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 - 21

WORKSHOP ON:
PROTEINS, MEMBRANES and their INTERACTIONS

22 JULY - 2 AUGUST 1996

"Prediction of protein structure"

Jean GARNIER
Laboratoire de Biologie Cellulaire et Moleculaire
Biotechnologies, INRA
78352 Jouy-en-Josas
FRANCE

Prediction of protein structure

I Introduction

- basic building blocks of living matter: self-organizing hetero-polymers, source of specific interactions
- Degeneracy of the code relating amino acid residues and 3D structure.

II Prediction of secondary structures[1]

- Statistical methods: Information Theory (GOR)
- Artificial Intelligence methods: _ Neural network
  _ Nearest-neighbor
- Evolutionary Information

III Prediction of the 3-D structure

- Comparative modeling[2]
  - Amino Acid sequence alignments
  - Loop modeling: database search
    Canonic rules
    ab initio (ex MC)
- Fold recognition techniques[3]
  - Profiles, pair potential
  - threading (NP complete)

1 Review in Garnier + Levin (1991) Comput. Applic. Biosci., 7, 133-142

2 " " Greer (1991) Meth. in Enzymology, 202, 239-252 + Mosimann et al. Proteins, 13 (1992) 301-317

3 " " Lemer et al. (1995), 23, 337-355

Slides 5

SPECIFICITY OF PROTEIN STRUCTURE

Flexibility around torsional bonds

Large molecules

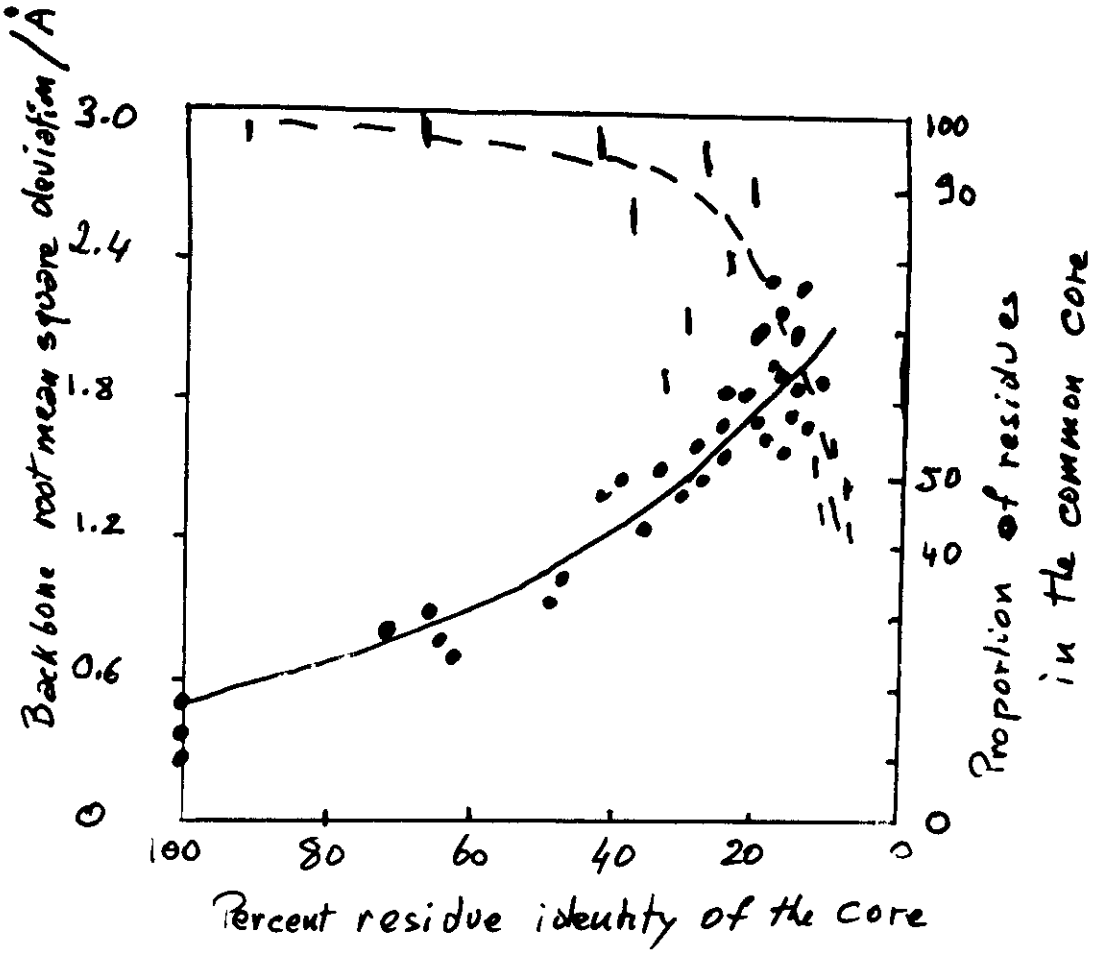
Dihydrofolate reductase :
2492 atoms--159 amino acids

Large possible sequence diversity

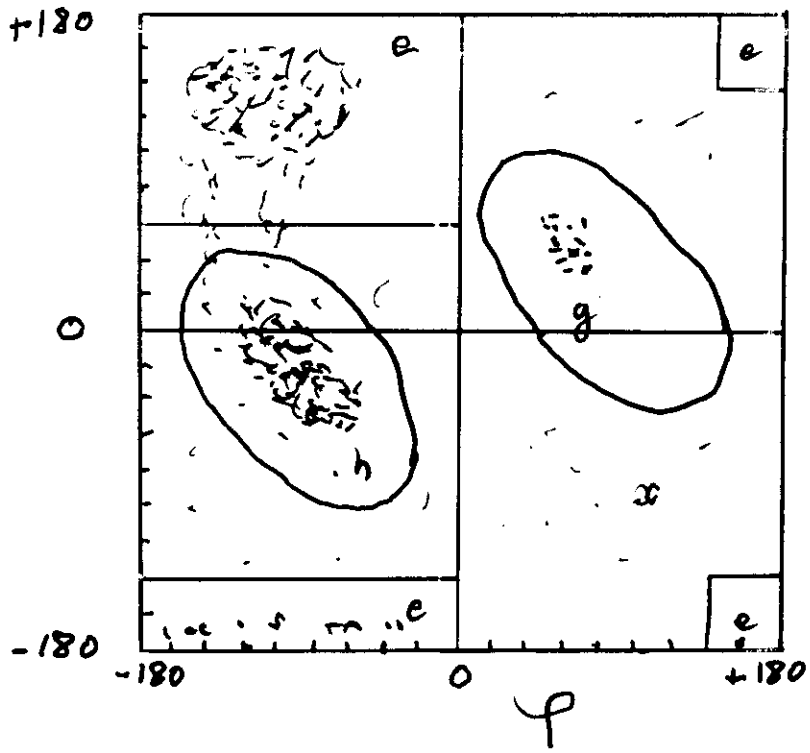
Peptide of 20 amino acids :
20^{20} possible sequences
OR 1.04 10^{26} combinations

Degeneracy of the code relating structure and amino acids


A:124
101
CA
CA_5



(Lesk + Chothia)



Ramachandran map of observed dihedral angles

~ 4561 residues except G

DISTRIBUTION OF SOME AMINO ACID RESIDUES ACCORDING TO THEIR CONFORMATION

Amino Acid	Total	Ramachandran %		map
		h	e	g+x
G	981	18.8	47.7	33.5
A	926	58.7	36.1	5.2
V	802	39.8	57.5	2.8
P	508	43.3	54.5	2.2
ALL	11051	45.3	46.5	8.1

DISTRIBUTION OF SOME AMINO ACID RESIDUES IN APERIODIC (COIL) CONFORMATION

Amino Acid	C/Total %	Ramachandran map %			
		h	e	g	x
G	72.1	10.2	44.4	43.0	2.4
A	40.5	37.8	50.9	4.0	7.2
V	32.8	26.6	65.8	2.7	4.9
P	74.2	35.5	61.8	0.3	2.4
ALL (C)	49.7	31.6	52.7	9.3	6.3
ALL (H+E+C)		45.3	46.5	4.8	3.3

Table I: Distribution of the residues in the Ramachandran zones according to their secondary structure *

Secondary structure: H

	h	e	g	x
G	112	7	2	0
A	401	1	0	0
V	237	0	0	1
L	332	1	0	0
I	172	1	0	0
S	165	0	0	0
T	164	0	1	0
D	174	0	2	2
E	219	1	0	0
N	112	1	2	1
Q	142	1	0	0
K	238	0	1	0
H	81	0	0	0
R	109	0	0	0
F	151	0	0	0
Y	106	1	1	0
W	61	0	0	0
C	63	0	0	0
M	68	0	0	1
P	86	0	0	0

Secondary structure: E

	h	e	g	x
G	0	147	5	1
A	2	145	1	0
V	12	200	0	1
L	8	199	0	1
I	5	177	0	1
S	6	149	0	2
T	0	188	0	1
D	4	47	1	1
E	5	74	0	0
N	3	62	3	0
Q	6	77	0	0
K	1	108	0	1
H	2	49	0	0
R	0	87	0	0
F	4	114	0	1
Y	1	126	0	0
W	0	51	0	0
C	0	92	0	0
M	0	51	0	0
P	0	44	0	1

Table I: (continued)

Secondary structure: C

	h	e	g	x
G	72	314	304	17
A	141	190	15	27
V	70	173	7	13
L	90	169	6	18
I	47	109	1	13
S	219	248	18	40
T	132	196	4	18
D	149	198	16	33
E	95	120	7	19
N	112	163	61	31
Q	68	105	8	12
K	129	160	19	23
H	38	76	13	19
R	69	98	6	9
F	42	91	7	10
Y	55	102	12	15
W	26	30	0	4
C	43	93	6	14
M	16	41	5	4
P	134	233	1	9

Secondary structure: H+E+C

	h	e	g	x
G	184	468	311	18
A	544	334	16	32
V	319	461	7	15
L	430	369	6	19
I	224	287	1	14
S	390	397	18	42
T	304	378	5	19
D	327	246	19	36
E	319	195	7	19
N	227	226	66	32
Q	216	183	8	12
K	368	268	20	24
H	121	125	13	19
R	178	185	6	9
F	197	205	7	11
Y	162	229	13	15
W	87	81	0	4
C	106	185	6	14
M	84	92	5	5
P	220	277	1	10

* with Kabsch and Sander algorithm[27]

		h	e	g	x	PRED	OBS
1	R	0.45	0.47	0.05	0.03	Z	Z
2	P	0.36	0.58	0.00	0.02	e	e
3	D	0.48	0.41	0.04	0.06	h	h
4	F	0.35	0.59	0.01	0.01	e	h
5	C	0.30	0.60	0.01	0.07	e	h
6	L	0.22	0.68	0.02	0.02	e	h
7	E	0.10	0.85	0.00	0.05	e	e
8	P	0.10	0.84	0.00	0.03	e	e
9	P	0.31	0.43	0.00	0.20	e	e
10	Y	0.34	0.46	0.04	0.14	e	e
11	T	0.34	0.51	0.04	0.18	e	h
12	G	0.03	0.86	0.02	0.06	e	e
13	P	0.38	0.55	0.00	0.06	e	h
14	C	0.36	0.55	0.01	0.02	e	e
15	K	0.39	0.48	0.11	0.01	e	h
16	A	0.29	0.56	0.33	0.01	e	e
17	R	0.33	0.59	0.03	0.01	e	e
18	I	0.21	0.70	0.00	0.01	e	e
19	I	0.46	0.48	0.00	0.00	e	e
20	R	0.43	0.57	0.00	0.01	e	e
21	Y	0.46	0.49	0.01	0.05	e	e
22	F	0.41	0.48	0.02	0.01	e	e
23	Y	0.40	0.54	0.01	0.02	e	e
24	N	0.50	0.38	0.02	0.06	h	e
25	A	0.52	0.34	0.06	0.00	h	h
26	K	0.56	0.36	0.04	0.00	h	h
27	A	0.65	0.26	0.10	0.00	h	h
28	G	0.20	0.38	0.39	0.02	g	g
29	L	0.39	0.53	0.01	0.02	e	e
30	C	0.33	0.60	0.03	0.01	e	e
31	D	0.41	0.55	0.00	0.01	e	e
32	T	0.27	0.67	0.01	0.04	e	e
33	F	0.31	0.61	0.00	0.04	e	e
34	V	0.22	0.65	0.10	0.03	e	e
35	Y	0.31	0.55	0.02	0.08	e	e
36	G	0.20	0.60	0.05	0.01	e	h
37	G	0.07	0.64	0.27	0.00	e	g
38	C	0.32	0.57	0.03	0.02	e	e
39	R	0.44	0.40	0.75	0.01	g	g
40	A	0.29	0.61	0.06	0.01	e	e
41	K	0.38	0.50	0.02	0.03	e	e
42	R	0.43	0.47	0.01	0.01	e	h
43	N	0.34	0.35	0.10	0.15	e	e
44	N	0.30	0.42	0.14	0.06	e	e
45	F	0.30	0.68	0.00	0.00	e	e
46	K	0.47	0.45	0.01	0.04	h	h
47	S	0.34	0.58	0.00	0.02	e	e
48	A	0.58	0.34	0.00	0.03	h	h
49	E	0.66	0.25	0.04	0.01	h	h
50	D	0.53	0.31	0.04	0.01	h	h
51	C	0.56	0.39	0.02	0.01	h	h
52	M	0.52	0.41	0.04	0.00	h	h
53	R	0.45	0.52	0.00	0.02	e	h
54	T	0.39	0.52	0.02	0.05	e	h
55	C	0.30	0.58	0.05	0.06	e	h
56	G	0.25	0.35	0.45	0.08	g	h
57	G	0.09	0.64	0.26	0.01	e	e
58	A	0.45	0.47	0.05	0.03	Z	Z

Information theory

A. 1 state S, 1 residue R, 20 R, 3 S, S = helix, extended or coil (operiodic)

Fano, Brillouin, Shannon

$$I(S;R) = \log\left(\mathcal{P}(S|R)/\mathcal{P}(S)\right) \quad \text{and} \quad \mathcal{P}(S|R) = \mathcal{P}(S,R)\,\mathcal{P}(R)$$
$$= \log\left(\mathcal{P}(S,R)/\mathcal{P}(R)\,\mathcal{P}(S)\right)$$
$$= \log\left(\left(n_{S,R}/n_R\right)/\left(n_S/N\right)\right)$$

N = total number of Amino Acids

$\left(n_{S,R}/n_R\right)/\left(n_S/N\right)$: Chou + Fasman parameter

B. 2 states, S and $\bar{S}$ (non S), n residues R (typically $n = 17$)

$$I(\Delta S; R_1, \ldots, R_n) = I(S; R_1, \ldots, R_n) - I(\bar{S}; R_1, \ldots, R_n)$$

$$= \log\left(\mathcal{P}(S, R_1, \ldots, R_n)/\mathcal{P}(\bar{S}, R_1, \ldots, R_n)\right) + \log\left(\mathcal{P}(\bar{S})/\mathcal{P}(S)\right)$$

or $$\mathcal{P}(S, R_1, \ldots, R_n)/\mathcal{P}(\bar{S}, R_1, \ldots, R_n) = \mathcal{P}(S)/\mathcal{P}(\bar{S})\, e^{I(\Delta S; R_1, \ldots, R_n)}$$

and $$\mathcal{P}(S, R_1, \ldots, R_n) + \mathcal{P}(\bar{S}, R_1, \ldots, R_n) = 1$$

Prediction: the state of greater information $I(\Delta S; R_1, \ldots, R_n)$ or greater Probability $\mathcal{P}(S, R_1, \ldots, R_n)$ is predicted

N.B. As $\mathcal{P}(S, R_1, \ldots, R_n)$ cannot be exactly calculated various approximations have been used

Results: for 3-state prediction (H, E and C)

Q_3 = number of correctly predicted residues / number of predicted residues

Basic Approximations

Basic Approximations

1 Local sequence : for -8 < m < 8 or 17 amino acids

1 Local sequence : for -8 < m < 8 or 17 amino acids

2 GOR I (1978)

2 GOR I (1978)

$$I(\Delta S_j R_1,\ldots,R_n) \approx I(\Delta S_j;R_j) + \Sigma_{m,m\neq 0} I(\Delta S_j;R_{j+m})$$

$$I(\Delta S_j R_1,\ldots,R_n) \approx I(\Delta S_j;R_j) + \Sigma_{m,m\neq 0} I(\Delta S_j;R_{j+m})$$

3 GOR III (1987)

3 GOR III (1987)

$$I(\Delta S_j;R_1,\ldots,R_n) \approx I(\Delta S_j;R_j) + \Sigma_{m,m\neq 0} I(\Delta S_j;R_{j+m}|R_j)$$

$$I(\Delta S_j;R_1,\ldots,R_n) \approx I(\Delta S_j;R_j) + \Sigma_{m,m\neq 0} I(\Delta S_j;R_{j+m}|R_j)$$

With :

With :

$$n(s,R_j,R_{j+m}) = n(s,R_j,R_{j+m})obs + n(s,R_{j+m})d$$

$$n(s,R_j,R_{j+m}) = n(s,R_j,R_{j+m})obs + n(s,R_{j+m})d$$

$$n(s,R_{j+m})d + n(\bar{s},R_{j+m})d = 225 \text{ or } 200$$

$$n(s,R_{j+m})d + n(\bar{s},R_{j+m})d = 225 \text{ or } 200$$

4 GOR IV (1996)

Available by ftp anonymous at proline.jouy.inra.fr in /pub/GOR

Server at NIH http://absalpha.dcrt.nih.gov:8008/

Basic approximations of GOR IV

Basic approximations of GOR IV

- Local sequence (LocSeq) : for -8 < m < 8 or 17 amono acids

- Local sequence (LocSeq) : for -8 < m < 8 or 17 amono acids

- All pairs are counted

- All pairs are counted

$$\log\frac{P(S_j, LocSeq)}{P(n-S_j, LocSeq)} = \frac{2}{17}\sum_{m=-8}^{+8}\log\frac{P(S_j, R_{j+m}, R_{j+n})}{P(n-S_j, R_{j+m}, R_{j+n})} - \frac{15}{17}\sum_{m=-8}^{+8}\log\frac{P(S_j, R_{j+m})}{P(n-S_j, R_{j+m})}$$

GOR IV RESULTS

	# H segments	# E segments	Total #
Obs.	1989	**2587**	**4576**
Pred.	2148	2043	4191

	Ave length H	Ave. length E
Obs.	10.9	5.9
Pred.	10.6	4.1

Overlap	H segments	E segments
75%	51.1	23.7
50%	70.0	42.0
25%	75.7	50.2

	H	E	C
Qprd.	64.7	60.7	65.1
Qobs.	67.0	36.5	75.8

Q3 = 64.4% ± 1.2% (2 σ) **267 proteins or 63 566 AA residues**

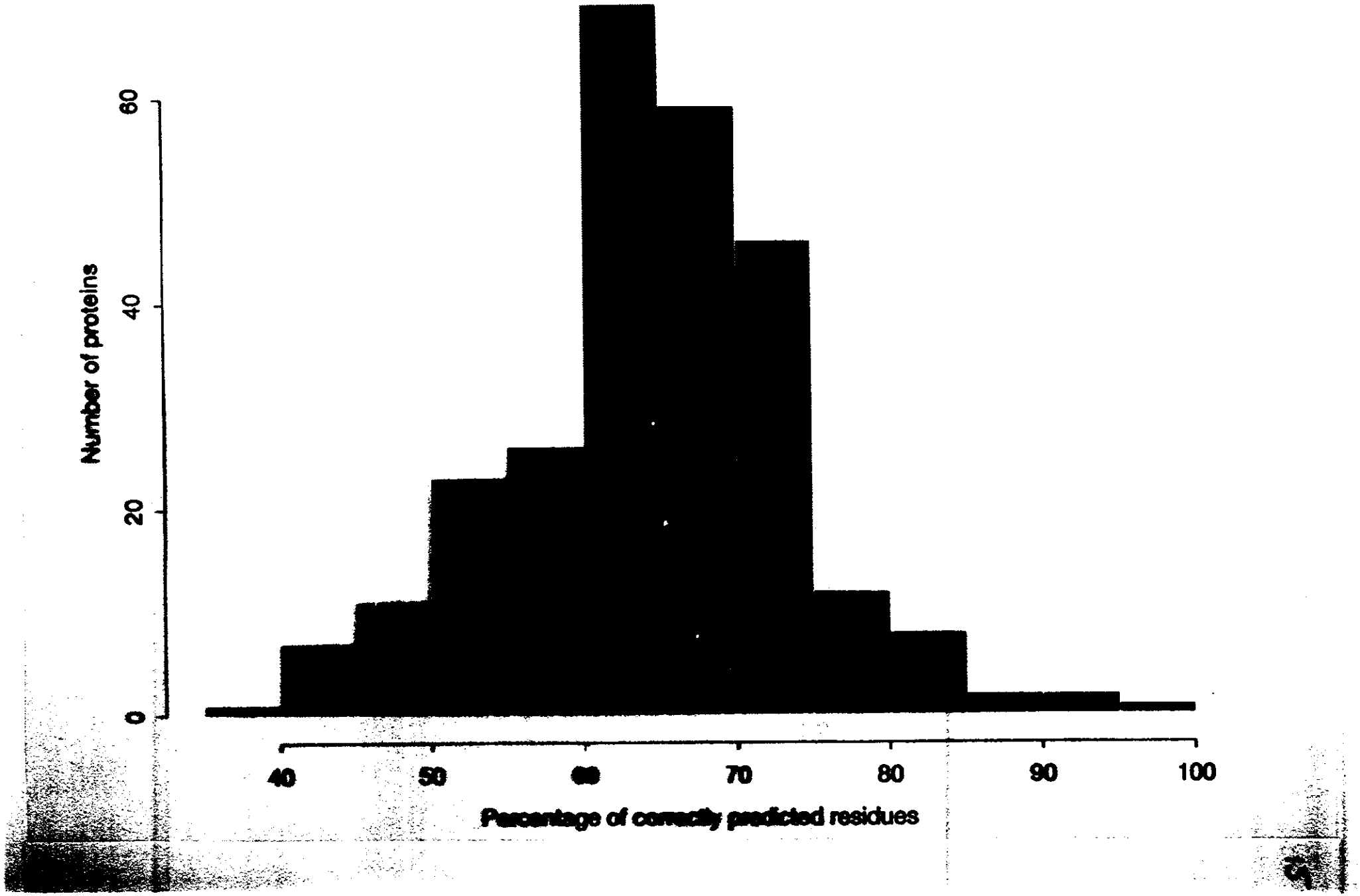
ftp ~~pebiomhm.nih.gov~~ ~~/gibrat/GOR~~

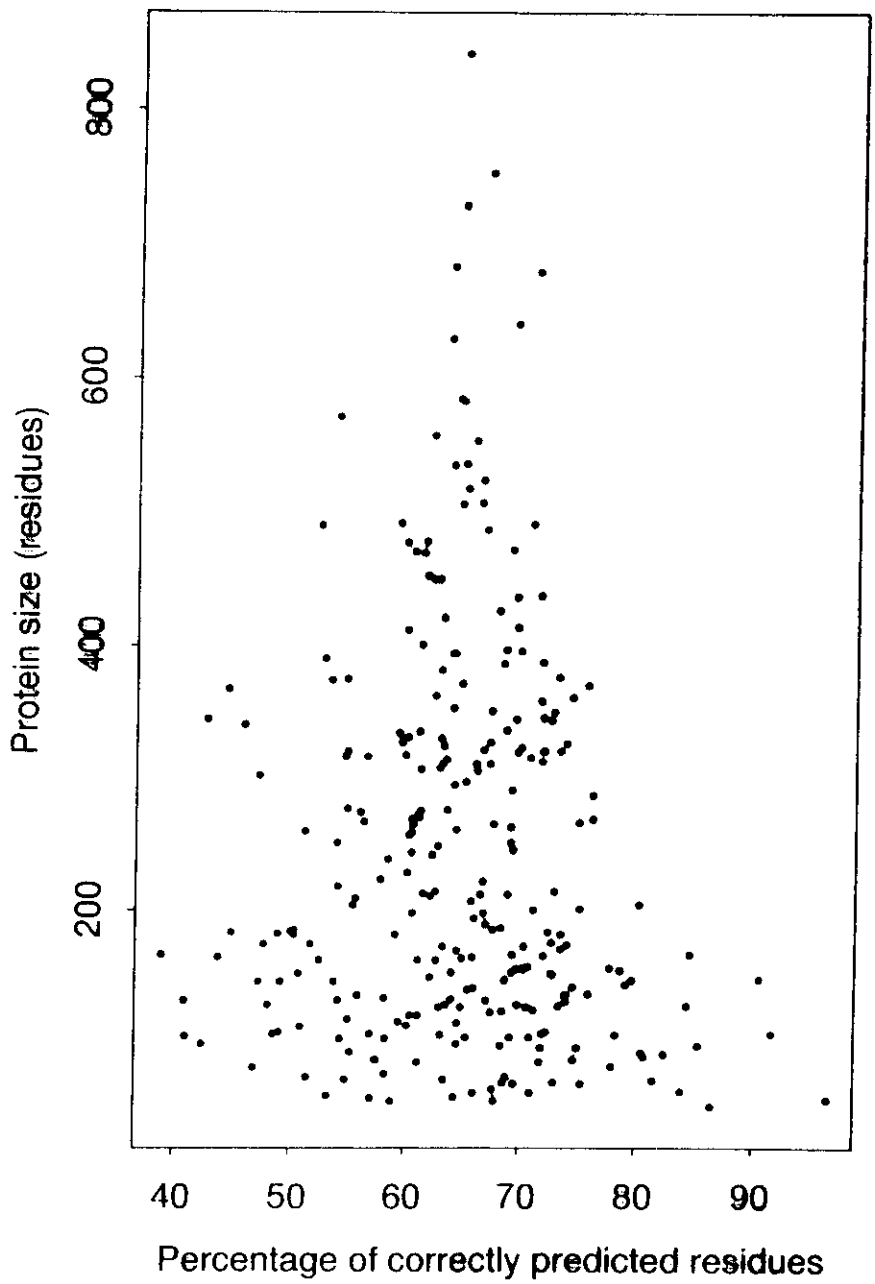
proline.jouy.inra.fr /pub/GOR

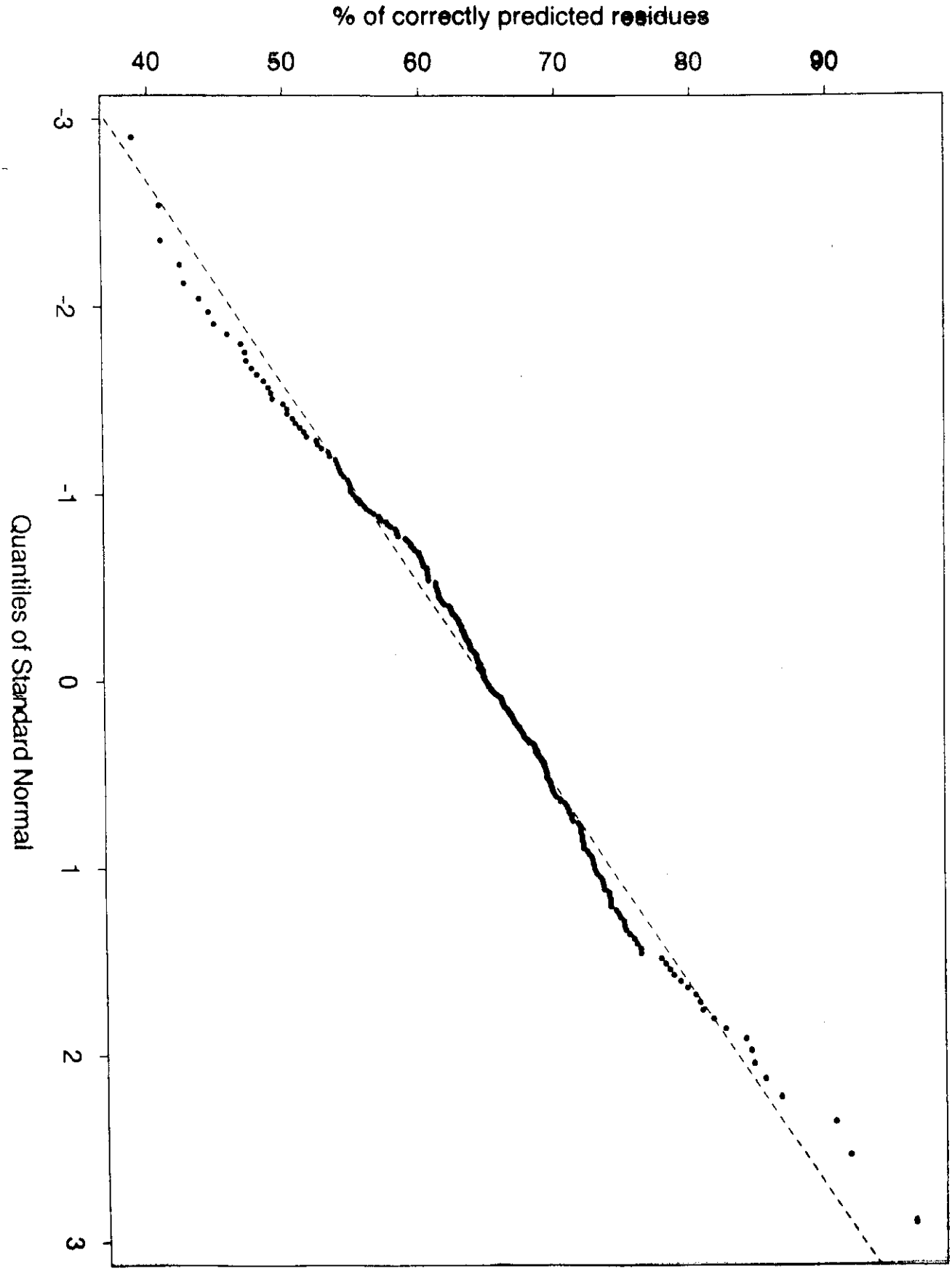
Table III

Prediction of Eglin

Seq	Obs	Prd	pH	pE	pC
T	X	C	0.00	0.00	1.00
E	X	C	0.00	0.02	0.98
F	X	C	0.00	0.08	0.92
G	X	C	0.01	0.13	0.87
S	X	C	0.02	0.14	0.84
E	X	C	0.04	0.24	0.72
L	X	C	0.09	0.33	0.59
K	C	C	0.15	0.24	0.61
S	C	C	0.19	0.16	0.65
F	C	C	0.12	0.12	0.77
P	C	C	0.29	0.12	0.59
E	C	C	0.35	0.26	0.39
V	C	C	0.37	0.30	0.34
V	C	C	0.35	0.24	0.41
G	C	C	0.25	0.20	0.55
K	C	C	0.26	0.27	0.47
T	C	C	0.24	0.34	0.42
V	H	C	0.37	0.19	0.44
D	H	H	0.54	0.08	0.38
Q	H	H	0.58	0.11	0.30
A	H	H	0.61	0.11	0.28
R	H	H	0.56	0.19	0.25
E	H	H	0.50	0.27	0.24
Y	H	H	0.46	0.35	0.19
F	H	H	0.34	0.44	0.22
T	H	H	0.29	0.38	0.32
L	H	C	0.20	0.35	0.44
H	H	C	0.09	0.22	0.69
Y	C	C	0.03	0.11	0.86
P	C	C	0.05	0.06	0.89
Q	C	C	0.09	0.15	0.76
Y	C	C	0.08	0.29	0.63
N	E	C	0.07	0.31	0.61
V	E	E	0.06	0.65	0.30
Y	E	E	0.04	0.75	0.21
F	E	E	0.02	0.76	0.22
L	E	C	0.01	0.30	0.69
P	E	C	0.02	0.09	0.88


Number of proteins
0
20
40
60
40
50
70
80
90
100
Percentage of correctly predicted residues


Protein size (residues)
200
400
600
800
40
50
60
70
80
90
Percentage of correctly predicted residues


% of correctly predicted residues
40
50
60
70
80
90
Quantiles of Standard Normal
-3
-2
-1
0
1
2
3

A nearest-neighbor approach to protein secondary structure prediction:

Similarity Peptide Analysis (SIMPA)

Principie :

Similar peptides of same sequence length n will have same secondary structure

Define :

a similarity matrix between amino acids: blosum62 *

a window length of n= 15 amino acids*

a cutoff of similarity = 13, above it, the observed conformation is assigned to the test sequence with this similarity score*

References

Levin, Robson and Garnier, FEBS Lett., 205, 303-308 (1986)

Levin and Garnier, BBA, 955, 283-295 (1988)

*Levin and Garnier (1996) to be published

Similarity Peptide Analysis (SIMPA)*

Principle :

similar peptides of same sequence length n will have similar secondary structure tendencies.

Define :

a similarity matrix of 20x20 amino acids

a window length of n=17 amino acids

a cutoff of 7

* Levin and Garnier, 1988.

A simplified example of SIMPA (I)

(Unitary matrix, n = 7, cutoff=2, DC=1)

Unknown sequence: S L C G T V A D L K...

Reference database: A L P G C T A L E G A V...
C C E E E E H H H H H H...

Search:

Step 1

A L P G C T A L E G A V...
S L C G T V A

Score=3

Step 2

A L P G C T A L E G A V...
S L C G T V A

Score=1

All the database

Step i+1

A L P G C T A L E G A V...
L C G T V A D

Score=0

All the database and
All the sequence length to be predicted

A simplified example of SIMPA (II)

Search results for : S L C G T V A D L K....

Obs. C C E E E E H H H H H H....
Seq. A L P G C T A L E G A V....
S L C G T V A Score=3
L C G T V A D Score=3
etc...

Match-score table and prediction

Unknown	H	E	C	Prediction
S	0	0	3	C
L	0	0	3+3	C
C	0	3+3	0	E
G	0	3+3	0	E
T	0	3+3	0	E
V	0	3+3	0	E
A	3+3	0	0	H
D	3	0	0	H

etc...

Table 1

Secondary structure prediction of homologous proteins with the SIMPA method*

File Name	Identity %	Basic Prediction %	Homologous Prediction %
2 ACT	48	53.7	83.5
2 ALP	39	50.0	77.8
2 ABX	42	81.1	75.7
5 CHA (1)	34-43	65.7	89.3
5 CHA (2)	41-45	60.8	87.6
3 C2C	40-44	64.3	87.5
2 EST	30-43	62.1	86.7
2 EBX	42	71.0	96.7
1 FDX	43	70.4	75.9
2 FDI	43	64.2	70.8
2 HHB (1)	35-45	61.7	92.9
2 HHB (2)	45	56.9	89.7
2 LHB	35	73.2	88.6
1 MCP (2)	44	67.6	88.7
1 FB4 (2)	44	75.6	87.3
2 PKA (1)	34-42	63.8	96.3
2 PKA (2)	30-41	60.5	91.5
1 MBN	25	80.4	88.9
1 PPD	48	63.7	92.0
2 SGA	39	54.1	76.8
3 RP2	34-38	56.7	84.8
1 TPO	36-45	61.4	89.7
Global: 22	**25-48**	**64.5 (SD 8%)**	**86.3 (SD 6.7%)**

*Selected from Levin and Garnier (7). The basic prediction was obtained when all homologous proteins of more than 22% identity were removed from the data base. The homologous prediction is explicitly using the homologies in the data base.

Similarity Peptide Analysis (SIMPA) RESULTS

	# H segments	# E segments	Total #
Obs.	2515	3430	5945
Pred.	2510	3316	5826

	Ave length H	Ave. length E
Obs.	10.8	5.3
Pred.	10.6	4.1

	H	E	C
Qprd.	70.0	62.0	67.6
Qobs.	68.4	46.8	76.8

Q_3 = 67.5% ± 1.2% (2 σ) ~ 320 proteins or 82 527 AA residues

With evolutionary information (Consensus)

Q_3 = 71.3% ± 1.2% (2 σ) ~ 190 proteins or 49 244 AA residues

[redacted]

Comparison of secondary structure predictions

A Single sequence

	Probability Index (%) 3 states, H, E & C
Information Theory	
GOR I (1)	56.9
GOR III (2)	63.0
GOR IV +	64.4
Similarity analysis	
SIMPA (3)	63.7 (+homol.)
SIMPA* (1995)	67.3 (+homol.) 67.5
Yi + Lander (1994)	68.0
Neural network	
Qian & Sejnowski (4)	64.3
Joint predictions	
COMBINE (5)	65.5

NB. SD on the *means* about 1-2%

1 Garnier et al., JMB, 120, 97-120,(1978), 1988.
2 Gibrat et al., JMB, 198, 425-443,(1987), Biochemistry, 30, 1578-1586, (1991)
3 Levin & Garnier, BBA, 955, 283-295, (1988)
4 Qian & Sejnowski, JMB, 202, 865-884, (1988)
5 Bleu et al., Prot. Eng., 2, 185-191,(1988)
+ Garnier et al. Methods in Enzymology (1995, sous presse)

B multiple homologous sequences

	%
. SIMPA* (Levin et al. 1995)	71.3
. PHD (Rost et al. 1994)	71-72 (61.7 sans [illegible])

Anti-progesterone antibody DB3*, CDR-H1, -H2, -H3

CLUSTAL W(1.4) multiple sequence alignment

```
                                  H1                              H2
1DBAb_FAB'    QIQLVQSGPELKKPGETVKISCKASGYAFTNYGVNWVKEAPGKELKWMGWINIYTG--EP
1MCPb_IMMU    EVKLVESGGGLVQPGGSLRLSCATSGFTFSDFYMEWVRQPPGKRLEWIAASRNKGNKYTT
2F19b_FAB     QVQLQQSGAELVRAGSSVKMSCKASGYTFTSYGVNWVKQRPGQGLEWIGYINPGKG--YL
2FB4b_IMMU    EVQLVQSGGGVVQPGRSLRLSCSSSGFIFSSYAMYWVRQAPGKGLEWVAIIWDDGS--DQ
2FBJb_IG*A    EVKLLESGGGLVQPGGSLKLSCAASGFDFSKYWMSWVRQAPGKGLEWIGEIHPDSG--TI
3HFLb_IG*G    -VQLQQSGAELMKPGASVKISCKASGYTFSDYWIEWVKQRPGHGLEWIGEILPGSG--ST
7FABb_LAMB    AVQLEQSGPGLVRPSQTLSLTCTVSGTSFDDYYWTWVRQPPGRGLEWIGYVF-YTG--TT
3HFMb_IG*G    DVQLQESGPSLVKPSQTLSLTCSVTGDSITSDYWSWIRKFPGNRLEYMGYVS-YSG--ST
                ..* .**  . .   .. ..*  .*        *..   **  *   .
                                                                H3
1DBAb_FAB'    TYVDDFKGRFAFSLETSASTAYLEINNLKNEDTATYFCTRGDYVN-----WYF--DVWGA
1MCPb_IMMU    EYSASVKGRFIVSRDTSQSILYLQMNALRAEDTAIYYCARNYYGST------WYFDVWGA
2F19b_FAB     SYNEKFKGKTTLTVDRSSSTAYMQLRSLTSEDAAVYFCARSFYGGSDLAVYYF--DSWGQ
2FB4b_IMMU    HYADSVKGRFTISRNDSKNTLFLQMDSLRPEDTGVYFCARDGGHGFCSSASCFGPDYWGQ
2FBJb_IG*A    NYTPSLKDKFIISRDNAKNSLYLQMSKVRSEDTALYYCARLHYYG--------YNAYWGQ
3HFLb_IG*G    NYHERFKGKATFTADTSSSTAYMQLNSLTSEDSGVYYCLHGNYD--------F--DGWGQ
7FABb_LAMB    LLDPSLRGRVTMLVNTSKNQFSLRLSSVTAADTAVYYCARNLIAGG--------IDVWGQ
3HFMb_IG*G    YYNPSLKSRISITRDTSKNQYYLDLNSVTTEDTATYYCANWDG-----------DYWGQ
                  . .     .     . . .   *.  *.*                   **
```

*In red are observed beta-strands except for DB3 (1DBA) which is predicted with Q_3 = 94% (59.8).

Anti-progesterone antibody DB3*, CDR-H3

CLUSTAL W(1.4) multiple sequence alignment

```
1DBAb_FAB'        EDTATYFCTRGDYVN-----WYF--DVWGA
1MCPb_IMMU        EDTAIYYCARNYYGST------WYFDVWGA
2F19b_FAB         EDAAVYFCARSFYGGSDLAVYYF--DSWGQ
2FB4b_IMMU        EDTGVYFCARDGGHGFCSSASCFGPDYWGQ
2FBJb_IG*A        EDTALYYCARLHYYG--------YNAYWGQ
3HFLb_IG*G        EDSGVYYCLHGNYD--------F--DGWGQ
7FABb_LAMB        ADTAVYYCARNLIAGG--------IDVWGQ
3HFMb_IG*G        EDTATYYCANWDG-----------DYWGQ
```

Corrected multiple sequence alignment

```
1DBAb_FAB'        EDTATYFCTRGD---YVN----WYFDVWGA
1MCPb_IMMU        EDTAIYYCARNYY--GST----WYFDVWGA
2F19b_FAB         EDAAVYFCARS-FYGGSDLAVYY-FDSWGQ
2FB4b_IMMU        EDTGVYFCARDGGHGFCSSASCFGPDYWGQ
2FBJb_IG*A        EDTALYYCARLH---YYG-----YNAYWGQ
3HFLb_IG*G        EDSGVYYCL----HGNYD------FDGWGQ
7FABb_LAMB        ADTAVYYCARN----LIAGG----IDVWGQ
3HFMb_IG*G        EDTATYYCAN-----WDG-------DYWGQ
```

* In red are observed beta-strands except for DB3 (1DBA) which is predicted.

Anti-progesterone antibody DB3*, CDR-L1.

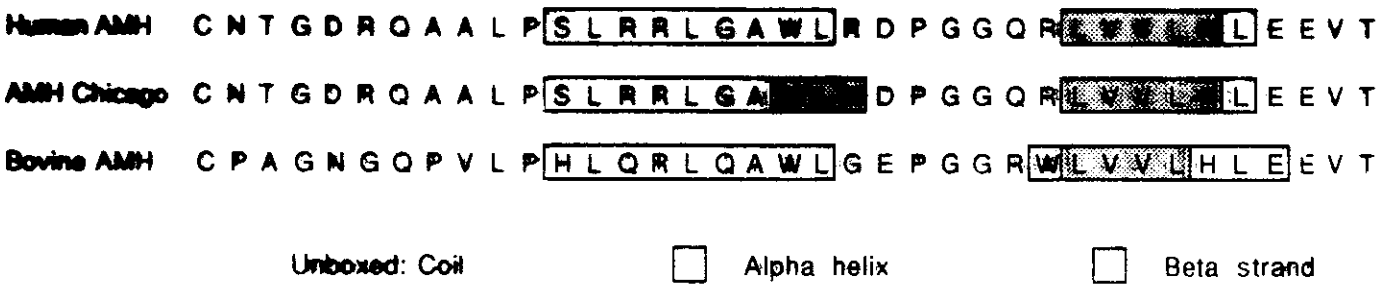
CLUSTAL W(1.4) multiple sequence alignment

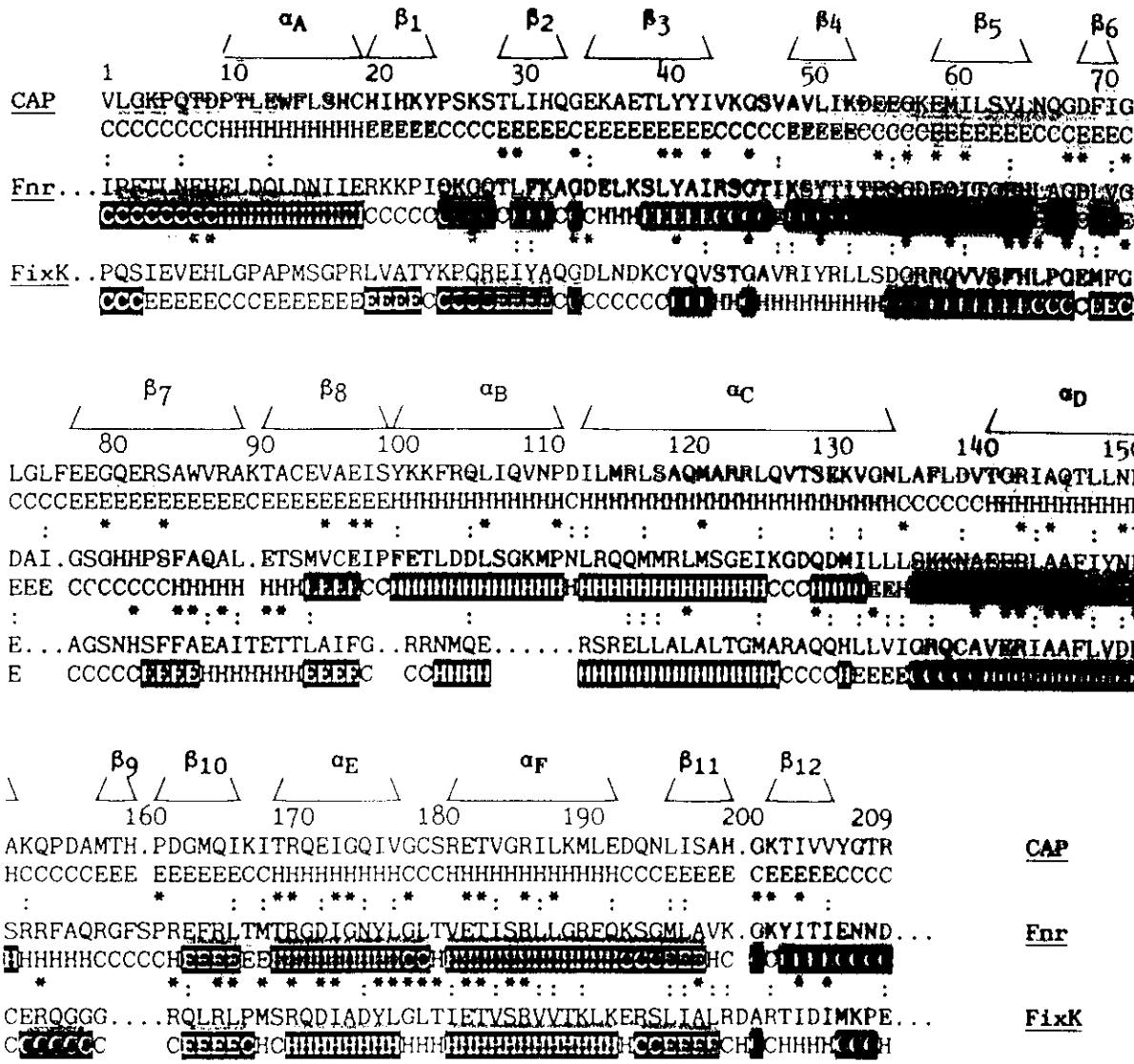
```
1DBAa_FAB'      GDQASISCRSSQSLIHSNGN-TYLHWYLQKPGQ
1MCPa_IMMU      GERVTMSCKSSQSLLNSGNQKNFLAWYQQKPGQ
1REIa_BENC      GDRVTITCQASQDIIKY------LNWYQQTPGK
2F19a_FAB       GDRVTISCRASQDISNY------LNWYQQKPDG
2FB4a_IMMU      GQRVTISCSGTSSNIGS----STVNWYQQLPGM
2FBJa_IG*A      GQKVTITCSASSSVSSL-------NWYQQKSGT
2RHEa_BENC      GQRVTISCTGSATDIGS----NSVIWYQQVPGK
3HFLa_IG*G      GEKVTMTCSASSSVNYM-------YWYQQKSGT
7FABa_LAMB      GQRVTISCTGSSSNIGAG---HNVKWYQQLPGT
3HFMa_IG*G      GNSVSLSCRASQSIGNN------LHWYQQKSHE
```

Corrected multiple sequence alignment

```
1DBAa_FAB'      GDQASISCRSSQSLIHSNGNTY-LHWYLQKPGQ
1MCPa_IMMU      GERVTMSCKSSQSLLNSGNQKNFLAWYQQKPGQ
1REIa_BENC      GDRVTITCQA---SQDIIKY---LNWYQQTPGK
2F19a_FAB       GDRVTISCRA---SQDISNY---LNWYQQKPDG
2FB4a_IMMU      GQRVTISCS--GTSSNIGSSTV--NWYQQLPGM
2FBJa_IG*A      GQKVTITCSA---SSSVS----SLHWYQQKSGT
2RHEa_BENC      GQRVTISCT--GSATDIGSNS--VIWYQQVPGK
3HFLa_IG*G      GEKVTMTCS---ASSSVNY----MYWYQQKSGT
7FABa_LAMB      GQRVTISCT-GSSSNIGAGHNV--KWYQQLPGT
3HFMa_IG*G      GNSVSLSCRA---SQSIGNN---LHWYQQKSHE
```

*In red are observed beta-strands except for DB3 (1dba) which is predicted.


Human AMH C N T G D R Q A A L P S L R R L G A W L R D P G G Q R L V V L L L E E V T
AMH Chicago C N T G D R Q A A L P S L R R L G A D P G G Q R L V V L L L E E V T
Bovine AMH C P A G N G Q P V L P H L Q R L Q A W L G E P G G R W L V V L H L E E V T
Unboxed: Coil
Alpha helix
Beta strand


αA β1 β2 β3 β4 β5 β6
1 10 20 30 40 50 60 70
CAP
Fnr...
FixK..
β7 β8 αB αC αD
80 90 100 110 120 130 140 150
DAI.GSGHHPSFAQAL.ETSMVCEIPFETLDDLSGKMPNLRQQMMRLMSGEIKGDQDMILLLSKKNAFRRLAAEIYNL
E...AGSNHSFFAEAITETTLAIFG..RRNMQE.....RSRELLALALTGMARAQQHLLVIGRQCAVERIAAFLVDL
β9 β10 αE αF β11 β12
160 170 180 190 200 209
AKQPDAMTH.PDGMQIKITRQEIGQIVGCSRETVGRILKMLEDQNLISAH.GKTIVVYGTR
SRRFAQRGFSPREFRLTMTRGDIGNYLGLTVETISRLLGRFQKSGMLAVK.GKYITIENND...
CERQGGG....RQLRLPMSRQDIADYLGLTIETVSRVVTKLKERSLIALRDARTIDIMKPE...

1- CAP promotor activated by mutant Fnr (Spiro &Guest)

Mutated Amino acids	Conserved amino acids
V_{180} R	E181
S_{184} G	T182
G_{188} K	R185

2- Promotor consensus sequences (Ebright et al.)

CAP T G T G A .....T C A C A

R G

180 184

Fnr T T T G A.....T C A A A

V S

180 184

Folding simulation

Principles:

Search for the lowest (free) energy

Problems:

lowest ?

number of variables:

cartesian or internal coordinates?

multiple minima problem

$\sim 10^N$

N number of residues

Which force field? 90% of computing time

$\sim n^2$, n = number of atoms

Robson. Ploth. JMB(1986), 188, 259-281

Potential energy for n atoms and m rotatable bonds

$$E_t(s) = \sum_{i=1}^{i=j} \sum_{j=2}^{j=n} e_{ij}(s) + \sum_{k=1}^{k=m} I(\alpha_x)$$

$$e_{ij} = A_{ij} s^9 - B_{ij} s^6 + C_{ij} s$$

$$s = \frac{1}{r'} \quad \text{with} \quad r' = \frac{(1 + 1.5r^2 + 0.0625r^4)}{(2 + 0.5r^2)}$$

Ex.

Separation A		r.s.w. energy Kcal/mole		Electrostatic energy kcal/mole	
r	r'	r^{13}	r'^{7}	r	r'
0.1	0.5	5.10^{13}	2.10^{7}	703	139
4.1	4.2	-0.157	-0.142	17.1	16.7
15	31	0	0	4.6	2.2

$$I(\alpha_x) = \frac{K_x}{2} \left(1 + \cos\left(N_x(\alpha_x - a_x)\right)\right)$$

ex → 35 AA 66 × dihedral angles

82% ≤ 2σ / X ray

18% ≠ X ray

10% → most frequent

6% no preference

2-x wrong

Program ESAP: Extended Simulated Annealing Procedure

Higo et al. Biopolymers (1992), 22, 33-43

Gibrat et al. Immunomethods (1992) 1, 107-125

1- Target Function

$$E_t = \lambda_i E_i + \lambda_c E_c \quad \text{with } T = 300/\lambda$$

$$E_i = \sum_{\text{non bonded}} E_{vdw} + E_{elec.} + E_{rot}$$

Robson-Platt potential

$$E_c = (r - r_x)^2$$

2- Monte Carlo Metropolis with Noguti-Go Algorithm

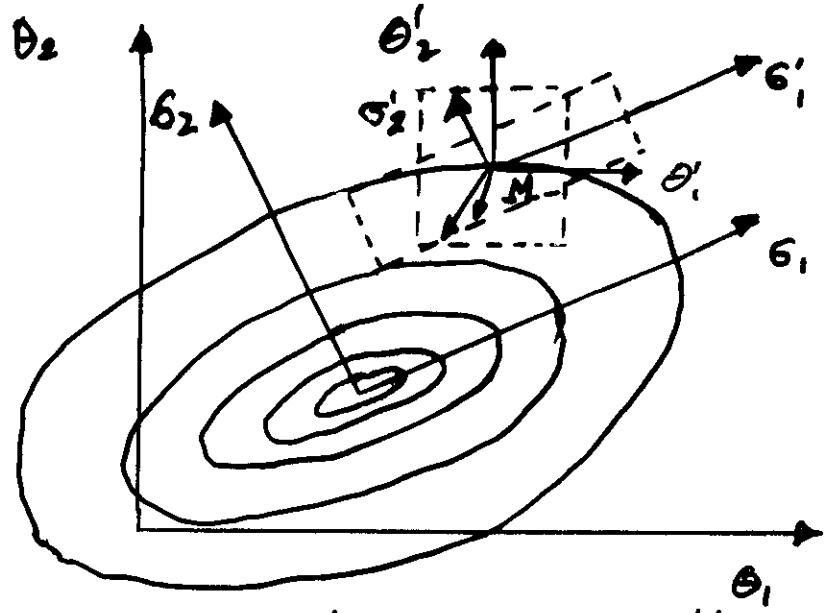
3- Annealing by increasing λ

stage	1	2	3	4	5	6
λ_c	0.01	0.1	0.5	2.5	10.0	10.0
λ_i	0.01	0.01	0.05	0.25	1.0	1.0
Nb steps/10^4	2	8	8	8	4	4
$T = 300/\lambda_i$ K	$3\,10^4$	$3\,10^4$	$6\,10^3$	$1.2\,10^3$	300	300

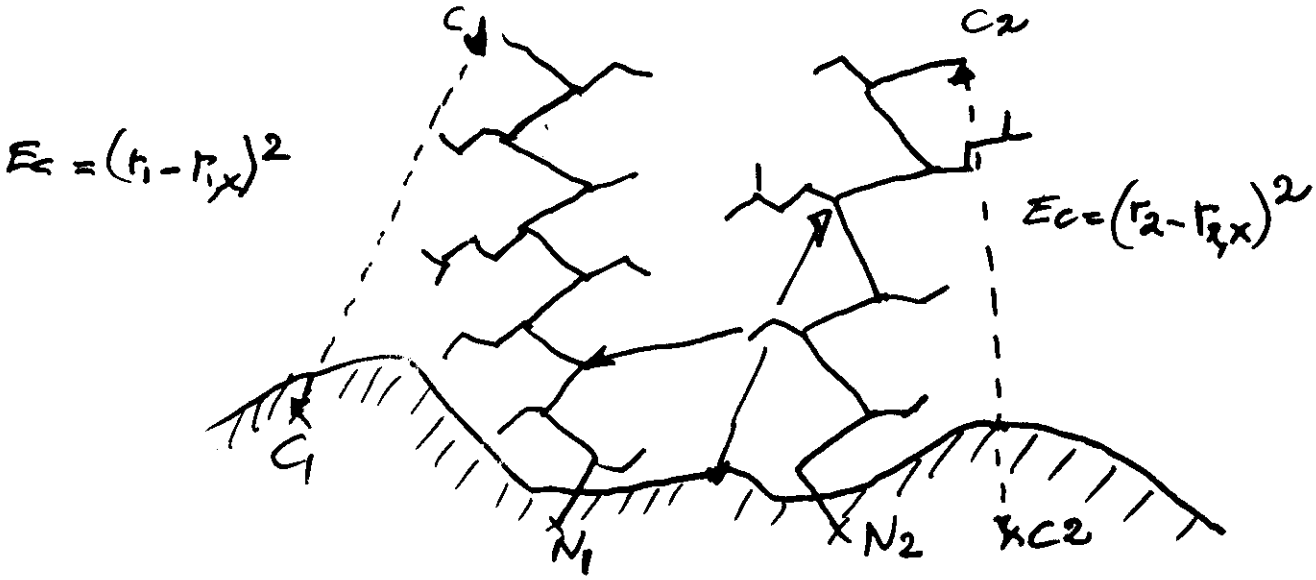
4- Results: Lowest Boltzman average of E_t with corresponding R

$$U = \langle E_t \rangle = N_{total}^{-1} \sum_{j=1,M} (n_j + 1)\, E_{t,j}$$

$$R = \langle r \rangle = N_{total}^{-1} \sum_{j=1,M} (n_j + 1)\, r_j$$



Noguti-GŌ Algorithm



Fixed (X-Ray coordinates)

ESAP starting conformation (2 Loops)

ESAP : Extended Simulated Annealing Process

	Loop	AA seq	dihedral angles used	Atoms used
L2	Bence Jones	YNDLLPS	26	67
H_1 + H_2	McPC603	TFSDF + ASRNKGNKY	61	148

MC CPU times* (min)

Stage:	1	2	3	4	5	6	Total
L2	3.5	40	40	40	20	20	163.5
H_1+H_2	16	160	160	160	80	80	656
MC Steps ($\times 10^4$)	2	8	8	8	4	4	$3.6\,10^5$

Modeling of hypervariable loops of Anti-progesterone antibody B83

Loops	Sequence	number of Amino Acids	RMSD in Å backbone	RMSD in Å All atoms
L1	SQSLLESRTRY	12	2.9	4.5
L2	KVSNRFTGVPDR	12	0.8	2.0
L3	SSHVPP	6	0.2	0.8
H1	GYAFTNYG	8	0.7	2.0
H2	NIYTGE	6	0.7	2.1
H3	DYVNWY	6	0.8	2.6
	TOTAL = 50		Average = 1.02	2.3
			1.06*	2.15*

* average on 18 modeled loops