

# Comparative Protein Modeling by Satisfaction of Spatial Restrictions

Sali A. and Blundell T, Journal

**Wonsik Joung, Roxanne Martinez**

# Outline

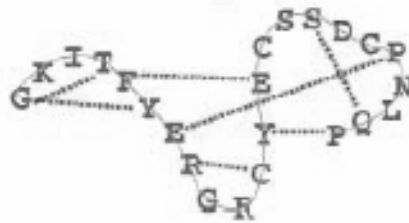
- **Introduction**
- **Derivation of Spatial Restraints**
- **Restraining Spatial Features**

# Introduction

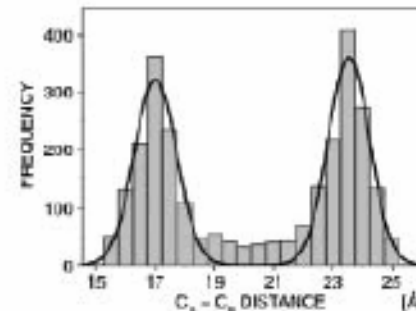
- **What's comparative protein modeling?**

3D GKITFYERGFQGHCYESDC-NLQP...  
SEQ GKITFYERG---RCYESDCPNLQP...

1. Extract spatial restraints



2. Satisfy spatial restraints



$$F(\mathbf{R}) = \prod_i p_i(f_i / l_i)$$

# Introduction

- **What is the most probable structure for a certain sequence given its alignment with related structures?**

# Introduction

- **The method stages:**
  - **Alignment of the sequence (not covered this paper)**
  - **Extraction of spatial restraints**
  - **Satisfaction of the restraints**

# Derivation of Spatial Restraints

- **Conditional pdf**
  - $P(x / a, b, \dots, c)$
  - Probability density for  $x$  when  $a, b, \dots, c$  are specified
- **For example**
  - $P(x_1 / \text{residue type}, \Phi, \Psi)$
  - To predict the side-chain dihedral angle  $x_1$  from the type of a residue and its main-chain angles  $\Phi$  and  $\Psi$

# Derivation of Spatial Restraints

- **Pdf approximation**

- $P(x / a, b, \dots, c) \approx W_{x, a, b, \dots, c}$

- $W_{x, a, b, \dots, c}$  is a table spanned by  $x, a, b, \dots, c$  that contains as its elements the observed relative frequencies for the occurrence of  $x$  given  $a, b, \dots, c$

# Derivation of Spatial Restraints

- **Pdf approximation (continue)**

- $W_{x,a,b,\dots,c} = W'_{x,a,b,\dots,c} / \sum_x W'_{x,a,b,\dots,c}$
- $W$  is calculated from the absolute frequencies  $W'$
- $W'$  are obtained directly by counting the number of occurrences of each combination of  $(x,a,b,\dots,c)$  values

# Derivation of Spatial Restraints

- **Local database**
  - **17 families, 80 protein structures**
  - **Grouped by homologous structures**
  - **Representative of all structural classes of proteins**

# Derivation of Spatial Restraints

- **Tabulating associations between protein features**
  - **The program MDT explores the local database and to derive the best pdfs**

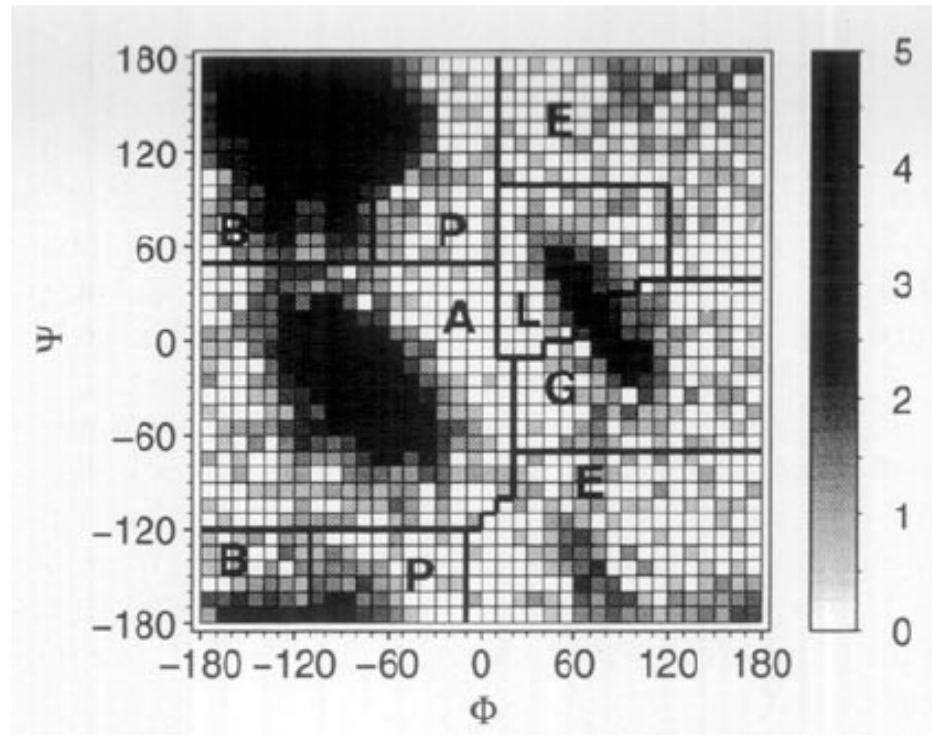
# Derivation of Spatial Restraints

## •Features used in this paper

Index	Variable	Feature
1	$r$	Amino acid residue type
2	$\Phi$	Main-chain dihedral angle $\Phi$
3	$\Psi$	Main-chain dihedral angle $\Psi$
4	$t$	Secondary structure class of a residue
5	$M$	Main-chain conformation class of a residue
6	$\alpha$	Fractional content of residues in the main-chain conformation class A
7	$\chi_i$	Side-chain dihedral angle $\chi_i$ , $i = 1, 2, 3, 4$
8	$c_i$	Side-chain dihedral angle $\chi_i$ class, $i = 1, 2, 3, 4$
9	$a$	Residue solvent accessibility
10	$\bar{a}$	Average accessibility of two residues in one protein
11	$s$	Residue neighbourhood difference between two proteins
12	$\bar{s}$	Average residue neighbourhood difference between two proteins
13	$i$	Fractional sequence identity between two proteins
14	$d$	C <sup><math>\alpha</math></sup> -C <sup><math>\alpha</math></sup> distance
15	$\Delta d$	Difference between two C <sup><math>\alpha</math></sup> -C <sup><math>\alpha</math></sup> distances in two proteins
16	$h$	Main-chain N-O distance
17	$\Delta h$	Difference between two main-chain N-O distances in two proteins
18	$b$	Average residue $B_{300}$
19	$R$	Resolution of X-ray analysis
20	$g$	Distance of a residue from a gap in alignment
21	$\bar{g}$	Average distance of a residue from a gap

# Derivation of Spatial Restraints

- **Main-chain conformation class of a residue (M)**



- **The plot shows  $W'(\Phi, \Psi)$  determined from the local database**

# pdf from a sparse data set

- **Calculation of probabilities  $W$  from frequencies  $W'$  is strictly valid only when the frequencies are very large**

$$p(x_i) = \omega_1 A(x_i) + \omega_2 W(x_i)$$

$$\omega_1 = \frac{1}{1 + N/(\sigma n_x)} \quad \text{and} \quad \omega_2 = 1 - \omega_1,$$

- **$N$  is the number of points in  $W'$**
- **$N_x$  is the number of bins i.e  $i=1,2,\dots,N_x$**
- **$\sigma$  is a parameter that determines the relative contributions of the a priori distribution**
- **$A$  is a priori distribution**

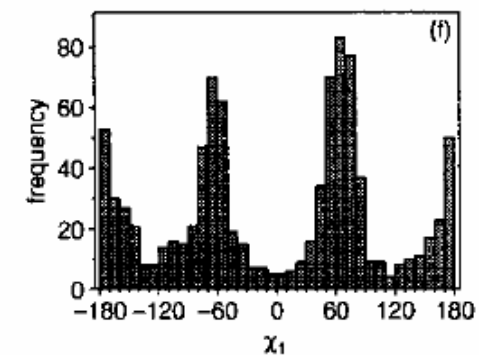
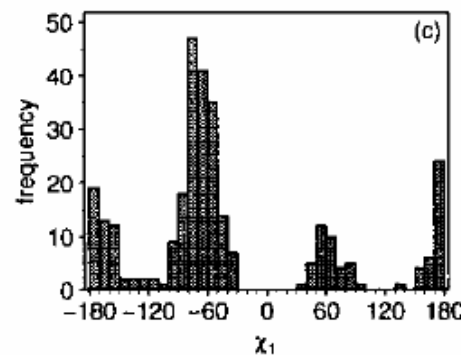
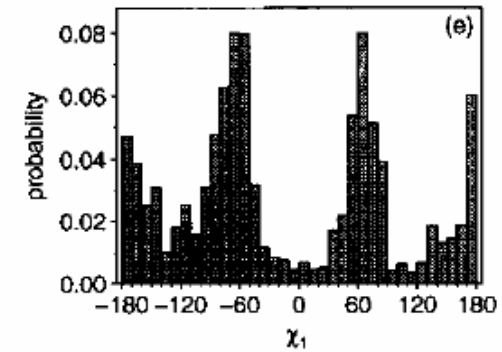
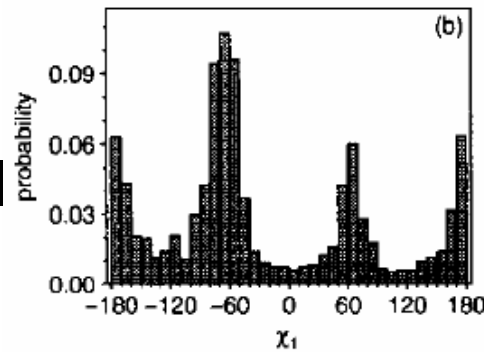
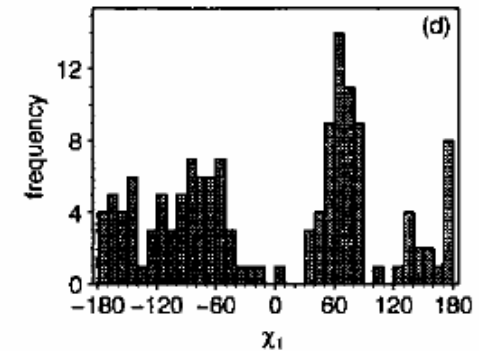
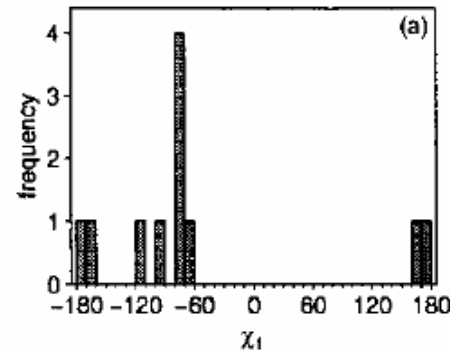
# pdf from a sparse data set

- X1 side-chain dihedral angles

- (a) from a small database of 8 proteins

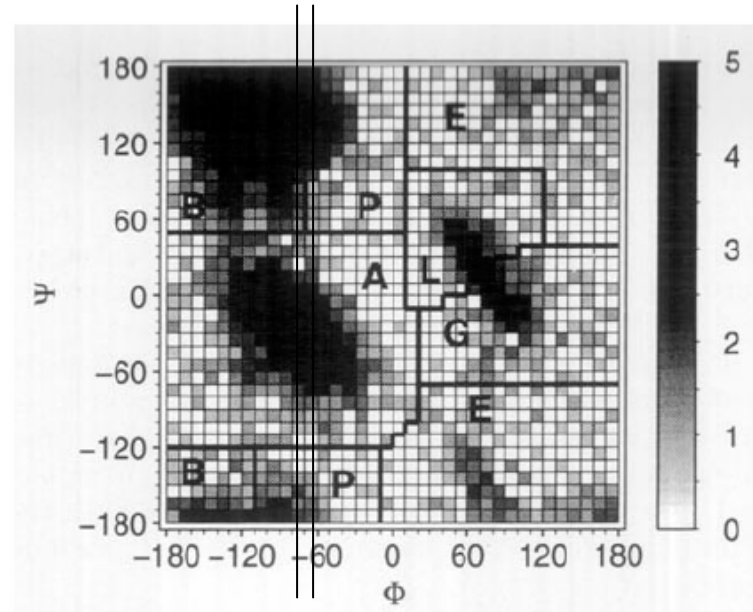
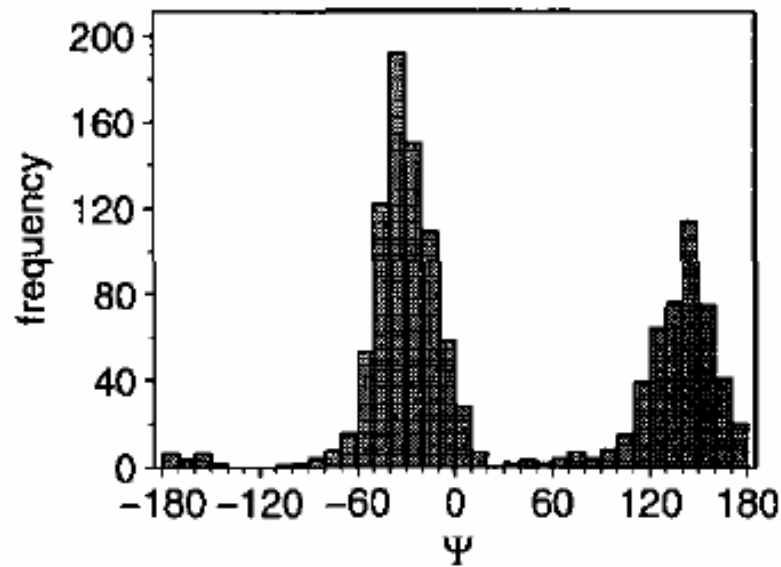
- (b) smoothing sparse distribution with parameter  $\sigma = 5$

- (c) from the whole local database of 80 protein structures without smoothing distribution



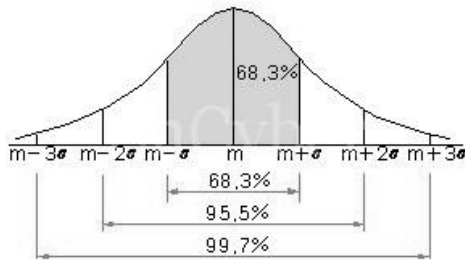
# Restraining residue main-chain conformation

- **Distribution of the  $\Psi$  main-chain dihedral angle at  $\Phi = [-80^\circ, -70^\circ]$  from the table  $W'(\Phi, \Psi)$**



# Restraining residue main-chain conformation

- The **distribution** of the main-chain dihedral angles approximates the **Gaussian distribution**



$$f(x) = \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\} \sqrt{2\pi\sigma^2}$$

	Mean ( $^\circ$ )		Standard deviation ( $^\circ$ )	
	$\Phi_i$	$\Psi_i$	$\sigma_i(\Phi)$	$\sigma_i(\Psi)$
A	-65	-41	15	15
B	-130	135	15	20
P	-65	140	15	15
G	60	40	10	10
L	90	-10	15	10
E	130	180	25	25

# Restraining residue main-chain conformation

$$p^m(\Phi) = \sum_{i=A, \dots, E} \omega_i N[\bar{\Phi}_i, \sigma_i(\Phi)]$$
$$p^m(\Psi) = \sum_{i=A, \dots, E} \omega_i N[\bar{\Psi}_i, \sigma_i(\Psi)],$$

- $N[\alpha, \sigma]$  stands for a Gaussian pdf with mean  $\alpha$  and standard deviation  $\sigma$
- $\omega$  is the probability that the restrained residue is in the main-chain conformation class  $i$