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

SMR.961 - 14

WORKSHOP ON:
PROTEINS, MEMBRANES and their INTERACTIONS

22 JULY - 2 AUGUST 1996

"Principles of protein structure"

Marc DELARUE
Institut Pasteur
25-28 Rue de Dr. Roux
75015 Paris
FRANCE

Principles of protein structure

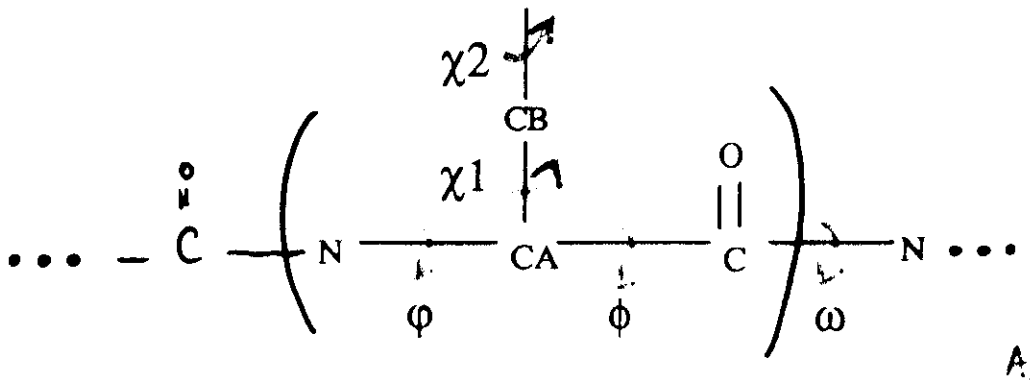
Overview.

1st course: Learn how to generate protein chains with correct
- stereochemistry
- shape
- distribution of (φ, ψ)
- self-avoiding.

2nd course: Learn how to generate side chains with correct packing - Mean Field theory -
What is the potential energy to be minimized - Full atom models -

3rd course: The so called "inverse protein folding problem" - Sequence space available to a given fold.
What is the amino acid - amino acid interaction table to be used - How to optimize it.

PROTEIN CONFORMATION



A) BACKBONE

1) defined by 3 dihedral angles : φ, ϕ, and ω

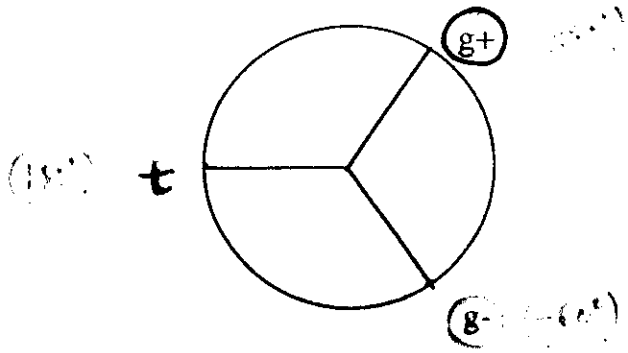
2) Planarity of the peptide bond : ω = 180° or ω = 0°

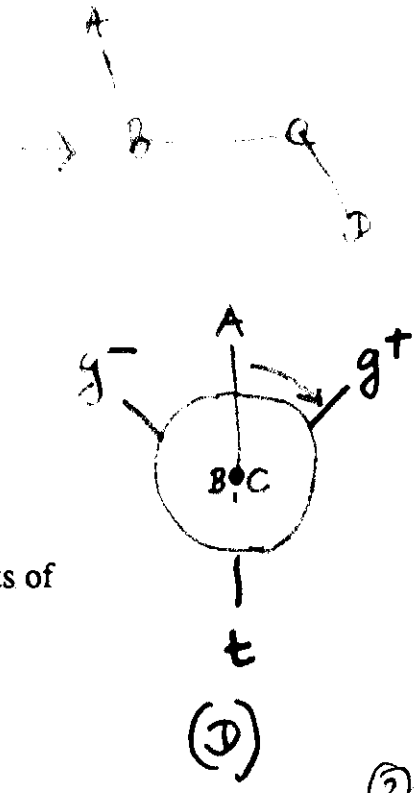
3) For most residues, the (φ, ϕ) conformational space is limited (Ramachandran plot)

B) SIDE-CHAINS

1) Defined by χ dihedral angles

3) 3 possible conformations for each χ angles



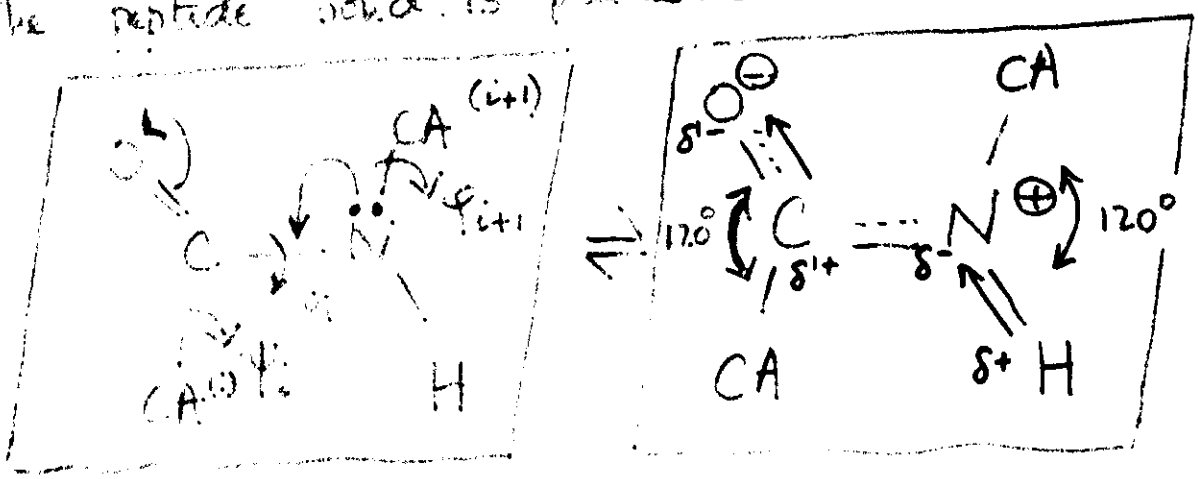


3) Side-chains are observed to cluster in limited sets of conformations, called rotamers

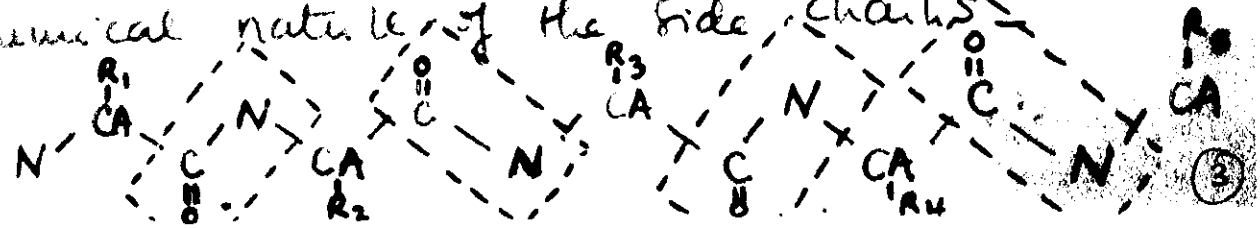
Protein representation

- The main chain

The peptide bond is planar -

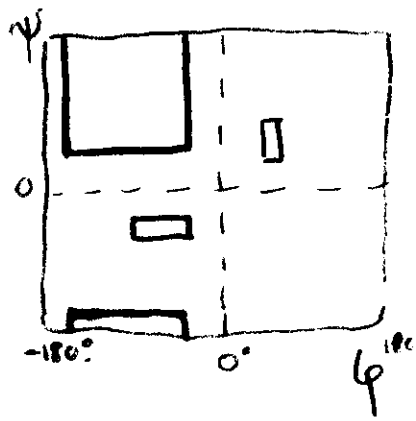


- $d(C_\alpha^i, C_\alpha^{i+1}) = 3.8\ \text{Å}$ always
- $\omega_i \simeq 180°$ always (trans ; cis is rare).
- There is a small dipolar moment associated with every peptide bond.
- The carbonyl is a good acceptor and the NH group is a good donor of hydrogen bonds ; this accounts for most regular structures (secondary structures) of the backbone : α helix & β strands -
- The specificity of the sequence is reflected in the chemical nature of the side chains



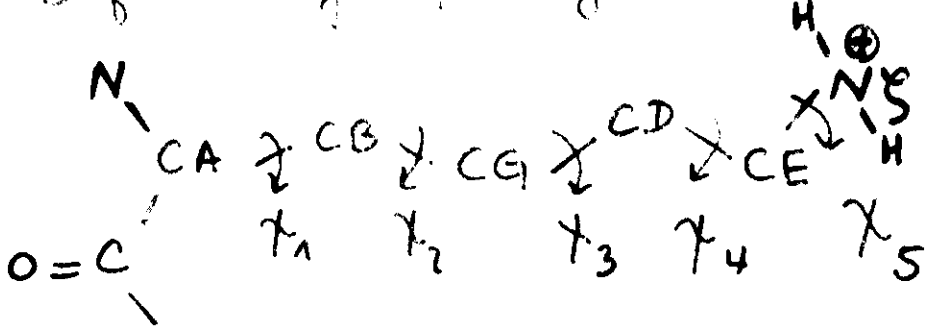
The side chains

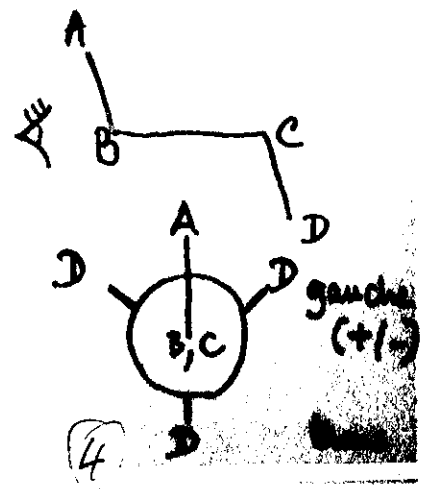
- small : Gly, Ala, Pro [cyclic) $\Rightarrow \phi = -60°$]
- hydrophobic : Val, Ile, Leu, Met, Cys (+ sulfur atoms)
- Aromatic : Phe, Trp, Tyr
- Charged ⊖ : Asp, Glu
  ⊕ : Lys, Arg, His
- Amide Groups : Asn, Gln
- Small & polar : Ser, Thr



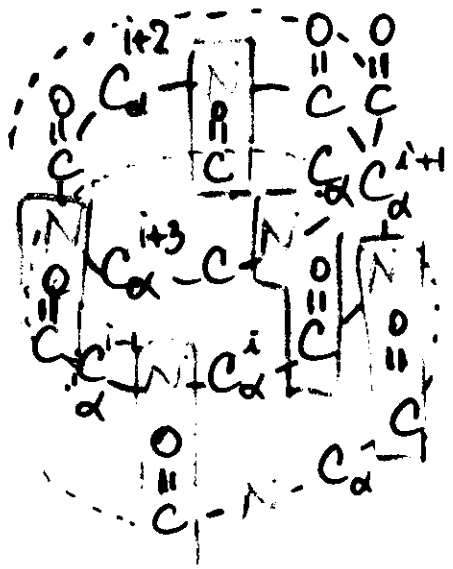
- All of them (except for Gly) have a C_β (at least one χ_1 angle). This means only some regions of the (ψ, ϕ) diagram (the Ramachandran plot) are accessible because of steric clashes between C_β & C(=O)

- Definition of χ angles :





Secondary structure elements



- α helix – Pauling (1953)

1.5 Å / residue

3.6 residues per turn

5.4 Å / turn

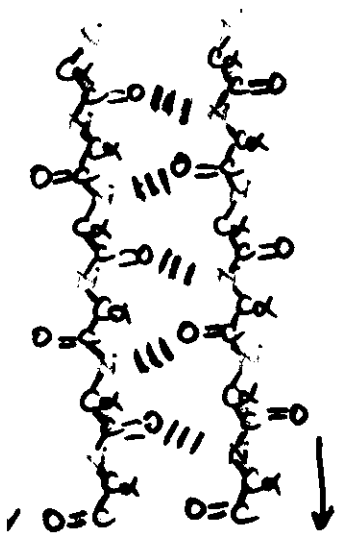
~ 5.6 Å wide

all peptide dipole moments are aligned.

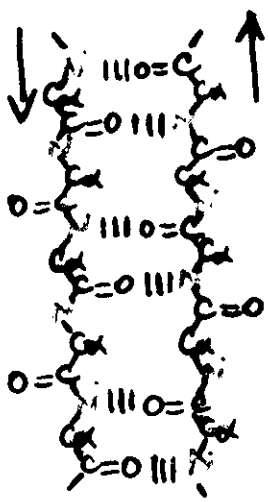
- β sheet

2 Å per residue

Parallel



Antiparallel



side chains pointing alternately up and down (above & below the plane of the sheet)

twist of β sheet difficult to predict.

Number of variables for a protein (Nseq)

- suppose $Nseq = 250$ aa $\Rightarrow$ MW $\simeq$ 28 kD

(110 Da / a.a).

- There are about 8 atoms / aa, on the average
  - 4 (N, C_α, C, O) for main chain
  - 4 ($C_\beta, C_\gamma, O_{\delta_1}, O_{\delta_2}$) for side chain.

average weight of an atom is 14

- if cartesian coords

$\Rightarrow 250 \times 8 \times 3 \sim \underline{6 \cdot 10^3}$ variables

- if internal (internal) coords

$\Rightarrow 250 \times 4 \sim \underline{10^3}$ variables.

The information missing is contained in the stereochemistry
- fixed bond lengths
- nature of peptide bond (planar, $\omega = 180°$)
- fixed bond angles : tetrahedral for C
  trigonal for N

- if globular & spherical $R \simeq 24\,\text{Å} - 25\,\text{Å}$.

$V \simeq 42\,000\,\text{Å}^3$

Other simplified representations of proteins

② CA only (with the constraint CA-CA = 3.8 Å)

⇒ it is the most coarse possible to reconstruct the whole backbone, using fragments of known proteins deposited in the PDB satisfying minimum chord with given Cα positions.

(chord check < 1 Å).

[illegible] between Cα's.

[illegible] : N×N.

: only binary information

$V_{adj}(i,j) = 1$ iff $d_{ij} <$ cutoff.

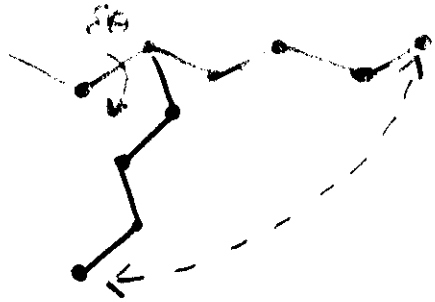
It is that kind of information that is obtained from NMR experiments.

Simulate the dynamical behaviour of a protein

(A) cartesian coords $m \frac{d^2 \vec{r}_i}{dt^2} = -\vec{\nabla}_{r_i} V$

- Many variables - Add solvent molecules -
- integration time step limited by movement of highest frequency - bond distance oscillations ...

(B) write down eqns in internal coords : pb of downstream consequences of one small perturbation of one angle

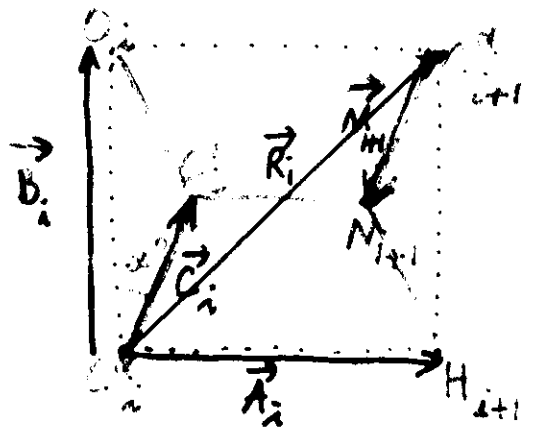


solutions - R. Abagyian (NY) - lagrangian etc ...
- AT Brünger - Robotics & Kalman Filters..
- The so-called "local moves" arrange to make moves that leave endpoints invariant.

The Knapp algorithm for "local moves"

Knapp, J. Comput. Chem., 14 : 19-29 (1993).

- idea representation

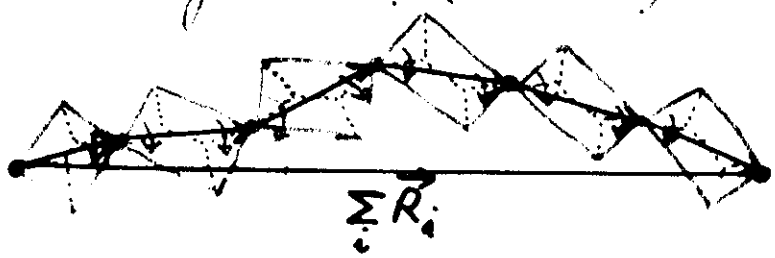


$l = 1.536\ \text{Å}$

α tetrahedral $\Rightarrow \vec{n}_i \cdot \vec{C}_i = -\frac{1}{3}$

$\vec{n}_i \cdot (\vec{C}_i \times \vec{C}_{i+1}^{\alpha}) = \vec{A}_i + \vec{B}_i$

- window algorithm (n windows)



$$\begin{cases} \vec{n}_{i+1} = \frac{1}{2}\vec{a}_i - \sqrt{3}/2\ \vec{b}_i \\ \vec{C}_i = \frac{1}{2}\vec{a}_i + \frac{\sqrt{3}}{2}\vec{b}_i \end{cases}$$

$$\begin{cases} \vec{a}_{j+1} = \frac{1}{3}\vec{a}_j - \frac{2\sqrt{2}}{3}\vec{b}_j \\ \vec{b}_{j+1} = \frac{2\sqrt{2}}{3}\vec{a}_j + \frac{1}{3}\vec{b}_j \end{cases}$$

window $\times$ 2 = 4 variables can be picked at random

the 4 remaining angles are given by:

$$\vec{C}_4 \cdot \prod_{i=1}^{3} \bar{C}_i \bar{N}_i \vec{n}_4^{(0)} = -\frac{1}{3} \quad \text{if n window} = 4$$

$$\sum_{i=1}^{3} \vec{R}_i = \sum_{j=1}^{3} \vec{R}_j^{(0)} \quad \text{where} \quad \vec{R}_j = \prod_{k=1}^{j} \bar{C}_k \bar{N}_k \vec{R}_j^{(0)}$$

- approximate $\bar{C}_j = 1 + \psi_j \vec{C}_j$ ⊗ and solve this linear set of eqns

(linear approx.) $\bar{N}_j = 1 + \varphi_j \vec{n}_j$ ⊗

[illegible] description of [illegible] moves algorithm

⇒ Most of the moves will generate steric clashes

1) Expand the molecule $d \to d(1+\varepsilon)$

2) [illegible] restraint at random, [illegible] eqns, & modify restraint → restraint + [illegible]

(3) check [illegible] [illegible] steric clashes)

Do 1, 2, 3 many times ...

check, from time to time, if it is possible to recompress the molecule. If not more [illegible] [illegible] [illegible] local minimum —

[illegible] [illegible] automatically enforce stereochem

Molecules

- [illegible] Lagrange multipliers instead of additional Energy terms. Solve interactively huge matrix (sparse one) —
- Excellent [illegible] [illegible] display interface

[illegible] [illegible] [illegible] commercially.

How do we approximate the main chain with one with idealized geometry?

1. : minimize some $\text{rmsd}^2 = \sum_i \left(x_i^{\text{real}} - x_i^{\text{idealized}} \right)^2$

which is a global criterion.

However, we have to grow the chain, one residue at a time, i.e. make choices based on local criteria!

It is not a choice that is locally optimal with respect to the criteria for best fitting of the whole chain.

: Grow a population of chains of length $n \rightarrow$

(... - Evaluate their Energy -

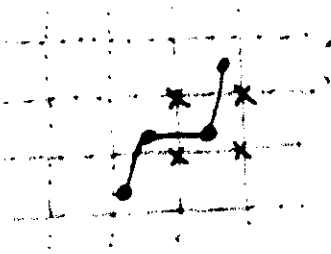
Replicate the most successful and give them weight $\propto \exp -\left(E^{(n+1)}_{\text{chain}\,\alpha} / k_B T \right)$.

Normalize to keep population constant.

$\rightarrow$ Highest Temperature as in Annealing -

$\rightarrow$ See Gunsteren, Berendsen for more Detail.

Projections of protein mainchains on lattices.



Minimize

$$F_{err}(R_1 \dots R_N) = \sum_{i=1}^{N-1} U_i (R_i - R_{i+1}) \quad \text{connectivity}$$

$$+ \sum_{i=1}^{N'} f(x_i - R_i) \quad \text{error function}$$

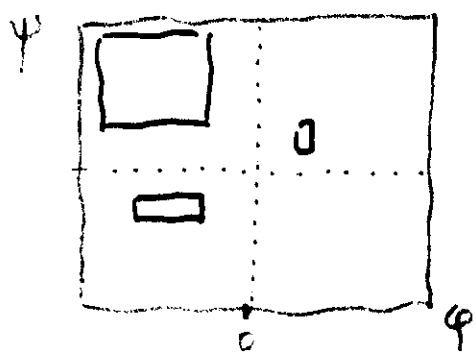
- This can be solved by linear dynamic programming methods (Finkelstein & Reva).
- The fold must strongly depends on the orientation of the protein with respect to the lattice
- Adding some other penalty for self avoidance rationally leads to search a solution in the framework of mean-field theory.
- The other lattices used by Shakhnovitch & coll, Skolnick & coll, mainly $3 \times 3 \times 3$ where all configurations are known & can be enumerated.

Ensemble of minimum criteria to be fulfilled by realistic protein main chains.

(1) Self avoiding

(2) Correct stereochemistry of peptide bonds : use (φ, ψ) angles. Fixed bond lengths & bond angles.

(3) Correct distribution of (φ, ψ) angles in Rama-chandran plot. Draw randomly (φ, ψ) pair with a bias given by a priori information.



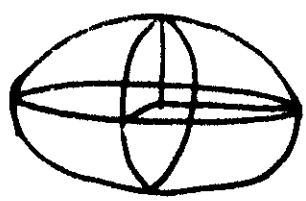
(4) overall shape : Remain in a given ellipsoïd

≠ volume = 168.x N^{res} (± 10%)

axial ratios $e_1 = 0.61 \pm 0.11$

$e_2 = 0.78 \pm 0.12$

(Rackovsky, Hao and Scheraga, 1993, PNAS)



YAM: Use the recursive method of Cohen & Sternberg

To build plausible protein models.

⓪ Randomly place residue #1 inside allowed volume

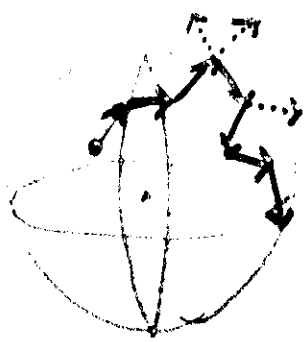
① Grow the main chain as long as all defined criteria are ~~violated~~ fulfilled - Try again and again until it works -

② Go back 3 or 5 residues upstream if no successful move has been accepted in the last 1000 trials

- Grow 3 residues at a time.
- Abandon process if more than $\times .10^6$ trials have already been made.
- Introduce one more level of recursivity ...

→ Method is slow but it works:

It can generate new structures that are acceptable overnight on a workstation -



Mean Field Theory : applications to optimization pbs in Mol. Struct. Biology.

(M.D & L. KOEHL).
I.L. Strasbourg.

What is Mean Field Theory and why do we need it.

Mean Field Theory helps to solve optimization pb. very rapidly. It is derived from Stat. Physics and is especially useful when the Energy contains 2-Body interaction terms. (Ising Model)

$$Z = \sum_{\{S_i\}} \exp -\beta H(\{S_i\})$$ where $S_i = \pm 1$.

2^N config if $i \in [1, N]$

and use saddle-point method to evaluate

H is $\Rightarrow$ each spin sees an "average" (MF)

Energy that is determined by a self consistent Equation... (on a set of Hamiltonians).

$$\langle \cdot \rangle_i = H_i\left(\beta E^i_{local}\right) \text{ where } E^i_{local} = \sum_j \omega_{ij} \langle S_j \rangle$$

if (2 body interactions) $H = \sum_i \sum_j \omega_{ij} S_i S_j$; $E^i_{local} = \frac{\partial H}{\partial S_i}$

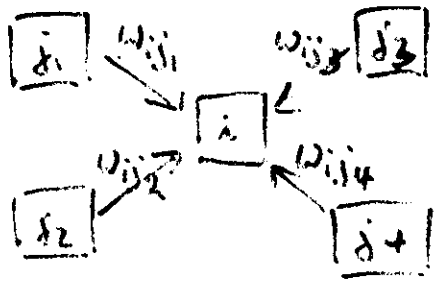
Application of Mean Field theory to predict protein side chain conformations.

- The optimization of side chain conf. cannot be treated by exploring exhaustively all possibilities.
- Side chain conformations are first discretized by building a library of rotamers, which are sufficient to describe most side chains of the PDB. (rmsd = 0.25 Å).
- Use mean field theory to explore all possibilities in only one passage.

  All possible rotamers are attached at each position, with a given weight that is updated from one cycle to the other.

  Each copy sees the average of the neighbouring environment (weighted by current weight).
- Only Van der Waals interactions are taken into account.
- update until CV is achieved.

- This method is also well known in the field of Neural Networks, where the response of each Neuron is a function of the input of all the interacting Neurons



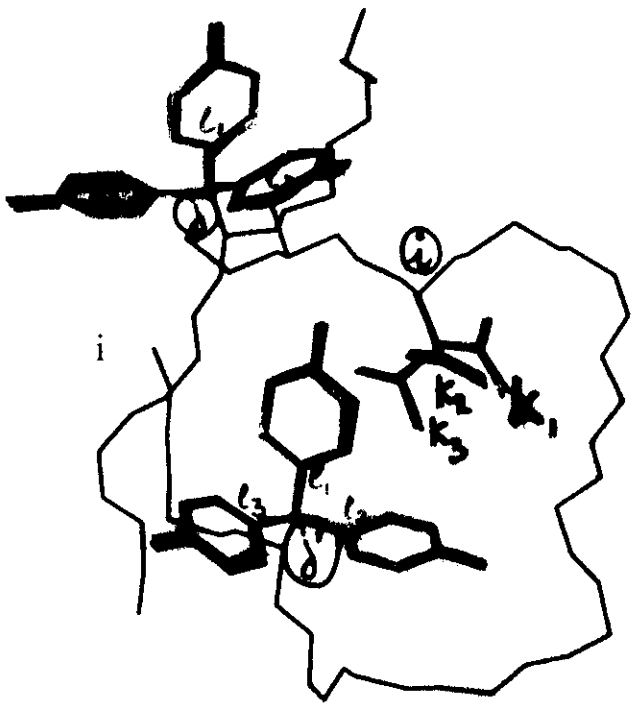
- In the core of a protein, each side chain "talks" to its neighbours and is influenced by them → optimization pb of combinatorial nature

We wish to replace the difficult problem of finding the global minimum (enumerate and evaluate each configuration in turn and exhaustively because the Energy landscape is rugged) by another problem, where the variables are the weights of the different possible copies at each position. (Make the pb discrete) -

It turns out that the Energy landscape is then much easier to explore.

Make the search LINEAR in term of the length of the Seq. N Seq. (Much more powerful than SA).
⇒ Very rapid and efficient search –

C) MEAN FIELD APPROXIMATION



Idea : generate multiple copies of the side–chains, which do not see each other, but see all possible copies of their neighbours, affected by their respective probability : CM(j,l) –

Mean energy of rotamer k for residue i :

position ↓ ↓ possible rotamer

$$E(i,k) = U(i,k) + U(i,k,B) + \sum_{j \in N_i} \sum_{l=1,20} CM(j,l)\, U(i,k,j,l)$$

Intra | Side–chain – Backbone | Side–chain – Side–chain

- How to derive the local field in the general case:

Minimize $F = U - TS$ (Finkelstein & Reva).

$$U = \sum_{\text{all interact. pairs}} \frac{1}{2} \varepsilon_{\eta\xi} f(r_i - r_j)\, p(\eta_i, \xi_j)$$

$$= \sum_i \sum_{j \in N_i} \sum_{\xi,\eta} \frac{1}{2} \varepsilon_{\eta,\xi} f(r_i - r_j) \underbrace{p(\eta_i)\, p(\xi_j)}_{\text{MF hypothesis}}$$

$$-TS = +k_B T \sum_i \sum_\eta p(\eta_i) \log p(\eta_i)$$

$$\delta F = \delta U - \delta(TS)$$

$$= \sum_i \sum_\eta \delta p(\eta_i) \left[k_B T \log p(\eta_i) + \sum_{j \in N_i} \sum_\xi \varepsilon_{\eta\xi} f(r_i - r_j)\, p(\xi_j) \right] + \text{cst}$$

$$= 0$$

$$\Rightarrow \boxed{p(\eta_i) = \exp -\frac{1}{k_B T} \sum_{j \in N_i} \sum_\xi f(r_i - r_j)\, \varepsilon_{\eta\xi}\, p(\xi_j)}$$

- Description of the algorithm itself

Start from equiprobable $p(\eta_i)$ and update until convergence is achieved. Some trick is needed.

$$\frac{\delta p}{\delta t} = k \left(p^{calc} - p^{old} \right) = \frac{p^{new} - p^{old}}{\delta t} \Rightarrow p^{new} = k \delta t\, p^{calc} + (1 - k\delta t) p^{old}$$

$$\Rightarrow \boxed{p^{new} = \lambda\, p^{calc} + (1 - \lambda)\, p^{old}} \quad \text{where } \lambda = k\delta t, \quad \lambda = 0.5$$

- In our implementation of the method, we never physically construct a stereochemically and ~~en~~ "make chain" packing correct chain. It is only at the end, when the algorithm has converged that we decide that, for each position, the rotamer to be constructed is the one that has the highest probability $p(n_i)$ → Energy minimization -

- In C. Lee implementation & M. Vasquez, actual physical realizations of a configuration $\{S_{ij}\}$ are constructed, using the so far derived probability distributions, to calculate E^i_{local}

E^i_{local} = ⟵ $E^i_{physical}$ ⟶ $p^{new}(n_i)$ by Boltzmann like formulae.

$\alpha \simeq 20$ ± Configuration picked up using $p^{old}(n_i)$

The system is always "physical" -

Method called "Ensemble optimization".

It is always better to conduct the optimization process with a POPULATION of CHAINS instead of just one chain.

SIDE-CHAIN CONFORMATION

1) ROTAMER LIBRARY : Tuffery et al (1991)

– a total of 113 rotamers for 19 residues

2) THE CONFORMATIONAL MATRIX: CM

For a protein of N residues : CM of size NRES x K

CM(i,j) is the probability that residue i adopts the conformation described by rotamer j

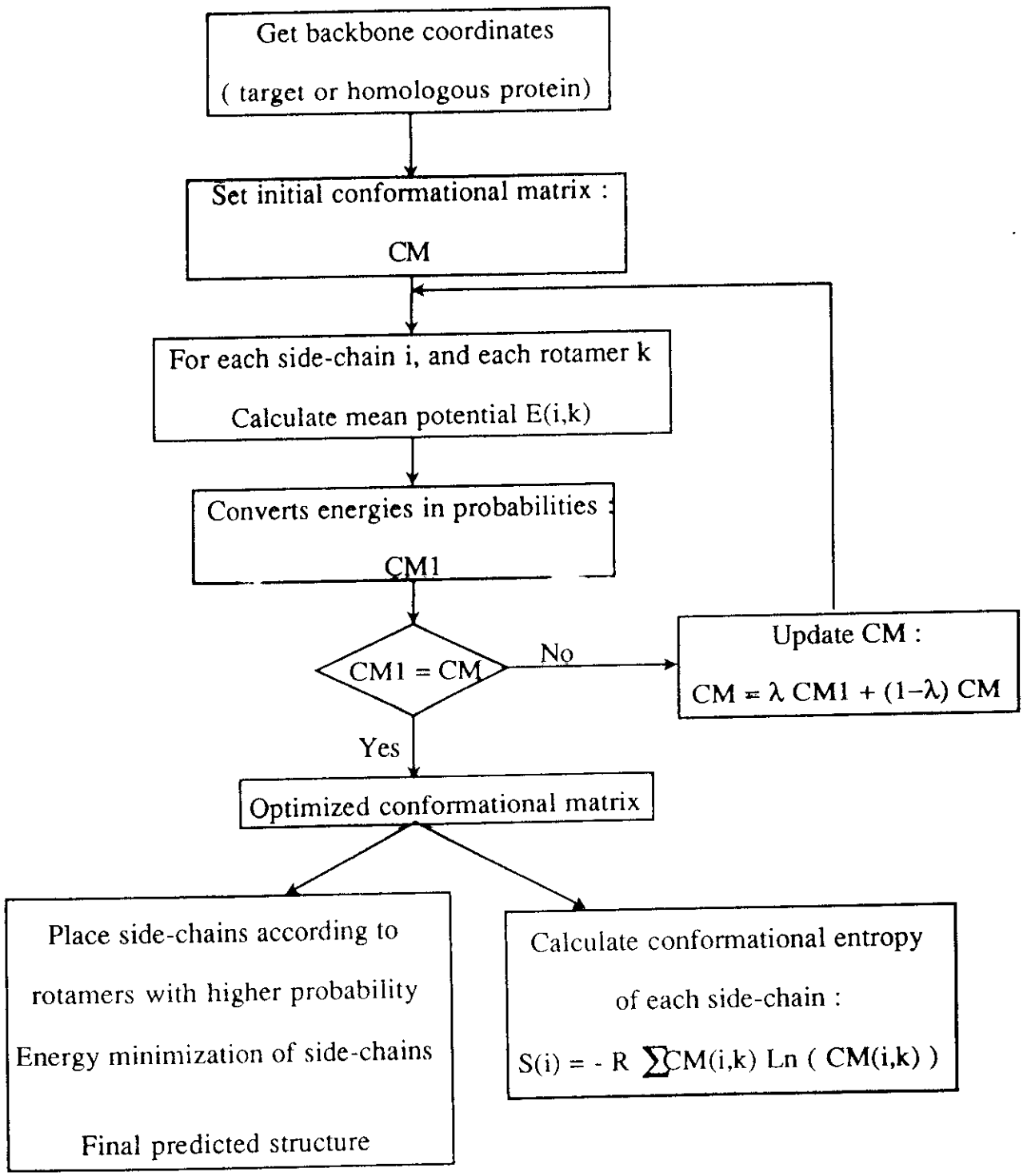
Two applications :

A) The rotamer with the highest probability describes the conformation of the side-chain

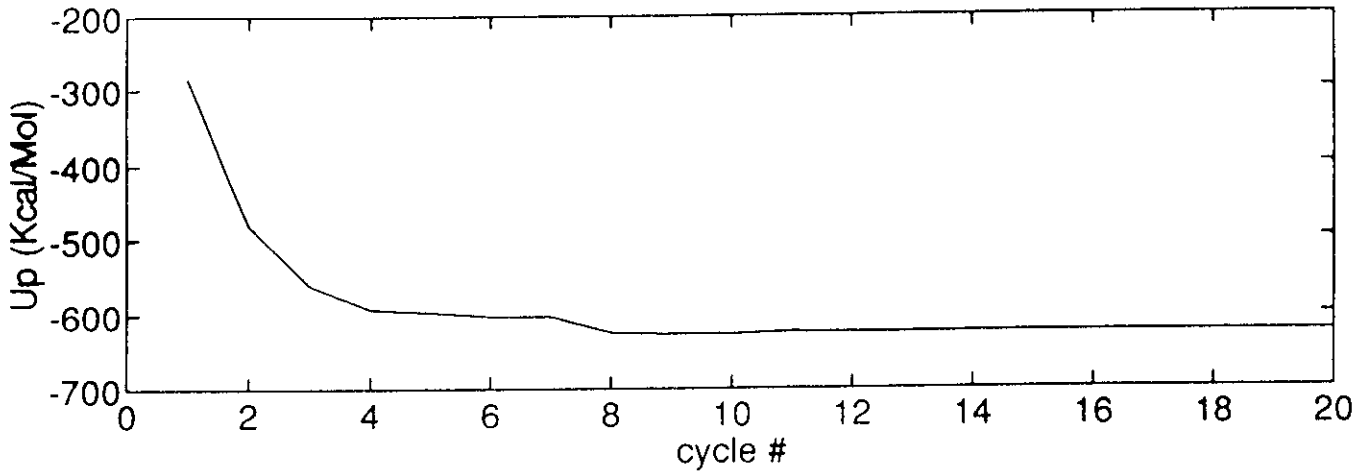
B) CM provides an estimate of the conformational entropy of the side-chains

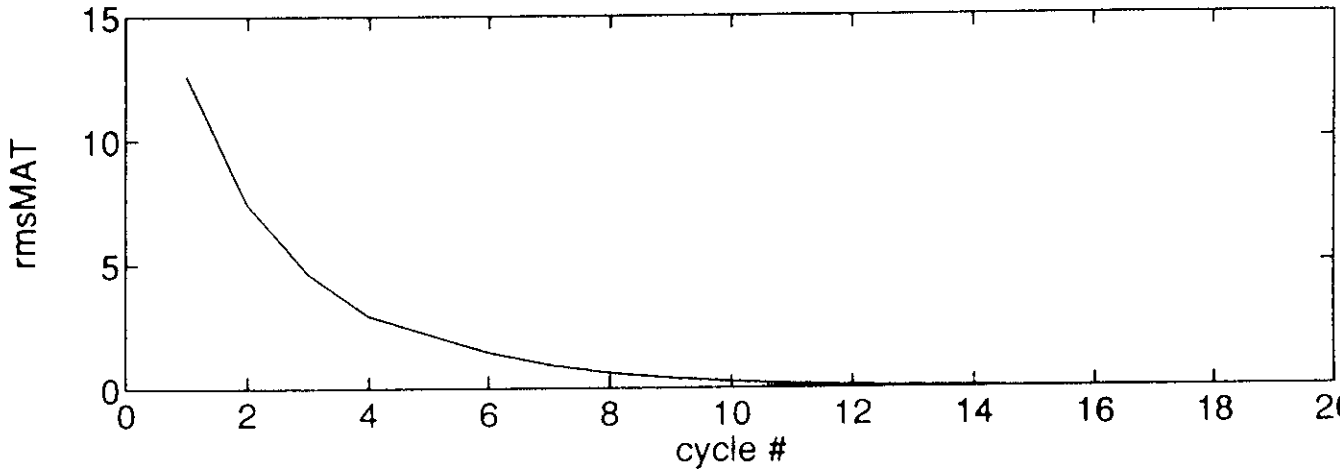
For a residue i $\quad S(i) = - R \sum_{k} CM(i,k)\ Ln\ [\ CM(i,k)\ ]$

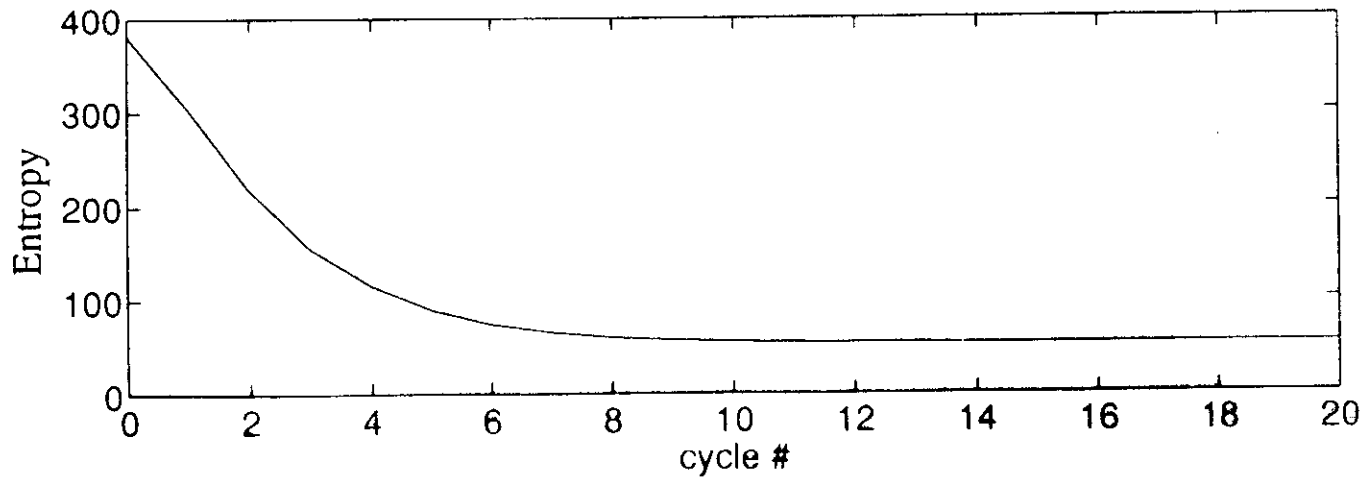
CALCULATION OF THE SIDE-CHAIN CONFORMATION MATRIX



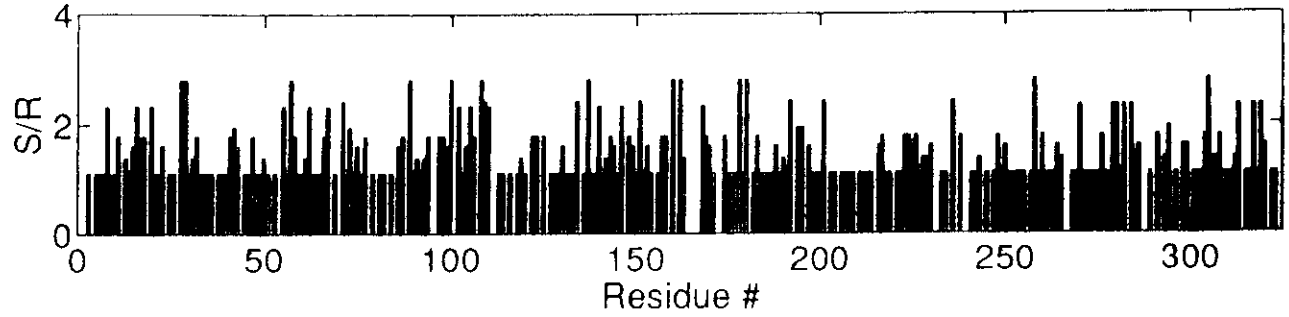
SIDECHAIN POSITIONING FOR RHIZOPUSPEPSIN



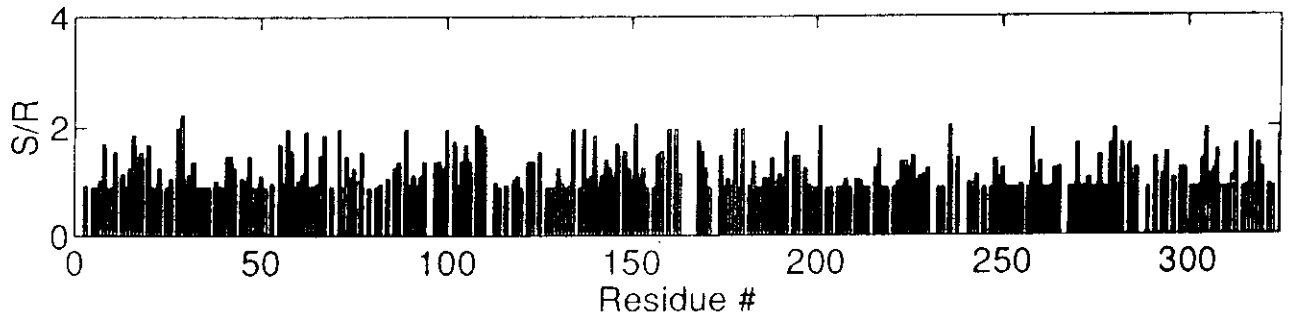




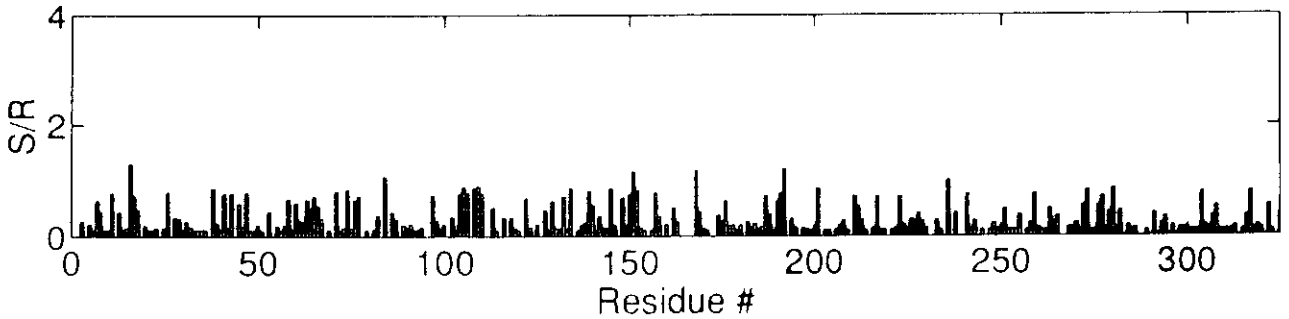
Initial entropy



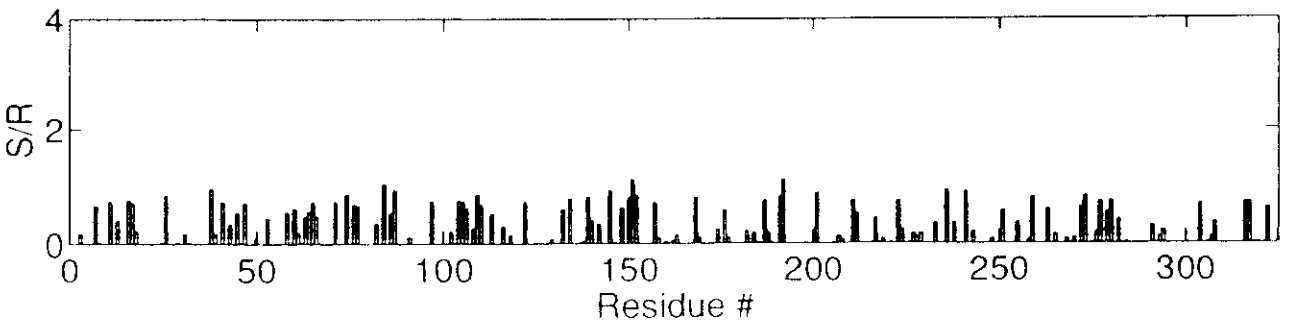
After 1st cycle

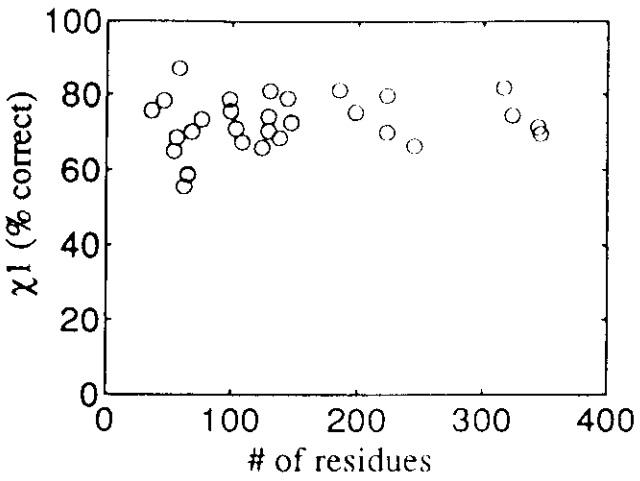


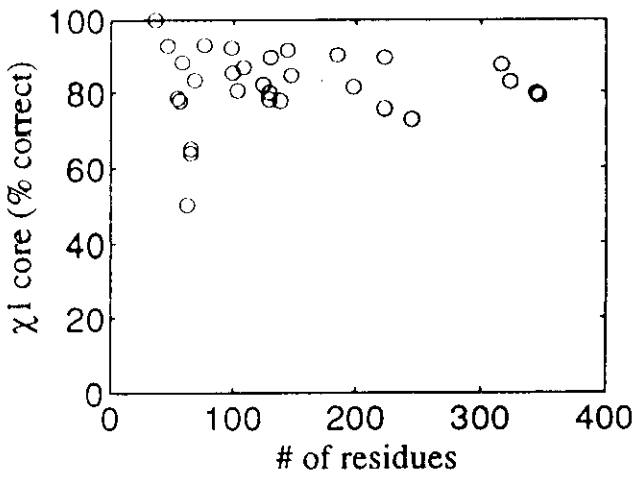
After 5 cycles

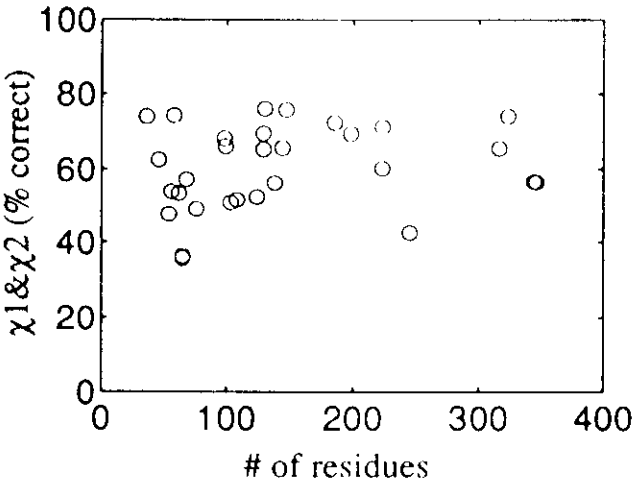


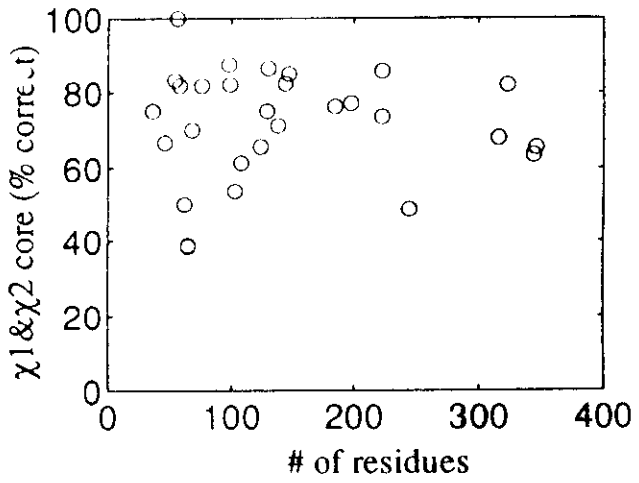
Final

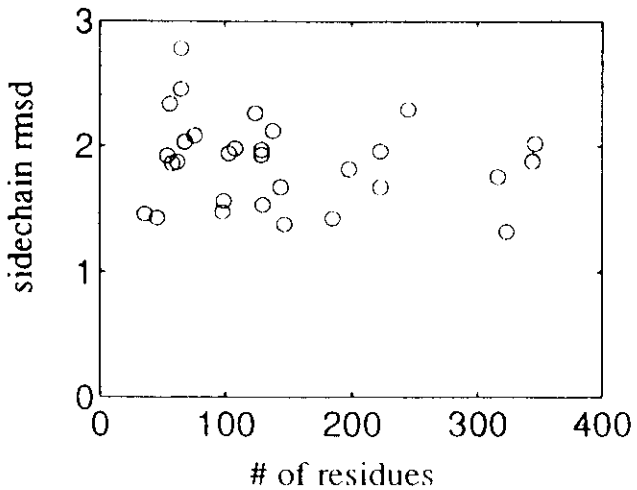


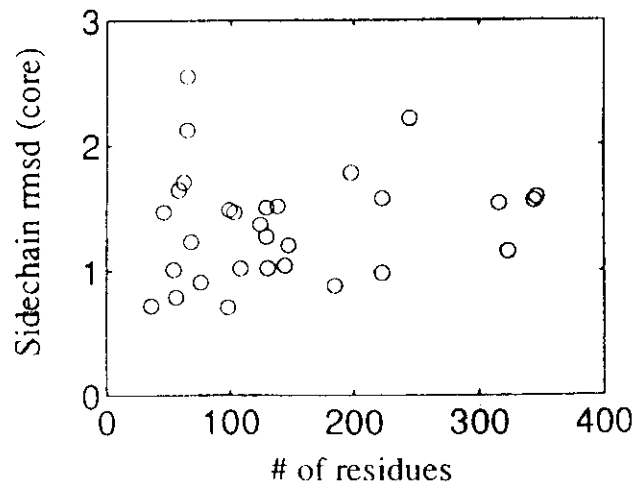

χ1 (% correct)
# of residues


χ1 core (% correct)
# of residues

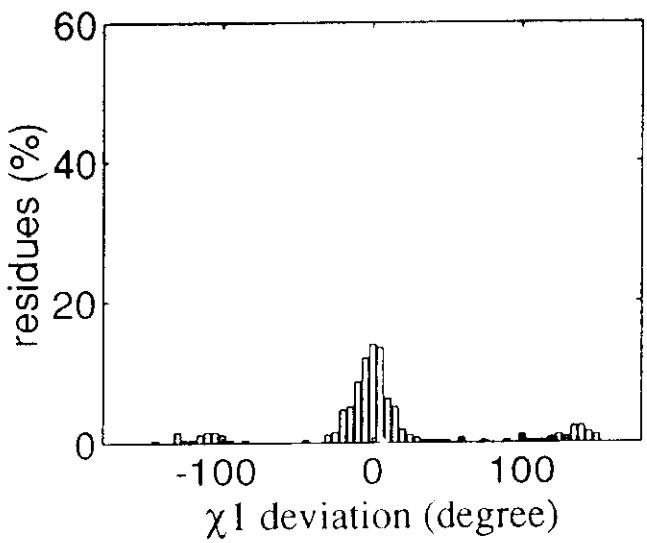

χ1&χ2 (% correct)
# of residues

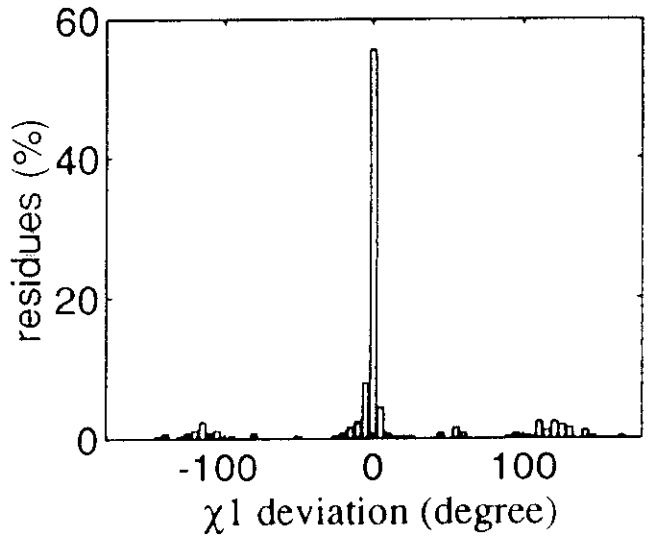

χ1&χ2 core (% correct)
# of residues

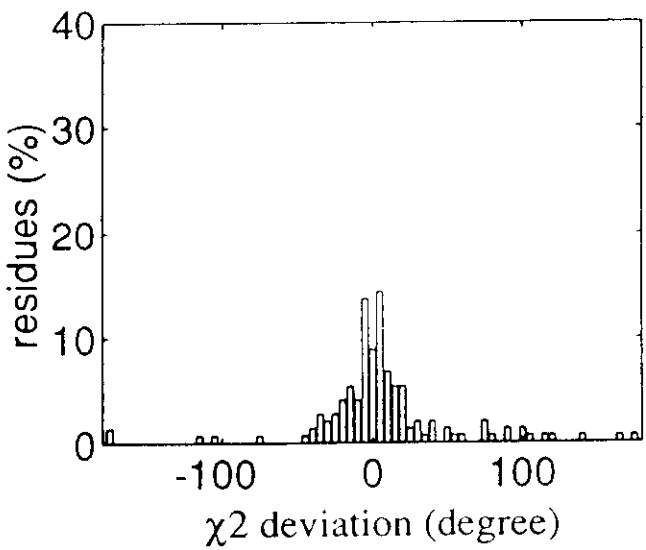

sidechain rmsd
# of residues

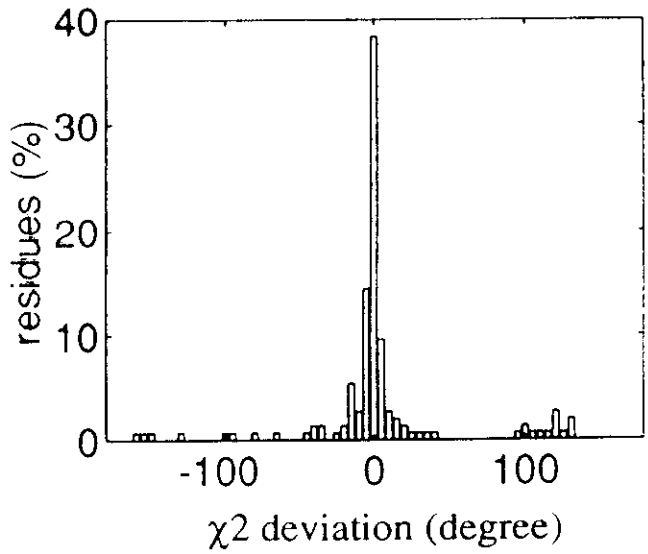

Sidechain rmsd (core)
# of residues

Effects of the energy minimization on the accuracy of the dihedral angles prediction









Accuracy of side-chain prediction for specific side-chain types

Residue	χ1 (%)	χ2 (%)	χ1+2 (%)	Rmsd (Å)
Large non polar				
VAL	84			0.92
ILE	82	77	68	1.22
LEU	85	72	68	1.24
PHE	88	94	85	1.61
MET	84	73	62	1.81
Aromatic side-chains				
TRP	87	72	69	2.30
TYR	93	97	92	1.45
HIS	81	51	49	1.95
Small side-chains				
PRO	55			0.84
CYS	81			1.18
SER	42			1.69
THR	80			0.97
Polar side-chains				
ASN	73	57	46	1.84
GLN	73	74	60	2.02
ASP	64	81	56	1.83
GLU	66	66	45	2.15
LYS	68	72	50	2.44
ARG	66	81	54	3.05

Possible causes for discrepancies

A) Incomplete force-field

1) No specific terms for hydrogen bonds (SER, THR ...)

2) Solvent not included (surface residues)

B) Inherent clashes due to the rotamer library

Test case : BPTI; energy : - 167 Kcal/Mol

"Ideal" structure : each sidechain replaced by the closest (r.m.s.d.) rotamer

		U (Kcal/Mol)			
Library	No. of rotamers	Ideal structure	Final model	$\chi 1$ (%)	$\chi 1+2$ (%)
Tuffery et al (1991)	303	81	-39.4	87	74
LIB1 (120°)	552	61	-56.7	85	67
LIB2 (60°)	3474	-75	-79.1	67	57
LIB3 (30°)	4512	-84	-110	65	54
BPTI			-167		

How do we know that a given sequence is compatible with a given 3D structure?

→ obvious answer if there is seq. homology.

→ "Phenomenology of sequence evolution"-

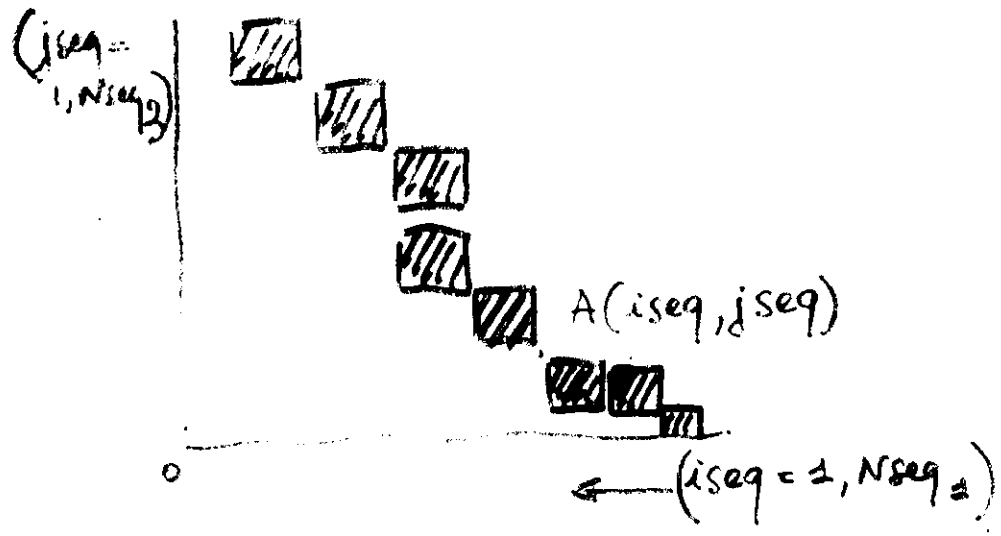
- Definition of conservative mutations
- The only strictly conserved residues for a family of proteins that has widely diverged are functional (only a handful of them ⇒ very difficult to detect)-
- Evolutionary information can and should be used to refine mutation matrix-

→ Molecular Mechanisms at stake

- point mutations - constant drift?
- Gene duplication & Gene fusion -

Aligning two sequences ...

① The algorithm: Dynamic programming Method

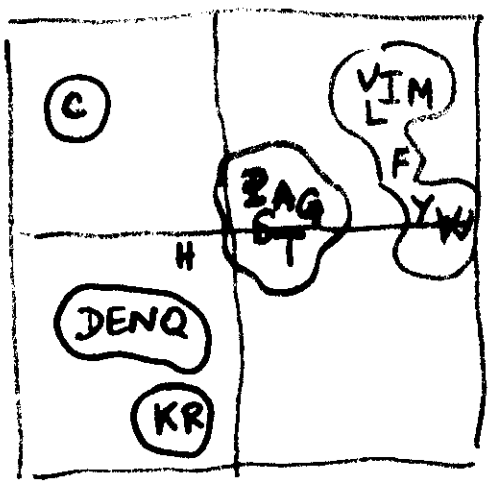


→ Table of amino acid substitution + gap penalties.

$M(iaa\ jaa)$: $iaa = 1, 20$
$jaa = 1, 20$

Purely empirical (M. Dayhoff, 70's)

Principal component analysis → projection on a plane



Applications

- Genome sequencing projects
  - Mycoplasma ~ 480 genes
  - Helicobacter pyloni
  - E. coli ~ 3000 genes (50%) - S. thyphimirium - B. subtilis
  - yeast 6 000 genes. 12 M bases.
  - Rice, C. elegans, A. thaliana, D. melanogast
  - Human genome . $3 \cdot 10^9$ bp ~ 10^5 genes.

- Automatic comparison of new seq. to existing Seq D (~ 60 000 Seq.)
  - → 1/3 with clear homolog of known function
  - → 1/3 with homolog of unknown function. (or improved)
  - → 1/3 with no homolog in Seq. Data Base.

Going beyond obvious Seq. homology (≥ 30%)

(1) Evolutionary information. M. Gribskov.
Define a position dependent Matrix $M(i,j)$
from previous multialignement - Profile Method,
confined to blocks of aligned sequences -

(2) Make use of structural information. J. Bowie

(a) define to which class each a.a. along the se
belongs, according to criteria calc. from
structure - solvent accessibility
- secondary structure

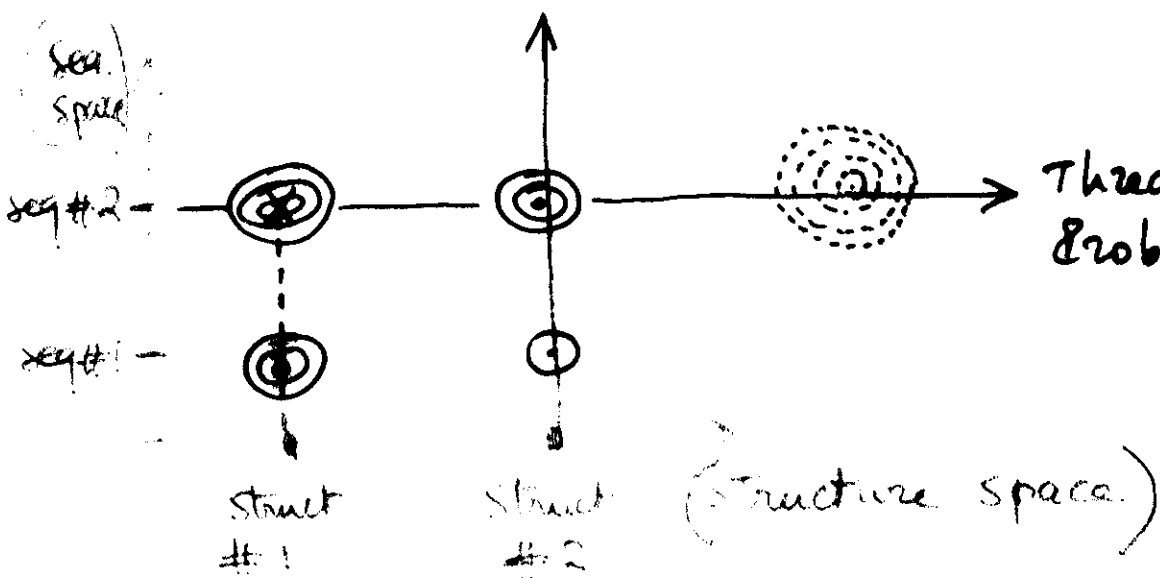
(b) convert belonging to one structural class
into mutability to other a.a. types

Critique of this approach -

- solvent accessibility really conserved amo
structurally aligned positions - M. Sternberg.
- Atomic properties - But Backbone can adjust

Sequence Design

A. Godzik.



- The threading problem requires:
  - good potentials
  - a method to align the gaps, and to get rid of the "frozen approximation"

- The sequence design problem requires:
  - good potentials
  - a method to avoid "overdesigning" and to select for other folds.

Answer: Maximize $\dfrac{E - \langle E \rangle_{\text{all other folds}}}{\sigma_{\text{all other folds}}}$

The potential Energy Functions

- Residue - Residue contact potentials - (statistical pot.)

$$E = \sum_{i,j} \varepsilon(r_i, r_j)\, f(\vec{r}_i - \vec{r}_j)$$

- Myiazawa & Jernigan, 1985 ; Sippl (dist. dep), 1990
  Bryant & Lawrence 1991 ...

- All of them can distinguish the right structure among an ensemble of wrong structures.

- They do not have the same reference state.
  from E_{ij}, define → $E_{ij}^{ideal} = \frac{E_{ii} + E_{jj}}{2}$ (Burial).
  [Godzik 1995] → $E_{ij}^{excess} = E_{ij} - E_{ij}^{ideal}$ (2-Body).

- They are not good for ~~[illegible]~~ Seq. design experiments (MC).
  → overdesigning.

- There are not truly speaking two body potentials that can be added directly. Need of 3 body-pot.

→ Control experiment with 3 x 3 x 3 lattices, where all compact SAW are know : 103, 346 of them...

Use Mean Field Theory to get $M(i,j)$

$$E = \sum_{i,j} \sum_{\eta_i,\eta_j} f(x_i - x_j)\, \varepsilon(\eta_i,\eta_j)\, p(\eta_i,\eta_j)$$

$$\langle E \rangle_{s.c.} = \sum_{i,j} \sum_{\eta_i,\eta_j} \langle f(x_i - x_j) \rangle_{c.s.} \varepsilon(\eta_i,\eta_j)\, p(\eta_i,\eta_j)$$

$T = \ldots$

→ Get $p(\eta_i)$ in a self-consistent way.

→ Independent of starting point (removes the "frozen approx.").

→ Converges in a few cycles.

→ maintains full atomic repl. of a.a. if desired.

Caution ① Adjust T

② Add Extra Term to maintain amino acid composition $\sum_{\eta=1}^{20} \left[\left(\sum_i p(\eta_i) \right) - N_{seq} \cdot \bar{c}(\eta) \right]^2$

③ Correct for rotamer unequal representation

④ optimize $\varepsilon(\eta_i,\eta_j)$ so that native fold is best recognized —

$$\text{Score} = -\sum_i \ln\left(\eta_i^{\text{wild type}}\right) -$$