⇓ # of gs small O(1)
  - Homopolymer ⇓ Many gs ~ N^{γ}

How many monomers do you need to get a unique gs? RNA-4
Lattice Models 2 — Experiments 2 (HP)

- Real Seq vs Random - 20^N seq what % folds?
  - Qualitative Ans ??
- Minimal Information to Solve folding Problem
  - Complexity NP?? (Lattice Proof)
    Problems? ⇒
- Inverse Folding Problem
  - may be easier
  - possible more useful comercially

Conclusions (Philosophical)

- Proteins are an extremely complex system, with a wide diversity in structure and function.

- On close examination, they appear to share common (universal) features. Theory (ies) of proteins seems possible.

- Simple systems can sometimes have very complex and interesting properties

  Maybe much of the apparent complexity of proteins can be understood from looking at simpler model systems

- We study proteins to study how living organisms work (to study life):

  But understanding life by looking at protein function may be as difficult as trying to understand proteins from e⁻-photon interactions