CONF-9507246--2

ISMB-95
ROBINSON COLLEGE, CAMBRIDGE

Tutorial Programme
Sunday 15 July 1995

TUTORIAL T2

Intelligent Systems for the Molecular Biologist

(Douglas L Brutlag)

DISCLAIMER

Third International Conference on
Intelligent Systems for Molecular Biology

Tutorial T2

Intelligent Systems for the Molecular Biologist

Douglas L. Brutlag
Stanford University

July 16, 1995
10:00 AM - 1:00 PM

ISMB - 1995
Intelligent Systems for Molecular Biologists
Doug Brutlag

General Reference Books and Reviews

Books

Adams, M. D., Fields, C. and Venter, J. C. (1994). *Automated DNA Sequencing and Analysis.* New York: Academic Press, 368 pages.

Altman, R., Brutlag, D., Karp, P., Lathrop, R. and Searls, D. (1994). *Second International Conference on Intelligent Systems for Molecular Biology.* Menlo Park, CA: AAAI Press, 388 pages.

Bishop, M. J. (1994). *Guide to Human Genome Computing.* London: Academic Press, 350 pages.

Doolittle, R. F. (1986). *Of Urfs and Orfs: A Primer on How to Analyze Derived Amino Acid Sequences.* University Science Books, Mill Valley, California.

Doolittle, R. F. (1990). Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences (1 ed.). Methods in Enzymology Volume 183, New York: Academic Press.

Fasman, G. D. (1989). *Prediction of Protein Structure and the Principles of Protein Conformation.* New York NY: Plenum Press,

James, M. (1985). *Classification Algorithms* (1st ed.). New York, NY: John Wiley and Sons.

Gribskov, M. and Devereux, J. (1991). *Sequence Analysis Primer.* New York: Stockton Press, 279.

Hunter, L. (1993). *Artificial Intelligence and Molecular Biology.* Menlo Park, CA: AAAI Press, 470 pages.

Hunter, L., Searls, D. and Shavlik, J. (1993). *First International Conference on Intelligent Systems for Molecular Biology.* Menlo Park, CA.: AAAI Press.

Lesk, A. (1991). *Protein Architecture: A Practical Approach* . Oxford: IRL Press at Oxford University Press.

Sankoff, D. and Kruskal, J. B. (1983). Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison . Reading, Massachusetts: Addison-Wesley.

Smith, D. W. (1994). *Biocomputing: Informatics and Genome Projects.* New York: Academic Press Inc., 336 pages.

Trifonov, E. N. and Brendel, V. (1986). *Gnomic: A Dictionary of Genetic Codes.* Balaban Publishers, Philadelphia, Pennsylvania.

von Heijne, Gunnar (1987). *Sequence Analysis in Molecular Biology: Treasure Trove or Trivial Pursuit,* Academic Press, New York.

Waterman, M. (1988). *Mathematical Methods for DNA Sequences,* CRC Press, Cleveland Ohio.

Reviews

Altschul, S. F., Boguski, M. S., Gish, W. and Wootton, J. C. (1994). Issues in searching molecular sequence databases. *Nat Genet* 6 (2), 119-29.

Chao, K.-M., Hardison, R. C. and Miller, W. (1994). Recent developments in linear-space alignment methods: A survey. *J. Computational Biology* 1 (4), 271-291.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* 22 , 521-565.

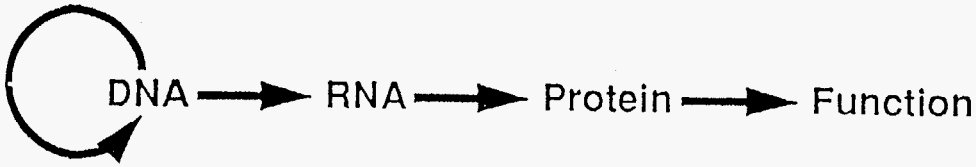
Garnier, J. and Levin, J. M. (1991). The protein structure code: what is its present status? *Comput Appl Biosci* 7 (2), 133-42.

Stormo, G. D. (1988). Computer methods for analyzing sequence recognition of nucleic acids. *Annu. Rev. Biophys. Biophys. Chem.* 17, 241-263.

Tyler, E. C., Horton, M. R. and Krause, P. R. (1991). A review of algorithms for molecular sequence comparison. *Comput Biomed Res,* 24(1), 72-96.

Molecular Biology

Flow of Genetic Information

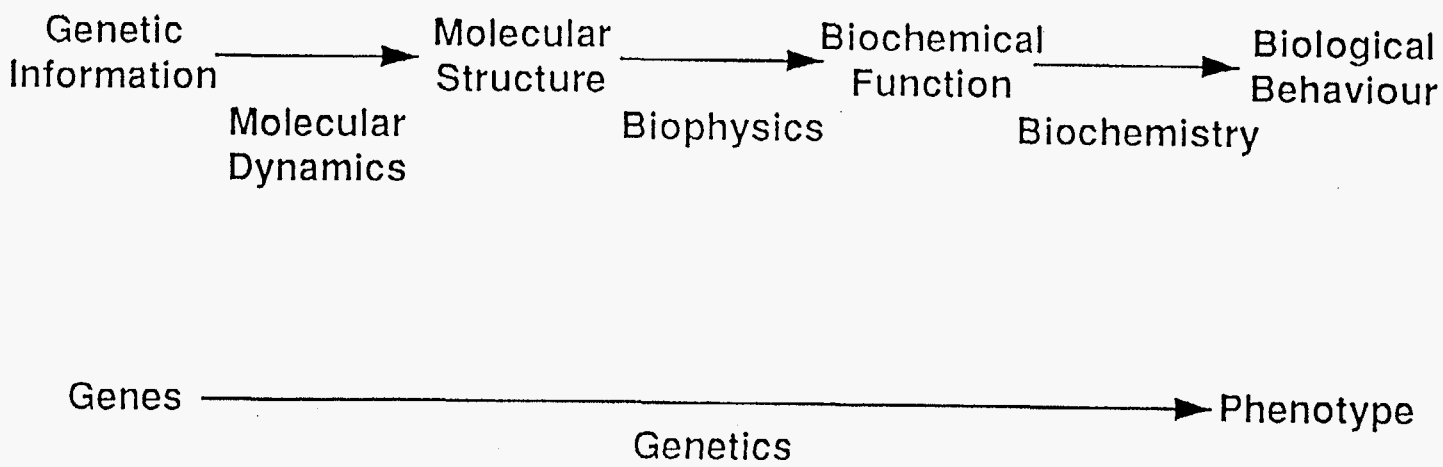


- Mechanism
- Specificity
- Regulation

Molecular Biology is an Information Science

Genetic Information → Molecular Structure → Biochemical Function → Biological Behavior

Molecular Biology is an Information Science



Challenges Understanding the Genetic Message

Genetic Information → Molecular Structure → Biochemical Function → Biological Behavior

- Genetic information is redundant
- Structural information is redundant
- Single genes have multiple functions
- Genes are one dimensional but gene products are three-dimensional

The Genetic Code

Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile
GCA GCC GCG GCT	CGA CGC CGG CGT AGA AGG	GAC GAT	AAC AAT	TGC TGT	GAA GAG	CAA CAG	GGA GGC GGG GGT	CAC CAT	ATA ATC ATT
Leu	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**
CTA CTC CTG CTT TTA TTG	AAA AAG	ATG	TTC TTT	CCA CCC CCG CCT	TCA TCC TCG TCT AGC AGT	ACA ACC ACG ACT	TGG	TAC TAT	GTA GTC GTG GTT

Redundancy in Genomic Sequences

- DNA is Double-Stranded
- Genetic Code
- Structural Redundancies
- Acceptable Amino Acid Replacements
- Intron-exon Variation
- Strain Variation
- Sequencing Errors

Challenges Understanding the Genetic Message

Genetic Information → Molecular Structure → Biochemical Function → Biological Behavior

- Genetic information is redundant
- Structural information is redundant
- Single genes have multiple functions
- Genes are one dimensional but gene products are three-dimensional

Representations of Protein Structure

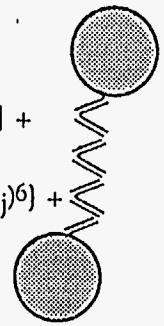
$$\sum_{\text{all bonds}} K_b (b_i - b_o)^2 +$$

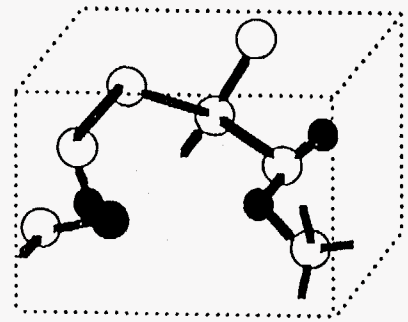
$$\sum_{\text{all angles}} K_\theta (\theta_i - \theta_o)^2 +$$

$$\sum_{\text{all torsion angles}} K_\phi [1 - \cos (n\phi_i + \delta)] +$$

$$\sum_{\text{all nonbonded pairs}} \varepsilon \{(r_o/r_{ij})^{12} - 2 (r_o/r_{ij})^6\} +$$

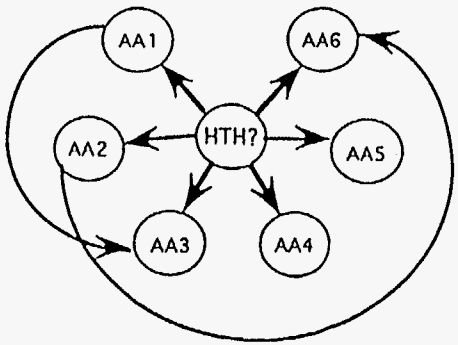
$$\sum_{\text{all partial charge pairs}} q_i q_j / r_{ij}$$





82% Hydrophobic

18% Hydrophilic




```
     40                    50
FPTTKTYFPHF-DLS-----HGS
:|  |:  :|   |  |||       |
YPWTQRFFESFGDLSTPDAVMGN
          40          50
```

==> Representation is Key to Understanding

Multiple Representations of Sequences

Weight Matrices, Blocks or Profiles

Position

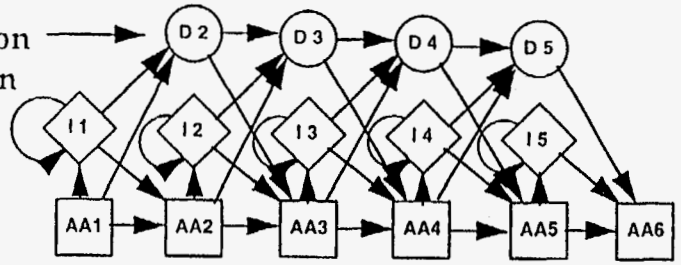
	1	2	3	4	5	6	7	8	9	10	11	12
A	2	1	3	13	10	12	67	4	13	9	1	2
R	7	5	8	9	4	0	1	16	7	0	1	0
N	0	8	0	1	0	0	0	2	1	1	10	0
D	0	1	0	1	13	0	0	12	1	0	4	0
C	0	0	1	0	0	0	0	0	0	2	2	1
Q	1	1	21	8	10	0	0	7	6	0	0	2
E	2	0	0	9	21	0	0	15	7	3	3	0
G	9	7	1	4	0	0	8	0	0	0	46	0
H	4	3	1	1	2	0	0	2	2	0	5	0
I	10	0	11	1	2	10	0	4	9	3	0	16
L	16	1	17	0	1	31	0	3	11	24	0	14
K	3	4	5	10	11	1	1	13	10	0	5	2
M	7	1	1	0	0	0	0	0	5	7	1	8
F	4	0	3	0	0	4	0	0	0	10	0	0
P	0	6	0	1	0	0	0	0	0	0	0	0
S	1	17	0	8	3	1	3	0	2	2	2	0
T	5	22	3	11	1	5	0	2	2	2	0	5
W	2	0	0	0	0	0	0	0	0	1	0	1
Y	1	0	4	2	0	1	0	0	2	4	0	1
V	6	3	1	1	2	15	0	0	2	12	0	28

Consensus Sequences

Zinc Finger (C2H2 type)
CX{2,4}CX{12}HX{3,5}H

Sequences of Common Structure or Function

Hidden Markov Model



Sequence Alignments

```
                    10        20        30        40                50
1     VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
       |:| :|: | |:||||     | |:||| |: : :|:|  :|  |       |:  |
2     HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
                    10        20        30        40        50
```

Initial Score	=	63	Optimized Score	=	98	Significance =	5.51
Residue Identity	=	14%	Matches	=	21	Mismatches =	22
Gaps	=	2	Conservative Substitutions			=	11

Consensus Patterns

- Active site of trypsin-like serine proteases

G D S G G

- Zinc Finger (C2H2 type)

C X{2,4} C X{12} H X{3,5} H

- N-Glycosylation Site

N [^P] [S T] [^P]

- Homeobox Domain Signature

[LIVMF] X{5} [LIVM] X{4} [IV] [RKQ] X W X{8} [RK]

Sequence Alignment

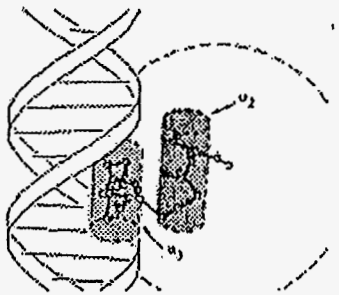
```
     X         220       230       240       250           X
     F--SGGNTHIYMNHVEQCKEILRREPKELCELVISGLPYKFRYLSTKE-QLK-Y
     |  : |::|||:||:|  |  | |||: : :| |     |   ::::: |:: |
LKPGDFIHTLGDAHIYLNHIEPLKIQLQREPRPFPKLRILRKVEKIDDFKAEDFQIEGYNPHPTIK
     X         260       270       280       290           X
```

$$\text{Score} = \sum_{\text{Region Start}}^{\text{Region End}} \text{Similarity-weights} - \sum_{\text{Region Start}}^{\text{Region End}} \text{Penalties}$$

where:

Penalty = Gap-penalty + Size-of-gap x Gap-size-penalty

Weight Matrix



Structural or functional motif

⇓

Examples of motif

```
HSGEQLAETLGMSRAAINKHIQ
VTLYDVAEYAGVSYQTVSRVVN
AMIKDVALKAKVSTATVSRALM
ATIKDVAKRAGVSTTTVSHVIN
ITIYDLAELSGVSASAVSAILN
LHLKDAAALLGVSEMTIRRDLN
TAYAELAKQFGVSPGTIHVRVE
GSLTEAAHLLGTSQPTVSRELA
MSQRELKNELGAGIATITRGSN
ITRQEIGQIVGCSRETVGRILK
FDIASVAQHVCLSPSRLSHLFR
LRIDEVARHVCLSPSRLAHLFR
MTRGDIGNYLGLTVETISRLLG
VTLEALADQVGMSPFHLHRLFK
```

⇒

Position

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
A	2	1	3	13	10	12	67	4	13	9	1	2	4	3	6	15	4	4	4	11	0	10
R	7	5	8	9	4	0	1	16	7	0	1	0	1	16	6	6	0	11	28	3	0	16
N	0	8	0	1	0	0	0	2	1	1	10	0	7	1	3	1	0	4	8	0	1	11
D	0	1	0	1	13	0	0	12	1	0	4	0	1	2	0	0	0	0	1	1	0	3
C	0	0	1	0	0	0	0	0	0	2	2	1	0	0	0	0	0	0	0	1	0	0
Q	1	1	21	8	10	0	0	7	6	0	0	2	1	17	7	7	0	2	12	5	2	4
E	2	0	0	9	21	0	0	15	7	3	3	0	1	6	11	0	0	2	0	1	13	6
G	9	7	1	4	0	0	8	0	0	0	46	0	6	0	7	1	0	3	1	1	0	4
H	4	3	1	1	2	0	0	2	2	0	5	0	3	3	0	2	0	2	4	5	0	2
I	10	0	11	1	2	10	0	4	9	3	0	16	0	2	0	1	26	1	0	8	16	0
L	16	1	17	0	1	31	0	3	11	24	0	14	0	2	0	1	21	1	1	12	20	0
K	3	4	5	10	11	1	1	13	10	0	5	2	1	4	1	1	0	1	8	4	5	14
M	7	1	1	0	0	0	0	0	5	7	1	8	0	0	2	0	2	0	0	2	0	1
F	4	0	3	0	0	4	0	0	0	10	0	0	0	0	1	0	0	1	1	1	11	0
P	0	6	0	1	0	0	0	0	0	0	0	0	1	12	7	0	0	0	0	0	0	3
S	1	17	0	8	3	1	3	0	2	2	2	0	37	1	24	5	0	29	3	0	1	3
T	5	22	3	11	1	5	0	2	2	2	0	5	16	4	2	38	0	4	1	0	4	3
W	2	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	2	10	0	0
Y	1	0	4	2	0	1	0	0	2	4	0	1	1	2	0	2	0	15	5	7	0	0
V	6	3	1	1	2	15	0	0	2	12	0	28	0	5	3	0	27	0	1	8	7	0

Sequence Profile

Probe

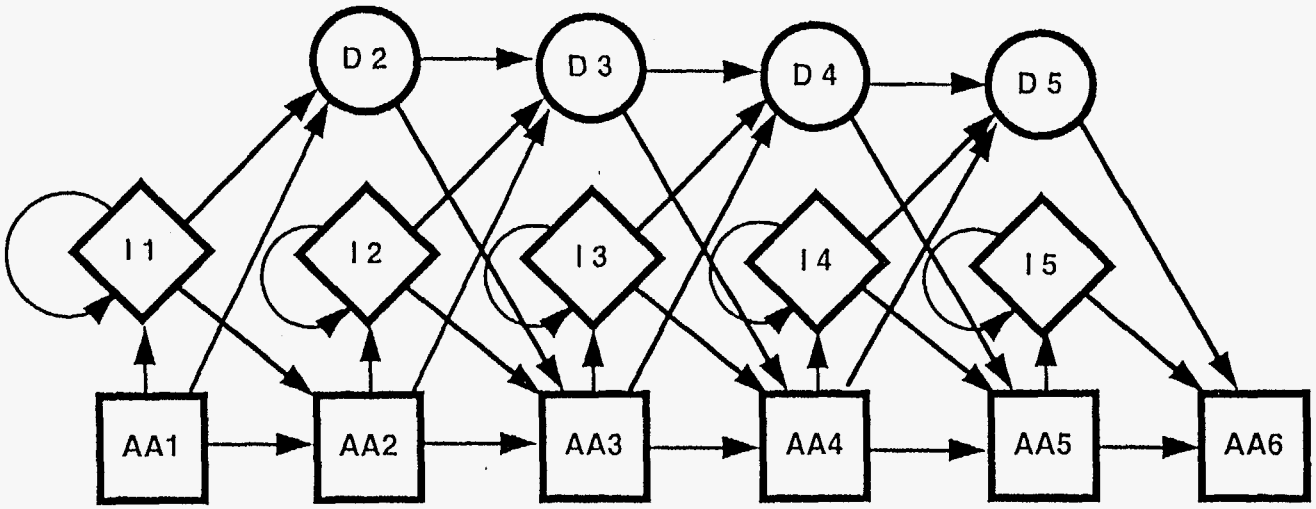
Position	8 - 37	38 - 67	68 - 98	130 - 159	160 - 188	189 - 214	216 - 246	247 - 276	Consensus
1	V	T	T	T	.	.	K	E	T
2	V	G	G	Q	.	.	E	E	E
3	Y	E	E	Q	.	Q	R	Q	Q
4	K	K	K	L	A	D	T	R	K
5	R	P	N	P	G	L	V	P	P
6	Y	F	F	Y	Y	A	Y	F	Y
7	I	P	T	E	P	V	L	V	V
8	C	C	C	C	C	C	C	C	C
9	S	K	D	P	K	D	P	E	D
10	F	E	S	H	K	.	R	H	H
11	A	E	D	E	D	.	D	A	D
12	D	G	G	G	D	.	G	G	G
13	.	.	.	.	S	V	.	.	S
14	C	C	C	C	C	C	C	C	C
15	G	E	D	D	.	N	D	G	D
16	A	K	L	K	.	R	R	K	K
17	A	G	R	R	S	K	S	C	S
18	Y	F	F	F	F	F	Y	F	F
19	N	T	T	S	V	R	T	A	T
20	K	S	T	L	G	H	T	M	T
23	H	L	K	P	K	K	A	K	R
24	W	H	A	S	T	D	F	K	S
25	K	H	N	R	W	Y	N	S	N
26	.	.	.	.	T	.	.	.	T
27	.	.	.	.	L	.	.	.	L
28	.	.	.	.	Y	.	.	.	S
29	L	L	M	L	L	L	L	L	L
30	Q	T	K	K	K	R	R	E	K
31	A	R	K	R	H	D	S	R	R
32	H	H	H	H	V	H	H	H	H
33	.	.	F	.	.	.	I	.	I
34	L	S	N	E	A	Q	Q	S	S
35	C	L	R	K	E	K	S	V	K
36	K	T	F	V	C	T	F	V	V
37	H	H	H	H	H	H	H	H	H
38									*

Profile

Position	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Gap Opening	Gap Extension
1	15	-3	12	13	-8	11	4	10	12	2	6	11	9	7	6	14	32	12	-22	-8	25	25
2	16	-8	22	24	-14	22	8	4	8	-1	2	15	8	16	3	15	13	9	-30	-12	25	25
3	11	-16	23	27	-16	7	22	0	15	-1	3	17	8	35	14	8	7	0	-19	-11	25	25
4	12	-16	13	13	-15	5	8	4	31	0	11	16	8	15	21	12	15	5	-8	-12	100	100
5	17	-6	9	8	-12	11	9	6	8	3	7	12	28	11	12	17	12	13	-21	-15	100	100
6	-5	20	-23	-19	57	-21	10	20	-20	29	10	-3	-24	-20	-18	-5	-5	8	38	58	100	100
7	15	-3	6	8	-2	8	4	20	3	14	14	5	17	6	4	12	16	25	-27	-8	100	100
8	32	142	-48	-49	-10	19	-7	29	-51	-73	-53	-25	9	-53	-27	78	28	30	-129	101	100	100
9	16	-13	25	22	-24	15	12	0	23	-8	1	19	21	20	16	20	15	2	-26	-18	100	100
10	6	-7	9	11	-4	2	21	3	11	2	3	13	6	13	15	11	7	2	-3	1	24	24
11	27	-15	40	37	-26	25	14	1	12	-8	-3	24	12	24	4	16	16	3	-42	-15	24	24
12	32	-3	43	30	-31	62	3	-7	6	-16	-9	26	14	17	-6	29	22	10	-53	-28	24	24
13	9	-1	5	5	0	6	4	9	5	5	6	7	5	4	5	12	8	10	-9	-1	24	24
14	32	142	-48	-49	-10	19	-7	29	-51	-73	-53	-25	9	-53	-27	78	28	30	-129	101	100	100
15	19	-13	40	31	-25	31	12	-1	11	-10	-5	27	9	21	3	18	15	4	-38	-16	28	28
16	9	-17	9	9	-15	0	11	2	35	-1	13	14	8	16	30	11	10	3	5	-14	28	28
17	15	3	9	8	-13	12	7	2	16	-7	2	13	13	8	20	27	12	4	-1	-10	100	100
18	-19	1	-46	-34	83	-33	3	41	-35	62	27	-19	-36	-36	-25	-10	-10	16	63	81	100	100
19	19	1	11	10	-9	14	4	10	11	1	5	15	12	6	8	19	33	12	-17	-9	100	100
20	13	-5	9	9	-3	10	7	9	11	7	10	11	8	8	7	15	22	10	-14	-6	100	100
23	13	-20	13	13	-20	5	9	1	42	-2	13	21	11	18	25	14	13	3	-10	-15	100	100
24	10	-7	7	6	-1	5	10	4	9	4	3	11	4	7	11	15	11	3	-6	2	100	100
25	5	-8	9	7	-2	2	16	0	14	-1	1	23	3	9	17	15	6	-1	2	3	100	100
26	8	-3	5	5	-1	5	5	6	5	4	5	8	4	5	5	9	11	7	-9	0	25	25
27	7	-6	3	4	3	2	5	8	4	9	8	6	3	5	5	7	6	8	-7	1	25	25
28	6	-1	3	3	4	2	7	6	3	5	5	7	2	4	4	7	6	6	-5	5	25	25
29	-1	-58	-32	-17	77	-28	-10	59	-17	107	92	-19	-16	-4	-19	-20	-3	61	23	18	100	100
30	7	-18	15	16	-22	3	14	-1	39	-6	9	18	9	21	33	13	14	0	0	-18	100	100
31	9	-10	12	11	-16	3	18	-1	24	-6	5	15	12	17	33	15	8	0	10	-13	100	100
32	0	-12	28	28	-7	-7	104	-13	11	-8	-15	37	16	50	35	-4	-2	-11	-15	17	100	100
33	6	-3	1	2	6	1	5	12	2	10	8	5	1	3	3	7	6	10	-6	5	30	30
34	17	-7	17	18	-11	13	12	4	11	1	4	18	11	21	8	22	11	5	-15	-10	100	100
35	9	-7	7	9	-8	5	7	8	19	2	10	11	6	9	15	15	11	10	-10	-5	100	100
36	9	2	-4	-2	10	3	1	23	1	17	13	2	1	-4	-1	10	20	23	-12	10	100	100
37	-4	-17	40	39	-12	-13	152	-27	15	-17	-27	53	21	73	50	-9	-7	-26	-14	25	100	100
38	11	19	13	14	14	12	21	2	25	16	2	7	8	7	16	13	15	11	2	9		

(After Gribskov)

Hidden Markov Model
(after Haussler)



ISMB - 1995
Intelligent Systems for Molecular Biologists
Doug Brutlag

Symbolic Pattern Matching In Biological Sequences

Abarbanel, R. M., Wieneke, P. R., Mansfield, E., Jaffe, D. A. and Brutlag, D. L. (1984). Rapid searches for complex patterns in biological molecules. *Nucleic Acids Res.* **12**, 263-280.

Aho, A. V. and Corasick, M. J. (1975). Fast pattern matching: An aid to bibliographic search. *Commun. ACM* **18**, 333-340.

Bairoch, A. (1993). The PROSITE dictionary of sites and patterns in proteins, its current status. *Nucleic Acids Res* **21** (13), 3097-103.

Gotoh, O. (1987). Pattern matching of biological sequences with limited storage. *Comput. Appl. Biosci.* **3**, 17-20.

Knuth, D. E. (1973). *The Art of Computer Programming, Volume 3. Sorting and Searching.* Reading Mass: Addison-Wesley.

Knuth, D. E., Morris, J. H. and Pratt, V. R. (1977). Fast pattern matching in strings. *SIAM J. Comput.* **6**, 323-350.

Landau, G.M., Vishkin, U. and Nussinov, R. (1986). An efficient string matching algorithm with k differences for nucleotide and amino acid sequences. *Nucleic Acids Res.* **14**, 31-46.

Nussinov, R. (1983). An efficient code searching for sequence homology and DNA duplication. *J. Theor. Biol.* **100**, 319-28.

Nussinov, R. (1983). Efficient algorithms for searching for exact repitition of nucleotide sequences. *J. Mol. Evol.* **19**, 283-5.

Saurin, W. and Marliere, P. (1987). Matching relational patterns in nucleic acid sequences. *Comput Appl Biosci*, **3** (2), 115-20.

Sibbald, P. R. and Argos, P. (1990). Scrutineer: a computer program that flexibly seeks and describes motifs and profiles in protein sequence databases [published erratum appears in Comput Appl Biosci **6**, 431]. *Comput Appl Biosci*, **6** (3), 279-88.

Smith, H. O., Annau, T. M. and Chandrasegaran, S. (1990). Finding sequence motifs in groups of functionally related proteins. *Proc Natl Acad Sci U S A*, **87** (2), 826-30.

Smith, R. (1988). A finite state machine algorithm for finding restriction sites and other pattern matching applications. *Comput Appl Biosci*, **4** (4), 459-65.

Smith, R. F. and Smith, T. F. (1992). Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Eng* 5 (1), 35-41.

Staden, R. (1991). Screening protein and nucleic acid sequences against libraries of patterns. *Dna Seq*, 1 (6), 369-74.

Sternberg, M. J. (1991). PROMOT: a FORTRAN program to scan protein sequences against a library of known motifs. *Comput Appl Biosci*, 7 (2), 257-60.

13

CONSENSUS PATTERNS

Active site of trypsin-like serine proteases

G D S G G

Zinc Finger (C2H2 type)

C X{2,4} C X{12} H X{3,5} H

N-Glycosylation Site

N [^P] [S T] [^P]

Homeobox Domain Signature

[LIVMF] X{5} [LIVM] X{4} [IV] [RKQ] X W X{8} [RK]

BRUTE FORCE STRING SEARCH

```
A STRING SEARCHING EXAMPLE CONSISTING OF ...
STING
 STING
  STING
   STING
    STING
     STING
      STING
       STING
        STING
         STING
          STING
           STING
            STING
             STING
              STING
               STING
                STING
                 STING
                  STING
                   STING
                    STING
                     STING
                      STING
                       STING
                        STING
                         STING
                          STING
                           STING
A STRING SEARCHING EXAMPLE CONSISTING OF ...
```

WORST CASE BRUTE FORCE STRING SEARCH

```
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA . . .
AAAAT
 AAAAT
  AAAAT
   AAAAT
    AAAAT
     AAAAT
      AAAAT
       AAAAT
        AAAAT
         AAAAT
          AAAAT
           AAAAT
            AAAAT
             AAAAT
              AAAAT
               AAAAT
                AAAAT
                 AAAAT
                  AAAAT
                   AAAAT
                    AAAAT
                     AAAAT
                      AAAAT
                       AAAAT
                        AAAAT
                         AAAAT
                          AAAAT
                           AAAAT
                            AAAAT
                             AAAAT
                              AAAAT
                               AAAAT
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA . . .
```

BOYER-MOORE STRING SEARCH

```
A STRING SEARCHING EXAMPLE CONSISTING OF ...
STING
      STING
          STING
                STING
                      STING
                            STING
                                  STING
                                      STING
A STRING SEARCHING EXAMPLE CONSISTING OF ...
```

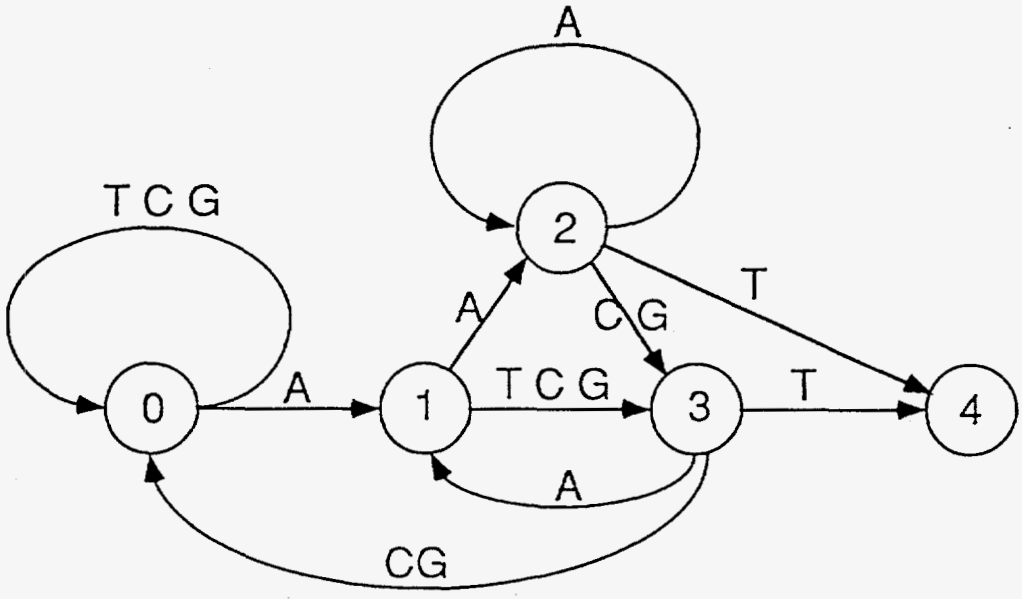
Finite State Machine for Pattern Searching

Pattern = A . T = [AAT or ACT or AGT or ATT]

Character	State
	0
A	1
AA	2
AC	3
AT	3
AG	3
AAA	2
AAT	4
ACT	4
AGT	4
ATT	4

	Original State			
	0	1	2	3
A	1	2	2	1
T	0	3	4	4
C	0	3	3	0
G	0	3	3	0

Finite State Automaton
To Find "A . T"



Codons

Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile
GCA GCC GCG GCT	CGA CGC CGG CGT AGA AGG	GAC GAT	AAC AAT	TGC TGT	GAA GAG	CAA CAG	GGA GGC GGG GGT	CAC CAT	ATA ATC ATT
Leu	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**
CTA CTC CTG CTT TTA TTG	AAA AAG	ATG	TTC TTT	CCA CCC CCG CCT	TCA TCC TCG TCT AGC AGT	ACA ACC ACG ACT	TGG	TAC TAT	GTA GTC GTG GTT

IUPAC Ambiguity Code

The IUPAC Code

A,C,G,T,U

R	A or G
Y	C or T or U
M	C or A
K	T or U or G
W	T or U or A
S	C or G
B	Not A
D	Not C
H	Not G
V	Not T or U
N	Either C, T, A, G, or U

Complements

A	T or U
B	V
C	G
D	H
G	C
H	D
K	M
M	K
N	N
R	Y
S	S
T	A
U	A
V	B
W	W
Y	R

Nomenclature Committee of IUB (NC-IUB) and IUPAC Joint Commission on Biochemical Nomenclature (JCBN) Codes for ambiguities in nucleotide sequences. (1985). Eur. J. Biochem. 146: 237-239.

ISMB - 1995
Intelligent Systems for Molecular Biologists
Doug Brutlag

Probabilistic Pattern Matching

Baldi, P., Chauvin, T., Hunkapillar, M. and McClure, A. (1992). Hidden Markov Models in Molecular Biology: New Algorithms and Applications. in Proceedings of the 1992 Neural Information Processing Systems. Denver Colorado. Eds. Morgan Kauffman , San Mateo CA., pp.

Baldi, P. and Chauvin, Y. (1994). Hidden Markov Models of G-protein-coupled receptor family. *J. Computational Biology* 1 (4), 311-336.

Bowie, J. U., Luthy, R. and Eisenberg, D. (1991). A Method to Identify Protein Sequences That Fold Into a Known Three-Dimensional Structure. *Science* 253 (164-170),

Brennan, R. G. and Matthews, B. W. (1989a). The helix-turn-helix DNA binding motif. *J Biol Chem,* 264 (4), 1903-6.

Brennan, R. G. and Matthews, B. W. (1989b). Structural basis of DNA-protein recognition. *Trends Biochem Sci,* 14 (7), 286-90.

Dodd, I. B. and Egan, J. B. (1990). Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. *Nucleic Acids Res* 18 (17), 5019-26.

Gribskov, M., Homyak, M., Edenfield, J. and Eisenberg, D. (1988). Profile scanning for three-dimensional structural patterns in protein sequences. *Comput Appl Biosci,* 4 (1), 61-6.

Gribskov, M., McLachlan, A. D. and Eisenberg, D. (1987). Profile analysis: Dectection of distantly related proteins. *Proc. Natl. Acad. Sci. USA,* 84, 4355-4358.

Haussler, D., Krogh, A., Mian, S. and Sjolander, K. (1992). *Protein Modeling using Hidden Markov Models* (UCSC-CRL-92-93). University of California, Santa Cruz.

Haussler, D., Krogh, A., Mian, S. and Sjolander, K. (1993). Protein Modeling using Hidden Markov Models: Analysis of Globins. in Twenty-Sixth Annual Hawaii International Conference on System Sciences: Architecture and Biotechnology Computing. Wailea, Hawaii. Eds. Mudge, T. N., Milutinovic, V. and Hunter, L. IEEE Computer Society Press, pp. 792-802.

Henikoff, S. and Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. *Nucleic Acids Res* 19 (23), 6565-72.

Luthy, R., Bowie, J. U. and Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature* 356 (6364), 83-85.

Luthy, R., McLachlan, A. D. and Eisenberg, D. (1991). Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins* 10 (3), 229-239.

Staden, R. (1988). Methods to define and locate patterns of motifs in sequences. *Comput Appl Biosci*, 4 (1), 53-60.

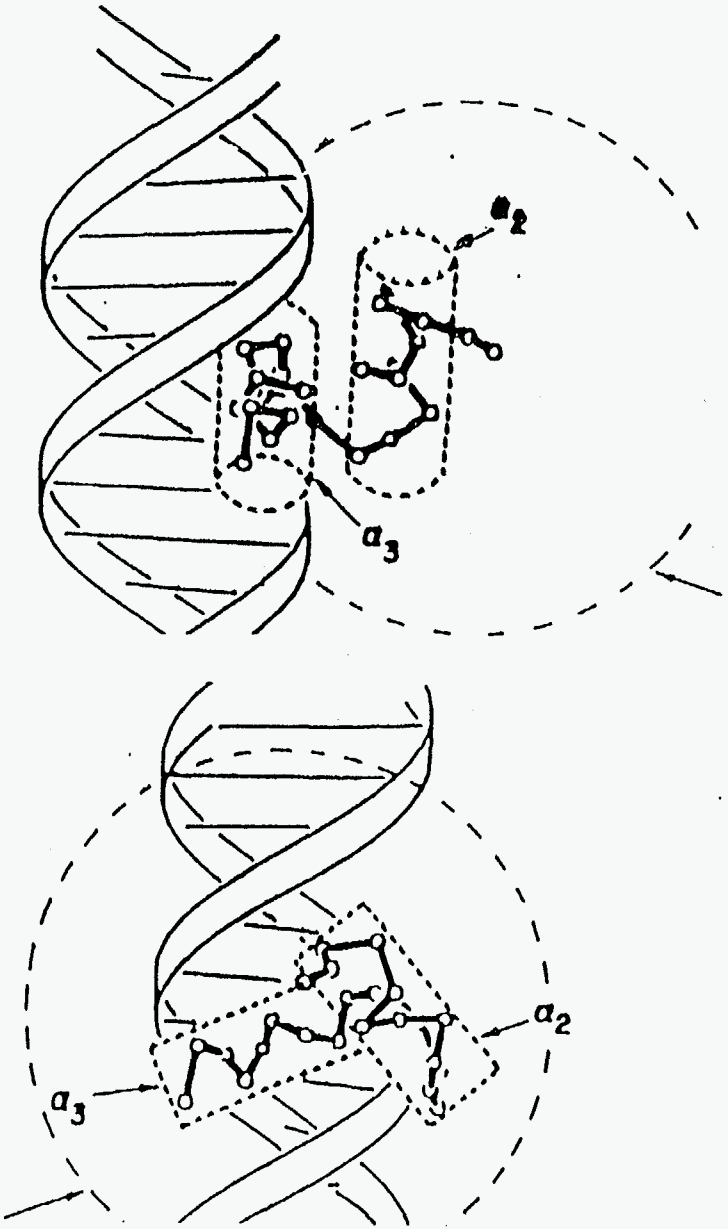
Staden, R. (1990). Searching for patterns in protein and nucleic acid sequences. *Methods Enzymol*, 183, 193-211.

Stormo, G. D. (1990). Consensus patterns in DNA. *Methods Enzymol*, 183, 211-21.

Wallace, J. C. and Henikoff, S. (1992). PATMAT: a searching and extraction program for sequence, pattern and block queries and databases. *Comput Appl Biosci*, 8 (3), 249-54.

The Helix-Turn-Helix Motif

Sequence	Helix																			
	|									|Turn|							Helix		|
RCRO$LAMBD	F	G	Q	T	K	T	A	K	D	L	G	V	Y	Q	S	A	I	N	K	A	I	H
RCRO$BP434	M	T	Q	T	E	L	A	T	K	A	G	V	K	Q	Q	S	I	Q	L	I	E	A
RCRO$BPP22	G	T	Q	R	A	V	A	K	A	L	G	I	S	D	A	A	V	S	Q	W	K	E
RPC1$LAMBD	L	S	Q	E	S	V	A	D	K	M	G	M	G	Q	S	G	V	G	A	L	F	N
RPC1$BP434	L	N	Q	A	E	L	A	Q	K	V	G	T	T	Q	Q	S	I	E	Q	L	E	N
RPC2$BPP22	I	R	Q	A	A	L	G	K	M	V	G	V	S	N	V	A	I	S	Q	W	E	R
RPC2$LAMBD	L	G	T	E	K	T	A	E	A	V	G	V	D	K	S	Q	I	S	R	W	K	R
LACR$ECOLI	V	T	L	Y	D	V	A	E	Y	A	G	V	S	Y	Q	T	V	S	R	V	V	N
CRP$ECOLI	I	T	R	Q	E	I	G	Q	I	V	G	C	S	R	E	T	V	G	R	I	L	K
TRPR$ECOLI	M	S	Q	R	E	L	K	N	E	L	G	A	G	I	A	T	I	T	R	G	S	N
RPC1$BPP22	R	G	Q	R	K	V	A	D	A	L	G	I	N	E	S	Q	I	S	R	W	K	G
GALR$ECOLI	A	T	I	K	D	V	A	R	L	A	G	V	S	V	A	T	V	S	R	V	I	N
Y77$BPT7	L	S	H	R	S	L	G	E	L	Y	G	V	S	Q	S	T	I	T	R	I	L	Q
TER3$ECOLI	L	T	T	R	K	L	A	Q	K	L	G	V	E	Q	P	T	L	Y	W	H	V	K
VIVB$BPT7	D	Y	Q	A	I	F	A	Q	Q	L	G	G	T	Q	S	A	A	S	Q	I	D	E
DEOR$ECOLI	L	H	L	K	D	A	A	A	L	L	G	V	S	E	M	T	I	R	R	D	L	N
RP43$BACSU	R	T	L	E	E	V	G	K	V	F	G	V	T	R	E	R	I	R	Q	I	E	A
Y28$BPT7	E	S	N	V	S	L	A	R	T	Y	G	V	S	Q	Q	T	I	C	D	I	R	K
IMMRE$BPPH12	S	T	L	E	A	V	A	G	A	L	G	I	Q	V	S	A	I	V	G	E	E	T
RFNR$ECOLI	M	T	R	G	D	I	G	N	Y	L	G	L	T	V	E	T	I	S	R	L	L	G
MERR$ECOLI	L	T	I	G	V	F	A	K	A	A	G	V	N	V	E	T	I	R	F	Y	N	R
IMMRE$BPPH11	L	T	Q	V	Q	L	A	E	K	A	N	L	S	R	S	Y	L	A	D	I	E	R
RP32$ECOLI	S	T	L	Q	L	E	A	D	R	Y	G	V	S	A	E	R	V	R	Q	L	E	K
LEUO$ECOLI	Q	N	I	T	R	A	A	H	V	L	G	M	S	Q	P	A	V	S	N	A	V	A
LYSR$ECOLI	G	S	L	T	E	A	A	H	L	L	H	T	S	Q	P	T	V	S	R	E	L	A
AMPR$ECOLI	L	S	F	T	H	A	A	I	E	L	N	V	T	H	S	A	I	S	Q	H	V	K
ANTP	M	P	Q	A	Q	T	N	G	Q	L	G	V	P	Q	Q	Q	Q	Q	Q	Q	Q	Q
VNU1$LAMBD	V	N	K	K	Q	L	A	D	I	F	G	A	S	I	R	T	I	Q	N	W	Q	E
VPB$BPMU	T	T	F	K	Q	I	A	L	E	S	G	L	S	T	G	T	I	S	S	F	I	N
DNAB$ECOLI	R	S	L	K	A	L	A	K	E	L	N	V	P	V	V	A	L	S	Q	L	N	R
BIRA$ECOLI	H	S	G	E	Q	L	G	E	T	L	G	M	S	R	A	A	I	N	K	H	I	Q
BPT7	K	Y	Q	E	D	L	A	A	L	E	G	T	S	D	R	I	I	S	D	L	R	S
DBH$RHIME	E	L	V	A	A	V	A	D	K	A	G	L	S	K	A	D	A	S	S	A	V	D
CYSB$ECOLI	L	N	V	S	S	T	A	E	G	L	Y	T	S	Q	P	G	I	S	K	Q	V	R
CYTR$ECOLI	A	M	I	K	D	V	A	L	K	A	K	V	S	T	A	T	V	S	R	A	L	M
MTA1$YEAST	K	E	K	E	E	V	A	K	K	C	G	I	T	P	L	Q	V	R	V	W	V	C
[illegible]	[illegible]																					



HTH Weight Matrix

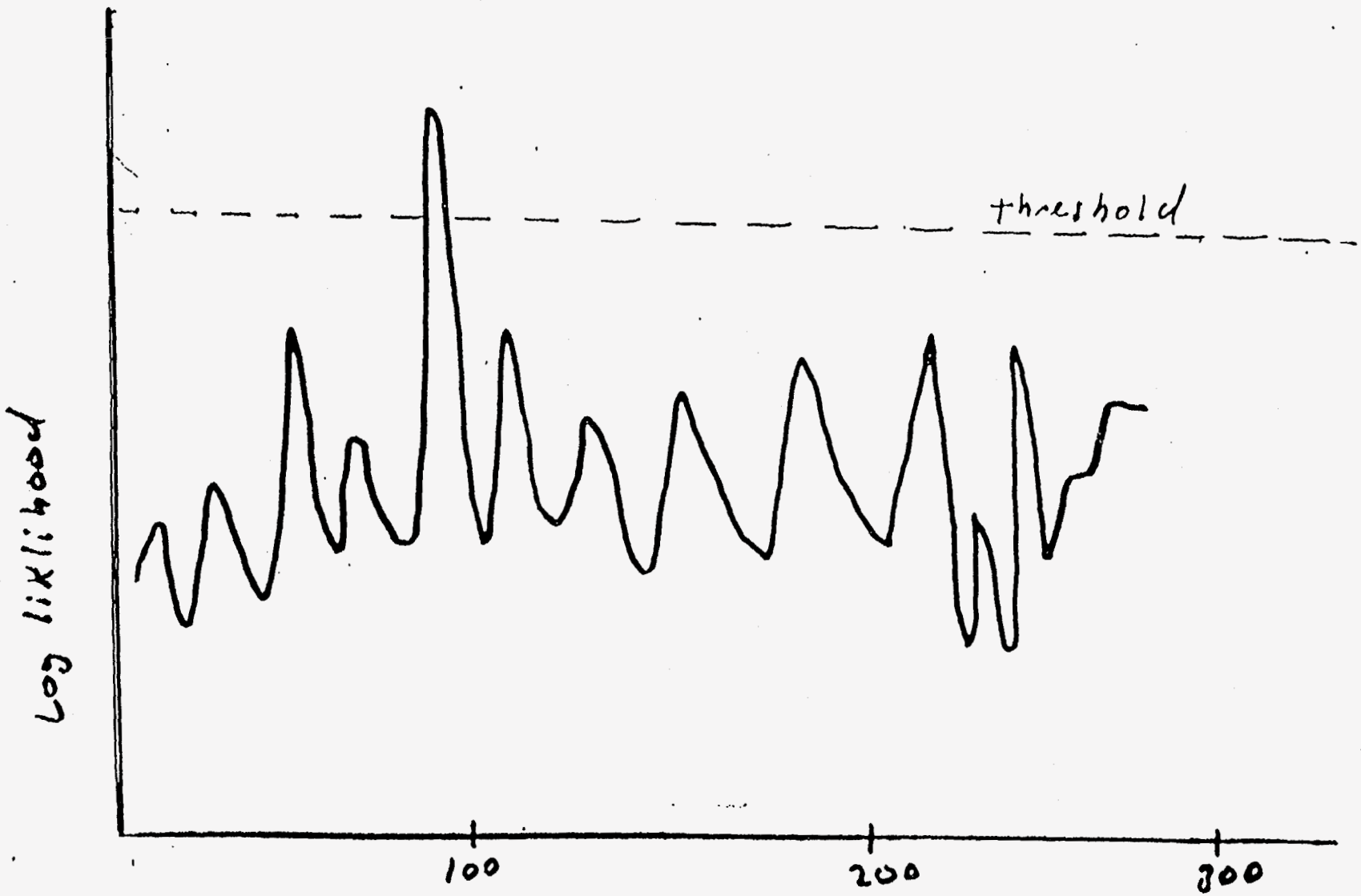
Position

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
A	2	0	0	5	5	4	29	2	5	7	0	2	0	1	6	9	2	1	1	4	0	4
R	3	1	3	5	1	0	0	2	1	0	0	0	0	4	2	2	0	5	11	0	2	6
N	0	4	1	0	0	0	1	2	0	0	3	0	2	1	0	0	0	2	2	0	2	7
D	1	0	0	0	6	0	0	6	1	0	0	0	1	2	0	1	0	0	3	1	1	1
C	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	1
Q	1	0	12	3	6	0	0	4	2	0	0	0	1	12	5	4	1	3	9	2	2	3
E	2	1	0	7	7	1	0	6	4	1	0	0	1	2	5	0	0	1	0	2	8	3
G	2	3	1	2	0	0	6	2	1	0	31	1	2	0	2	2	0	2	1	1	0	2
H	1	1	1	0	1	0	0	2	0	0	1	0	0	1	0	0	0	0	0	3	0	1
I	2	0	4	0	1	3	0	1	2	0	0	4	0	2	0	1	20	0	0	7	4	0
L	11	1	7	0	1	13	0	2	6	16	0	4	0	0	1	0	4	0	1	6	6	0
K	2	0	2	6	4	0	1	7	8	0	1	0	1	2	0	0	0	0	3	0	3	5
M	4	1	0	0	0	0	0	0	1	1	0	3	0	0	1	0	0	0	0	0	0	1
F	1	0	2	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0	1	1	1	0
P	0	1	0	0	0	0	0	0	0	0	0	0	3	1	4	0	0	0	0	0	0	0
S	2	9	0	1	4	0	0	0	0	1	0	0	19	0	9	2	0	18	2	0	1	2
T	1	13	2	5	0	4	0	1	2	1	0	4	6	2	0	15	0	2	0	0	0	1
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	6	0	0
Y	0	2	0	1	0	0	0	0	2	3	1	0	1	2	0	1	0	1	1	2	0	0
V	2	0	2	2	1	10	0	0	2	4	0	18	0	5	2	0	10	1	1	2	7	0

$$W_{ij} = \frac{\frac{N_{ij}}{N}}{f_i} \quad \text{where}$$

N_{ij} = number amino acid of type i at position j,
N = number of sequences in training set, and
f_i = frequency of amino acid of type i in database

$$\text{WM score for query of length L} = \sum_{j=1}^{L} \log W_{ij} = \sum_{j=1}^{L} \log\left(\frac{N_{ij}}{f_i}\right) - LN$$


threshold
Log likelihood
100
200
300

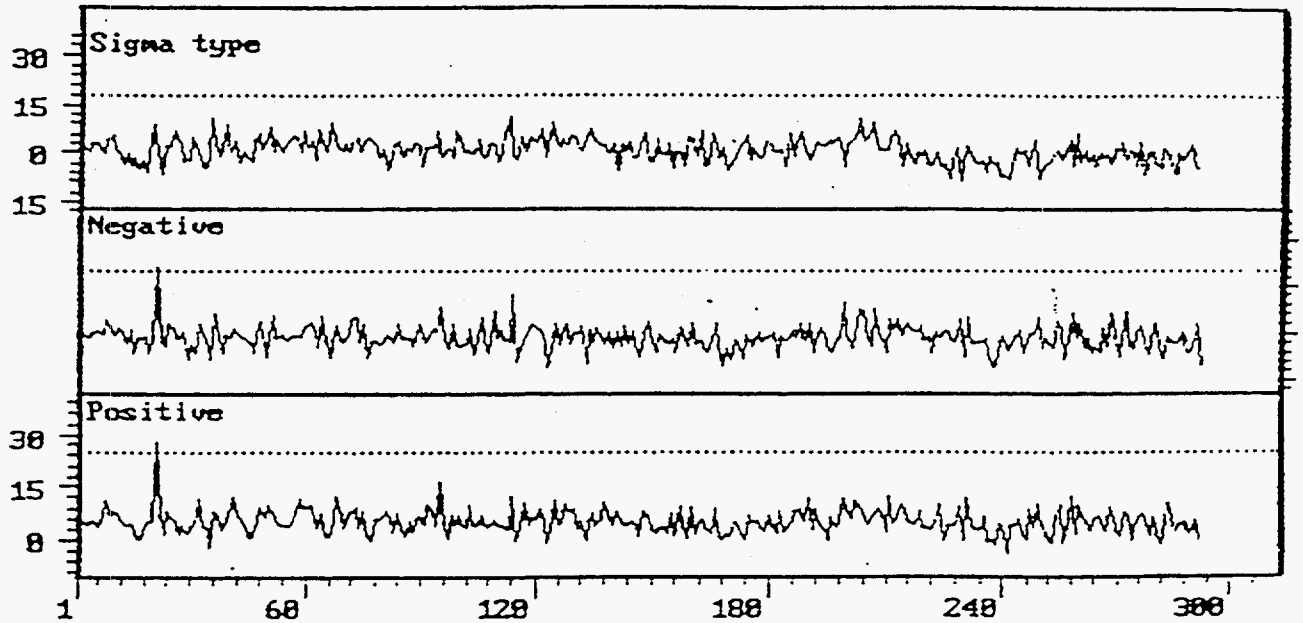
Program REGULAT.

Scan of: Positive regulator.
Negative regulator (repressor).
Sigma type regulator.

For sequence LYSR$ECOLI.

DE LYSA ACTIVATORY PROTEIN (GENE NAME: LYSR).
OS ESCHERICHIA COLI.

Total number of amino acids is: 311.



Plot of regulator(s) detection curve(s) for sequence LYSR$ECOLI.
From position 1 to 311.

Objectives:

- Identify properties of DNA sequences that determine their function, by computer-aided statistical analysis.

- Given a new sequence, accurately predict its function.

Examples:

- Regulatory regions: promoters

- Processing sites: poly-A sites, intron/exon boundaries

Related problem: predict protein structure and function from sequence.

Basic method for identifying signals:

Start with set of examples:

100 promoters
100 non-promoters

Find features that distinguish the two classes.

Use the features that distinguish the two classes to classify unknown sequences.

Two problems:

1) How can we determine weights for each piece of evidence?

2) How can we choose a good threshold to split the classes?

Discriminant analysis, weight-matrix methods, and perceptrons:

- all calculate a score for a sequence by multiplying a weight matrix times an evidence matrix.
- differ in how they determine what weights to use in the weight matrix.
- differ in how they choose a threshold.

Weight-matrix method

Weights correspond to the frequency of each base.

Procedure:

Count A,C,G, and T in each position in 100 *E. coli* promoters:

Base:	1	2	3	4	5	6
T	89	9	50	17	7	100
A	0	9	24	65	65	0
G	7	2	7	15	7	0
C	4	0	19	4	20	0

Divide count by 100 to get frequencies:

Base:	1	2	3	4	5	6
T	.89	.09	.50	.17	.07	1.0
A	0.0	.89	.24	.65	.65	0.0
G	.07	.02	.07	.15	.07	0.0
C	.04	0.0	.19	.04	.20	0.0

Recall linear algebra:

Calculate a score by matrix multiplication

W

Base:	1	2	3	4	5	6
T	.9	.1	.5	.2	.1	.8
A	.0	.8	.3	.7	.6	.0
G	.0	.0	-.1	.0	-.1	.0
C	.1	.0	-.4	.0	-.2	-.1

E

Base:	1	2	3	4	5	6
T	1	0	0	0	0	1
A	0	1	0	0	1	0
G	0	0	0	1	0	0
C	0	0	1	0	0	0

$$S = e_1 w_1 + e_2 w_2 + \ldots + e_n w_n = E \times W^T$$

$$S = .9 + .8 - .4 + 0 + .6 + .8 = 2.7.$$

Linear discriminant analysis

If the two classes you wish to separate have certain properties

- the data points are normally distributed
- the two classes differ only in their mean location

Then

- you can calculate the best dividing line analytically.

General notation:

e_i = i^{th} piece of evidence

w_i = weight for the ith piece of evidence

Score for a sequence is

$$S = e_1 w_1 + e_2 w_2 + \ldots + e_n w_n .$$

Decision rule:

If $S > T$, assign the sequence to one class.
If $S < T$, assign the sequence to the other class.
If $S = T$, can't assign, or assign arbitrarily.

Score is $S = e_1 w_1 + e_2 w_2$

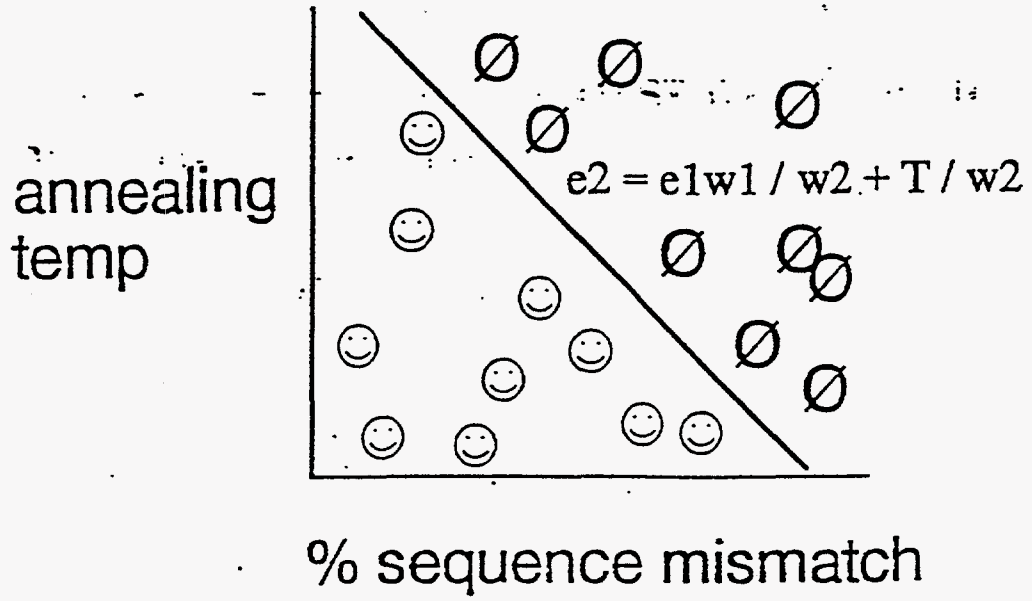
$S = T$ is the dividing line.

Let $S = T$, and re-arrange terms:

$$e_2 = \frac{w_1}{w_2} e_1 + \frac{T}{w_2}$$

This is the equation for a line that separates the two classes.

A linear partition to predict DNA hybridization



☺ hybridization

Ø no hybridization

Methods to select the threshold:

- minimize false negatives
- minimize false positives
- minimize total misclassified cases
- arbitrary

Perceptrons

Score, S, is a weighted linear function of its input variables.

Threshold, T, splits classes.

Use search to determine the values in the weight matrix.

S^+	S^-
S_1^+ : AGGCG	S_1^- : ACTCA
S_2^+ : CATCT	S_2^- : CGATT

$$W_1 = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline A & 8 & 4 & -8 & -3 & -1 \\ C & -4 & -7 & -3 & 2 & 4 \\ G & 3 & 2 & 1 & -4 & -2 \\ T & 5 & -4 & -6 & 7 & 3 \end{array}$$

$S_1^+ \cdot W_1 = 8 + 2 + 1 + 2 - 2 = 11$ OK

$S_1^- \cdot W_1 = 8 - 7 - 6 + 2 - 1 = -4$ OK

$S_2^+ \cdot W_1 = -4 + 4 - 6 + 2 + 3 = -1$ CHANGE W

$$W_1 + S_2^+ = W_2 = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline A & 8 & 5 & -8 & -3 & -1 \\ C & -3 & -7 & -3 & 3 & 4 \\ G & 3 & 2 & 1 & -4 & -2 \\ T & 5 & -4 & -5 & 7 & 4 \end{array}$$

$S_2^- \cdot W_2 = -3 + 2 - 8 + 7 + 4 = 2$ CHANGE W

$$W_2 - S_2^- = W_3 = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline A & 8 & 5 & -9 & -3 & -1 \\ C & -4 & -7 & -3 & 3 & 4 \\ G & 3 & 1 & 1 & -4 & -2 \\ T & 5 & -4 & -5 & 6 & 3 \end{array}$$

$S_1^+ \cdot W_3 = 11$ OK

$S_1^- \cdot W_3 = -2$ OK

$S_2^+ \cdot W_3 = 2$ OK

$S_2^- \cdot W_3 = -3$ OK - SUCCESS

Figure 1. We show an example of the perceptron algorithm applied to some nucleotide sequences. The sequences S^+ and S^- represent different classes. The threshold is 0. W_1 is an arbitrary starting point. The "Perceptron Convergence Theorem" guarantees that a solution will be found (if one exists) regardless of the starting W.

Nakata used linear discriminant analysis

Used four derived variables from the DNA sequence:

- Perceptron score for promoter or not, using base sequence
- Base composition
- Thermal stability (ease of separating two strand)
- DNA twist, roll, torsion (3-D structure)

90 promoters from Hawley and McClure collection, split into test and training set.

Correctly classified 75% (test set estimate)

Harr's weight matrices for -10 and -35 regions:

Base:	T	T	G	A	C	A
T	85	87	13	17	9	31
A	6	11	0	61	17	52
G	4	0	81	2	7	11
C	6	2	6	20	67	6

Base:	T	A	T	A	A	T
T	89	9	50	17	7	100
A	0	9	24	65	65	0
G	7	2	7	15	7	0
C	4	0	19	4	20	0

Apparent accuracy rate of 87%

But:

- resubstitution estimate (same sequences to build and test)
- used 48 parameters for 54 sequences

Abremski used a non-linear neural net

128 of 288 promoters from Harley and Reynolds

- only strong promoters
- not requiring special sigma factors
- not heat-shock promoters
- not in any other way irregular

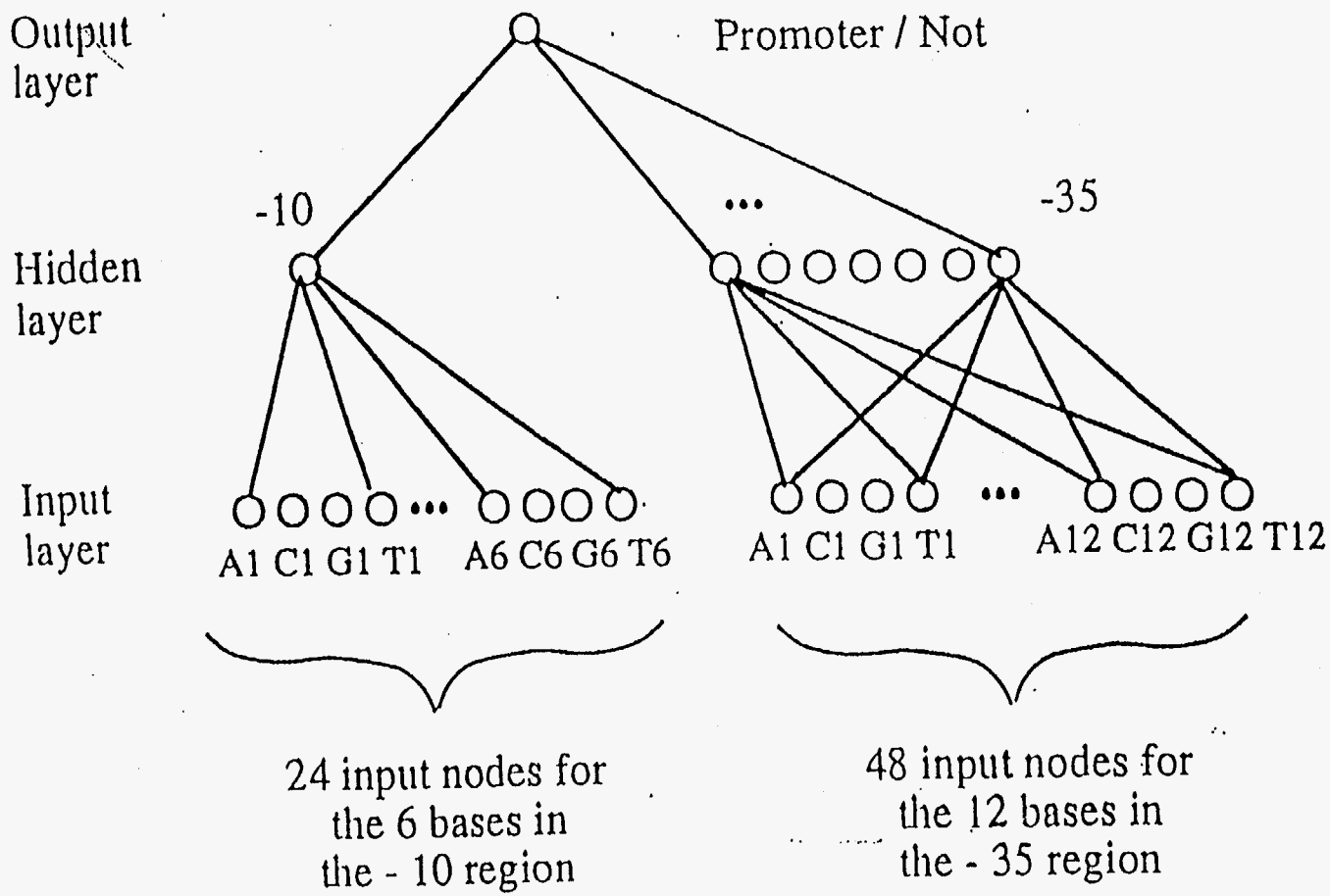
Non-promoter training sample: 515 sequences from phage T7 DNA believed to contain no promoters.

Features input to neural net:

bases in -10 and -35 regions

spacing between the two regions

Abremski's neural net for promoters



Output layer:

1 output node for promoter or not-promoter,

Input layer:

24 input nodes for the -10 region

48 input nodes for the -35 region

Hidden layer:

1 hidden node for the - 10 region

7 hidden nodes, each corresponding to the -35 region in one of the seven possible spacings from the - 10 region, i.e., spacing = {15,16,17,18,19,20,21}

Abremski's results

Correctly classified 100% of the training sequences

Correctly classified 98%, cross-validation

Required excluding all but the strong, regular promoters.

The problem of prior probability

Classification programs usually assume all classes are equally likely.

Example:

- equal prior probability of promoter or not promoter.

This assumption is usually wrong:

E. coli:

2000 promoter sites

4 million bases

1 promoter in 2000 bases =>

prior probability of promoter = 0.0005

What is the probability that the sequence actually is a promoter, given that the classifier says it is?

Bayes' rule:

$$P(\text{prom} \mid +\text{ve}) = \frac{P(+\text{ve} \mid \text{prom})P(\text{prom})}{P(+\text{ve})}$$

$$P(+\text{ve}) = P(+\text{ve} \mid \text{prom})P(\text{prom}) + P(+\text{ve} \mid \text{not prom})P(\text{not prom})$$

P(prom) = 0.0005

P(not prom) = 0.9995

P(+ve | prom) = 0.95

P(+ve | not prom) = 0.05

The probability that the sequence actually is a promoter, given that the classifier says it is = 0.01:

$$\mathbf{P(prom \mid +)} = \frac{0.95 * 0.0005}{0.95 * 0.0005 + 0.05 * 0.9995}$$

$$= 0.01$$

In 99 cases out of a 100, a sequence that our classifiers say is a promoter is not, in fact, a promoter.

How to deal with prior probability?

Gather additional evidence.

Find an ORF & look upstream from the proposed gene:

Prior probability of a promoter = 0.5

Probability that the sequence actually is a promoter, given that the classifier says it is:

$$P(\text{prom} \mid +\text{ve}) = \frac{0.95 * 0.5}{0.95 * 0.5 + 0.05 * 0.5}$$

$$= 0.95.$$

We must consider the prior probability.

Improper measures of accuracy

Watch out for people who use the same data to build the classifier that they use to test its predictive accuracy.

Use different data for training and test sets

Use several different training and test sets.

- Repeatedly divide the data into subsets with 90% for training and 10% for testing. Take the average accuracy.

As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality.

-- Albert Einstein

ISMB - 1995
Intelligent Systems for Molecular Biologists
Doug Brutlag

Alignment of Biological Sequences

Allison, L., Wallace, C. S. and Yee, C. N. (1992). Finite-state models in the alignment of macromolecules. *J Mol Evol*, 35 (1), 77-89.

Dayhoff, M. Schwartz, R. M. and Orcutt, B. C. (1978). A model of evolutionary change in Proteins. *Atlas of Protein Structure 1978*, 345-352

Dayhoff, M. O., Barker, W. C. and Hunt, L. T. (1983). Establishing Homologies in Protein Sequences, in *Methods in Enzymology*, 91, 524-545.

DeLisi, C. and Kanehisa, M. (1984). Assessing the Significance of Local Sequence Homologies. *Mathematical Biosciences* 69, 77-85.

Doolittle, R. and Fairchild. (1981). Similar amino acid sequences: chance or common ancestry? *Science* 214, 149-158.

Doolittle, R. F. (1986). *Of Urfs and Orfs: A Primer on How to Analyze Derived Amino Acid Sequences*. Mill Valley, California: University Science Books.

Feng, D.F., Johnson, M.S. and Doolittle, R.F. (1985). Aligning amino acid sequences: comparison of commonly used methods. *J. Mol. Evol.* 21, 112-125.

Gray, N. (1990). A program to find regions of similarity between homologous protein sequences using dot-matrix analysis. *J Mol Graph*, 8 (1), 11-5, 25.

Haussler, D., Krogh, A., Mian, S. and Sjolander, K. (1993). Protein Modeling using Hidden Markov Models: Analysis of Globins. in Twenty-Sixth Annual Hawaii International Conference on System Sciences: Architecture and Biotechnology Computing. Wailea, Hawaii. Eds. Mudge, T. N., Milutinovic, V. and Hunter, L. IEEE Computer Society Press, pp. 792-802.

Huang, X. Q., Hardison, R. C. and Miller, W. (1990). A space-efficient algorithm for local similarities. *Comput Appl Biosci* 6 (4), 373-81.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, 8 (3), 275-82.

Jurka, J. and Milosavljevic, A. (1991). Reconstruction and analysis of human Alu genes. *J Mol Evol*, 32 (2), 105-21.

5.

Lawrence, C. E. and Reilly, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7 (1), 41-51.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443-453.

Pearson, W. R. and Miller, W. (1992). Dynamic programming algorithms for biological sequence comparison. *Methods Enzymol*, 210, 575-601.

Pearson, W. R. (1991). Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, 11 (3), 635-50.

Rechid, R., Vingron, M. and Argos, P. (1989). A new interactive protein sequence alignment program and comparison of its results with widely used algorithms. *Comput Appl Biosci*, 5 (2), 107-13.

Reeck, G. R., de Haen, C., Teller, D. C., Doolittle, R. F., Fitch, W. M., Dickerson, R. E (1987). "Homology" in Proteins andNucleic Acids: A Terminology Muddle and a Way out of It. *Cell* 50, 667.

Schwartz, R.M. and Dayhoff, M. O. (1978). Matrices for Detecting Distant Relationships. *Atlas of Protein Structure 1978*, 353-358.

Smith, T. F. and Waterman, M. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195-197.

Smith, T., Waterman, M. and Fitch, W. (1981). Comparative biosequence metrics. *J. Mol. Evol.* 18, 38-46.

Streletc, V. B., Shindyalov, I. N., Kolchanov, N. A. and Milanesi, L. (1992). Fast, statistically based alignment of amino acid sequences on the base of diagonal fragments of DOT-matrices. *Comput Appl Biosci*, 8 (6), 529-34.

Waterman, M. S., Eggert, M. and Lander, E. (1992). Parametric sequence comparisons. *Proc Natl Acad Sci U S A*, 89 (13), 6090-3.

Sequence Alignment Problem

```
T C A T G
 / / / |
C A T T G

T C A T G
 / /  | |
C A T T G
```

Sequence Alignment

```
     X         220        230        240        250            X
     F--SGGNTHIYMNHVEQCKEILRREPKELCELVISGLPYKFRYLSTKE-QLK-Y
     |  : |::|||:||:|  |  | |||: : :| |     |   ::::: |:: |
LKPGDFIHTLGDAHIYLNHIEPLKIQLQREPRPFPKLRILRKVEKIDDFKAEDFQIEGYNPHPTIK
     X          260        270        280        290           X
```

$$\text{Score} = \sum_{\text{Region Start}}^{\text{Region End}} \text{Similarity-weights} - \sum_{\text{Region Start}}^{\text{Region End}} \text{Penalties}$$

where:

Penalty = Gap-penalty + Size-of-gap x Gap-size-penalty

Needleman Wunsch Alignment Algorithm

	A	D	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
D		1											
P												1	

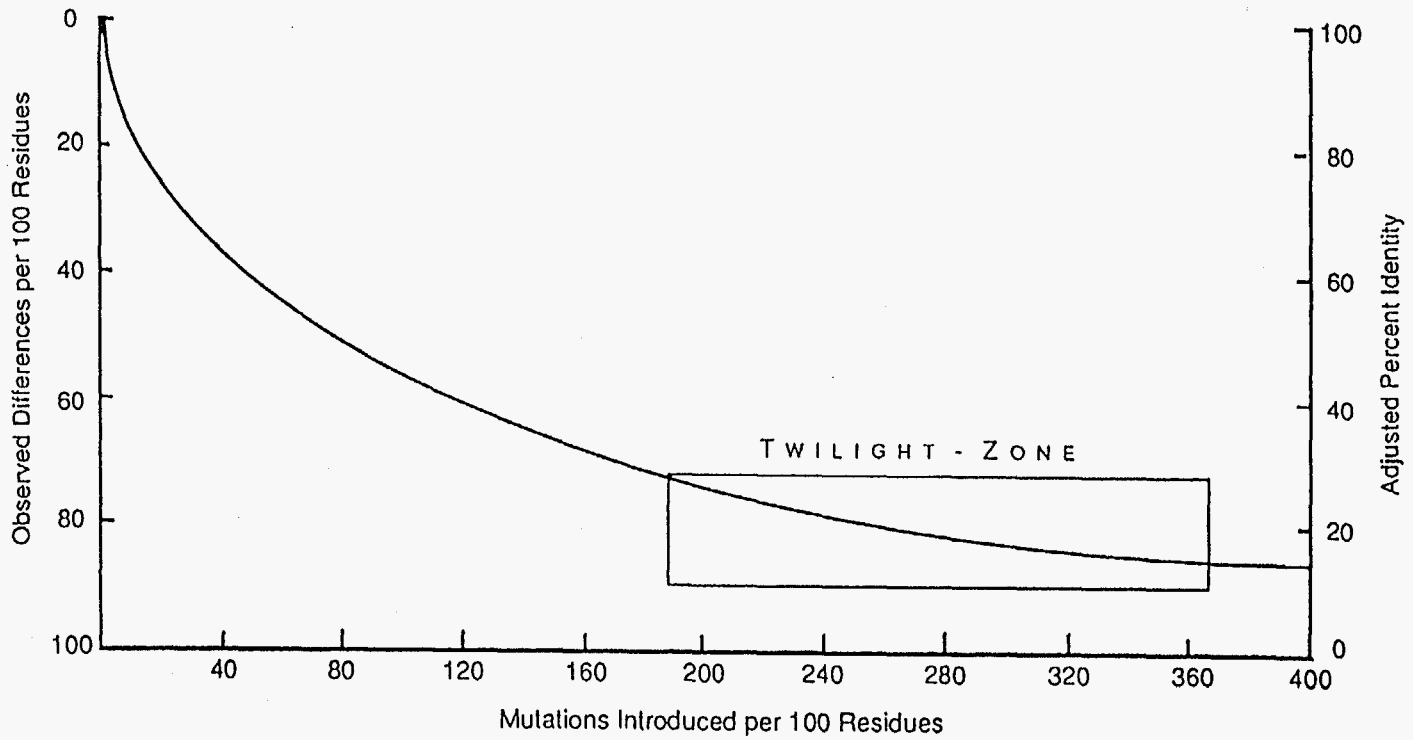
Needleman Wunsch Alignment Algorithm

	A	D	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
D	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Needleman Wunsch Alignment Algorithm

	A	D	C	N	Y	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
D	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Sequence Dissimilarity vs Evolutionary Distance



Acceptable Point Mutations
Log-Odds PAM 250 Matrix
(Adapted From Schwartz and Dayhoff)

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
A	Ala	.18																			
R	Arg	-.15	.61																		
N	Asn	.02	0	.20																	
D	Asp	.03	-.13	.21	.39																
C	Cys	-.20	-.36	-.36	-.51	1.19															
Q	Gln	-.04	.13	.08	.16	-.54	.40														
E	Glu	.03	-.11	.14	.34	-.53	.25	.38													
G	Gly	.13	-.26	.03	.06	-.34	-.53	.25	.38												
H	His	-.14	.16	.16	.07	-.34	.29	.07	-.21	.65											
I	Ile	-.05	-.20	-.18	-.24	-.23	-.20	-.20	-.26	-.24	.45										
L	Leu	-.19	-.30	-.29	-.40	-.60	-.18	-.34	-.41	-.21	.24	.59									
K	Lys	-.12	.34	.10	.01	-.54	.07	-.01	-.17	0	-.19	-.29	.47								
M	Met	-.11	-.04	-.17	-.26	-.52	-.10	-.21	-.28	-.21	.22	.37	.04	.64							
F	Phe	-.35	-.45	-.35	-.56	-.43	-.47	-.54	-.48	-.18	.10	.18	-.53	.02	.91						
P	Pro	.11	-.02	-.05	-.10	-.28	.02	-.06	-.05	-.02	-.20	-.25	-.11	-.21	-.46	.59					
S	Ser	.11	-.03	.07	.03	0	-.05	0	.11	-.08	-.14	-.28	-.02	-.16	-.32	.09	.16				
T	Thr	.12	-.09	.04	-.01	-.22	-.08	-.04	0	-.13	.01	-.17	0	-.06	-.31	.03	.13	.26			
W	Trp	-.58	.22	-.42	-.68	-.78	-.48	-.70	-.70	-.28	-.51	-.18	-.35	-.42	.04	-.56	-.25	-.52	1.73		
Y	Tyr	-.35	-.42	-.21	-.43	.03	-.40	-.43	-.52	-.01	-.09	-.09	-.44	-.24	.70	-.49	-.28	-.27	-.02	1.01	
V	Val	.02	-.25	-.17	-.21	-.19	-.19	-.18	-.14	-.22	.37	.19	-.24	.18	-.12	-.12	-.10	.03	-.62	-.25	.43

Comparison of Matrices for Scoring Sequence Similarities

Sequences Compared	Unitary Matrix	Genetic Code Matrix	Amino Acid Matrix	PAM 250 Matrix
Antibacterial substance A *Streptomyces* vs. Neocarzinostatin *Streptomyces*	3.1	3.2	2.6	2.9
Ferredoxin *Clostridium* vs Ferredoxin *Spirulina*	0.1	1.6	1.8	3.4
α–Hemoglobin Human vs. Myoglobin Human	5.8	6.6	9.9	10.7
α-Hemoglobin Human vs. Globin CTT-III Midge	2.0	2.4	3.2	3.5
Cytochrome C Horse vs. Cytochrome C_6 *Spirulina*	4.5	4.3	7.3	6.1
Cytochrome C Horse vs. Cytochrome C_{553} *Desulfovibrio*	0.2	0.4	0.4	3.9
β2-microglobulin Human vs. IG μ chain C4 region Human	3.6	3.3	4.7	4.8
Ig μ chain C4 region Human vs. Ig ε chain C4 Human	4.7	9.0	9.2	12.1

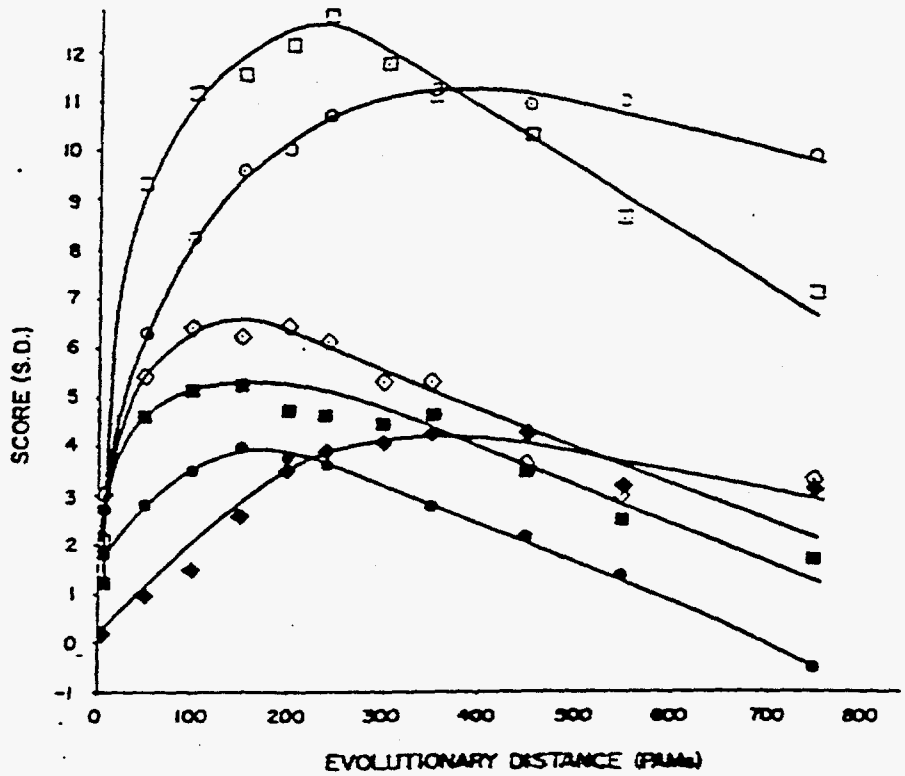


Figure 87. Alignment scores as a function of the evolutionary distance of the mutation data matrices. These log odds matrices, multiplied by 10, were calculated at 4, 50, 100, 150, 200, 242, 300, 350, 450, 550, and 750 PAMs. The gap penalty factor and matrix bias were both given values of 6 in all trials. All scores are based on 300 randomized sequence comparisons, and the standard deviations of the scores are therefore about 4% of their values. The following sequence comparisons were made: *open circle*, hemoglobin alpha chain—human vs. myoglobin—human; *solid circle*, hemoglobin alpha chain—human vs. globin CTT-III—midge larva; *open diamond*, cytochrome c—horse vs. cytochrome c_6—*Spirulina maxima; solid diamond*, cytochrome c—horse vs. cytochrome c_{553}—*Desulfovibrio gigas; open square*, Ig mu chain C4 homology region—human Gal vs. Ig epsilon chain C4 homology region—human Nd; *solid square*, Ig mu chain C4 homology region—human Gal vs. beta$_2$-microglobulin—human.

ISMB - 1995
Intelligent Systems for Molecular Biologists
Doug Brutlag

Rapid Database Search for Sequence Similarity

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). A Basic Local Alignment Search Tool. *J. Mol. Biol.*, 215, 403-410.

Altschul, S. F., Boguski, M. S., Gish, W. and Wootton, J. C. (1994). Issues in searching molecular sequence databases. *Nat Genet* 6 (2), 119-29.

Barsalou, T. and Brutlag, D. L. (1991). Searching Gene and Protein Sequence Databases. *MD Computing*, 8(3), 144-149.

Brutlag, D. L., Dautricourt, J. P., Maulik, S. and Relph, J. (1990). Improved sensitivity of biological sequence database searches. *Comput Appl Biosci*, 6(3), 237-45.

Collins, J. F., & Coulson, A. F. (1984). Applications of parallel processing algorithms for DNA sequence analysis. *Nucleic Acids Res*, 12, 181-192.

Collins, J. F., Coulson, A.F. W. and Lyall, A. (1988). The significance of protein sequence similarities. *CABIOS 4*, 67-71.

Galper, A. R. and Brutlag, D. L. (1990). *Parallel Similarity Search and Alignment with the Dynamic Programming Method* (KSL Report 90-74). Stanford University.

Gribskov, M., McLachlan, A. D. and Eisenberg, D. (1987). Profile analysis: Dectection of distantly related proteins. *Proc. Natl. Acad. Sci. USA 84*, 4355-4358.

Lipman, D.J. and Pearson, W.R. (1985). Rapid and Sensitive Protein Simlarity Searches. *Science 227*, 1435-1441.

Myers E. W. and Miller, W. (1988). Optimal alignments in linear space. *CABIOS 4*, 11-17.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci USA 85*, 2444-2448.

Pearson, W. J. (1986). Sensitivity and Selectivity in Protein Sequence Comparison. In *Methods in Protein Sequence Analysis*, Clifton, New Jersey: Humana Press.

Wilbur, W.J. and Lipman, D.J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA 80*, 726-30.

Sequence Homology Search

Query:	METR$SALTY		
Perfect Score:	1994		
Scoring parameters:	PAM 150	Gap open 20	Gap extend 6
Searched:	Swiss-prot 28	36,000 seqs	12,496,420 residues
Statistics:	Mean 44.062	Variance 67.8	

No.	Score	Match	Length	DB	ID	Description	Pred. No.
1	1994	100.0	276	4	METR_SALTY	TRANSCRIPTIONAL ACTIV	0.00e+00
2	1866	93.6	317	4	METR_ECOLI	TRANSCRIPTIONAL ACTIV	0.00e+00
3	285	14.3	299	2	CYNR_ECOLI	CYN OPERON TRANSCRIPT	8.36e-43
4	231	11.6	302	1	ALSR_BACSU	ALS OPERON REGULATORY	4.50e-30
5	216	10.8	289	1	AMPR_RHOCA	TRANSCRIPTIONAL ACTIV	1.22e-26
6	214	10.7	311	4	LYSR_ECOLI	TRANSCRIPTIONAL ACTIV	3.47e-26
7	214	10.7	300	4	NOCR_AGRT5	REGULATORY PROTEIN NO	3.47e-26
8	211	10.6	300	4	NOCR_AGRT7	REGULATORY PROTEIN NO	1.65e-25
9	208	10.4	290	1	AMPR_CITFR	TRANSCRIPTIONAL ACTIV	7.85e-25
10	208	10.4	290	1	AMPR_ENTCL	TRANSCRIPTIONAL ACTIV	7.85e-25
11	206	10.3	289	1	CATR_PSEPU	CATBC OPERON TRANSCRI	2.21e-24
12	205	10.3	297	3	ILVY_ECOLI	TRANSCRIPTIONAL ACTIV	3.70e-24
13	198	9.9	306	2	GLTC_BACSU	REGULATORY PROTEIN GL	1.35e-22
14	195	9.8	324	2	CYSB_ECOLI	CYS REGULON TRANSCRIP	6.24e-22
15	193	9.7	304	6	YAFC_ECOLI	HYPOTHETICAL 33.8 KD	1.73e-21

Sequence Homology Search

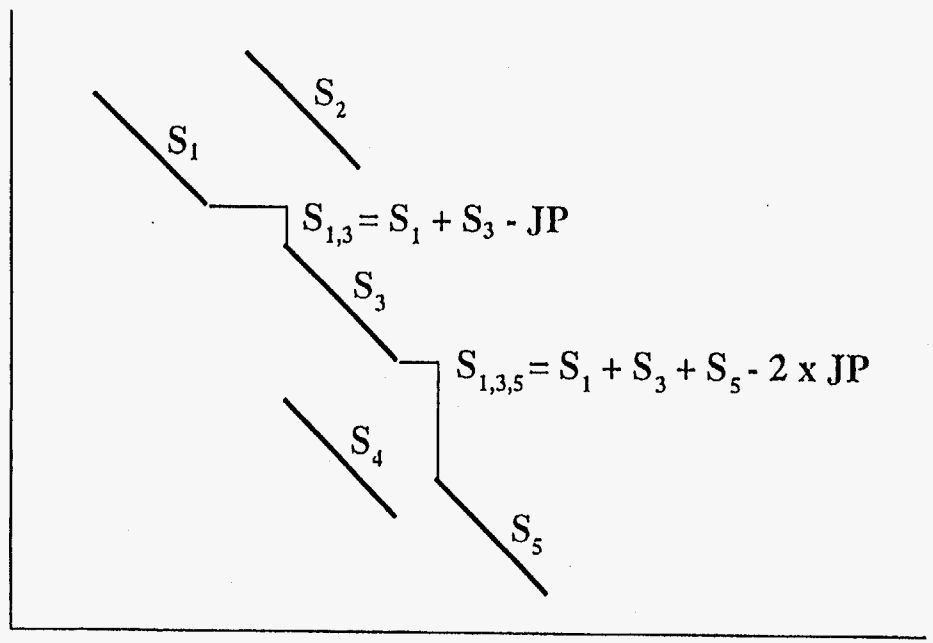
```
3. METR_SALTY (1-276)
   CYNR_ECOLI   CYNR ACTIVATORY PROTEIN.

Residue Identity =    26%  Matches            =     74  Mismatches   =   106
Gaps             =     19  Conservative Substitutions                =    76

       X   10        20        30        40        50        60        70
IEIKHLKTLQALRNSGSLAAAAAVLHQTQSALSHQFSDLEQRLGFRLFVRKSQPLRFTPQGEVLLQLANQVL
   :|:: : |: : ||:: ||: || :|:|||:|: :||: ||  || | :: :|:|  |||  | |:: |
MLSRHINYFLAVAEHGSFTRAASALHVSQPALSQQIRQLEESLGVPLFDRSGRTIRLTDAGEVWRQYASRAL
       X    10        20        30        40        50        60        70
```

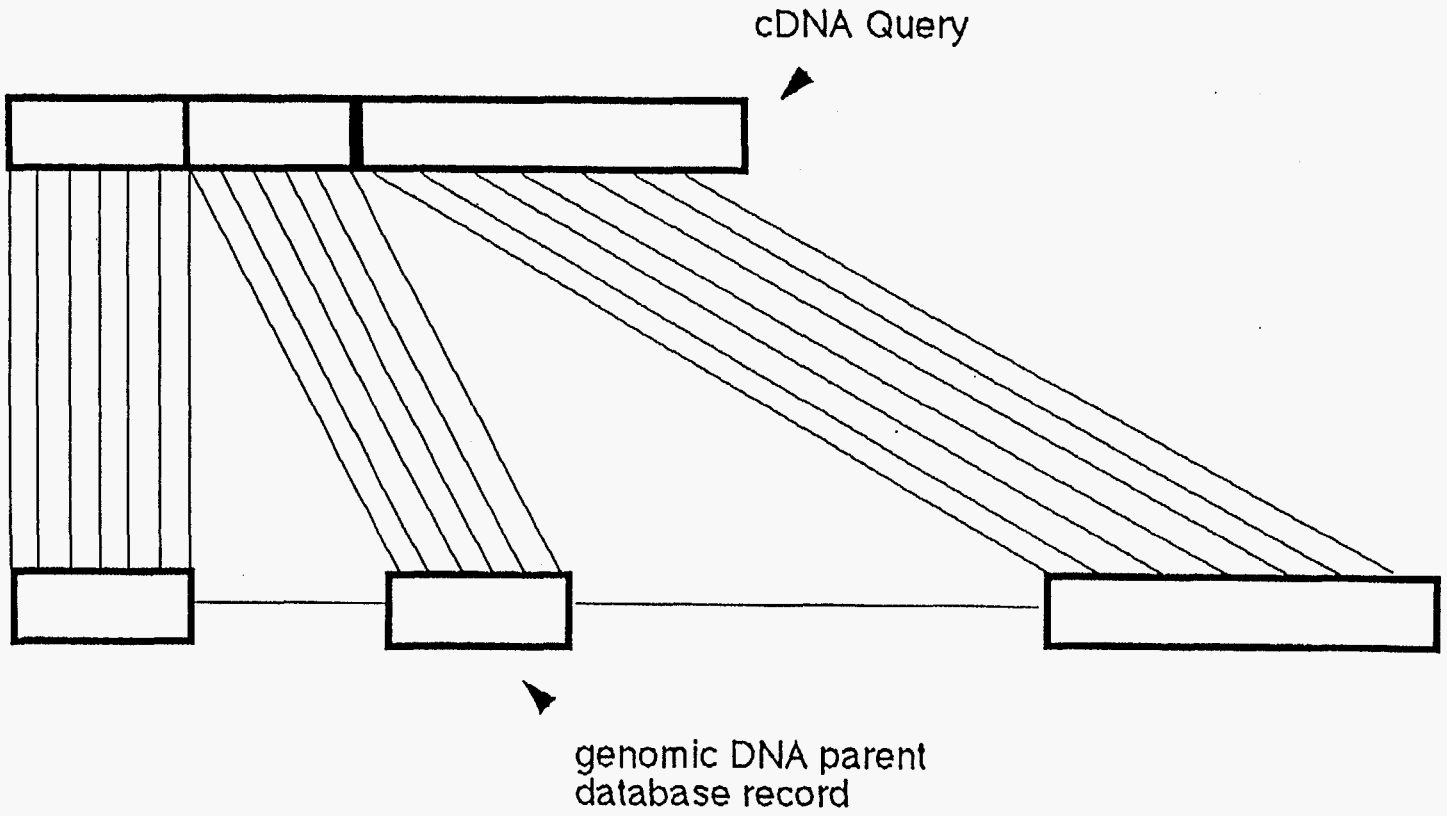
Sequence Name	Description	Length	Score	%Match	Exp
1. METR_SALTY	METR ACTIVATORY PROTEIN.	276	1356	100	0.000
2. METR_ECOLI	METR ACTIVATORY PROTEIN.	317	1285	95	0.000
3. CYNR_ECOLI	CYNR ACTIVATORY PROTEIN.	299	305	22	0.011
4. ILVY_ECOLI	ILVY ACTIVATORY PROTEIN.	297	294	22	0.022
5. AMPR_RHOCA	AMPR ACTIVATORY PROTEIN.	289	287	21	0.035

Joining Diagonals of Similarity

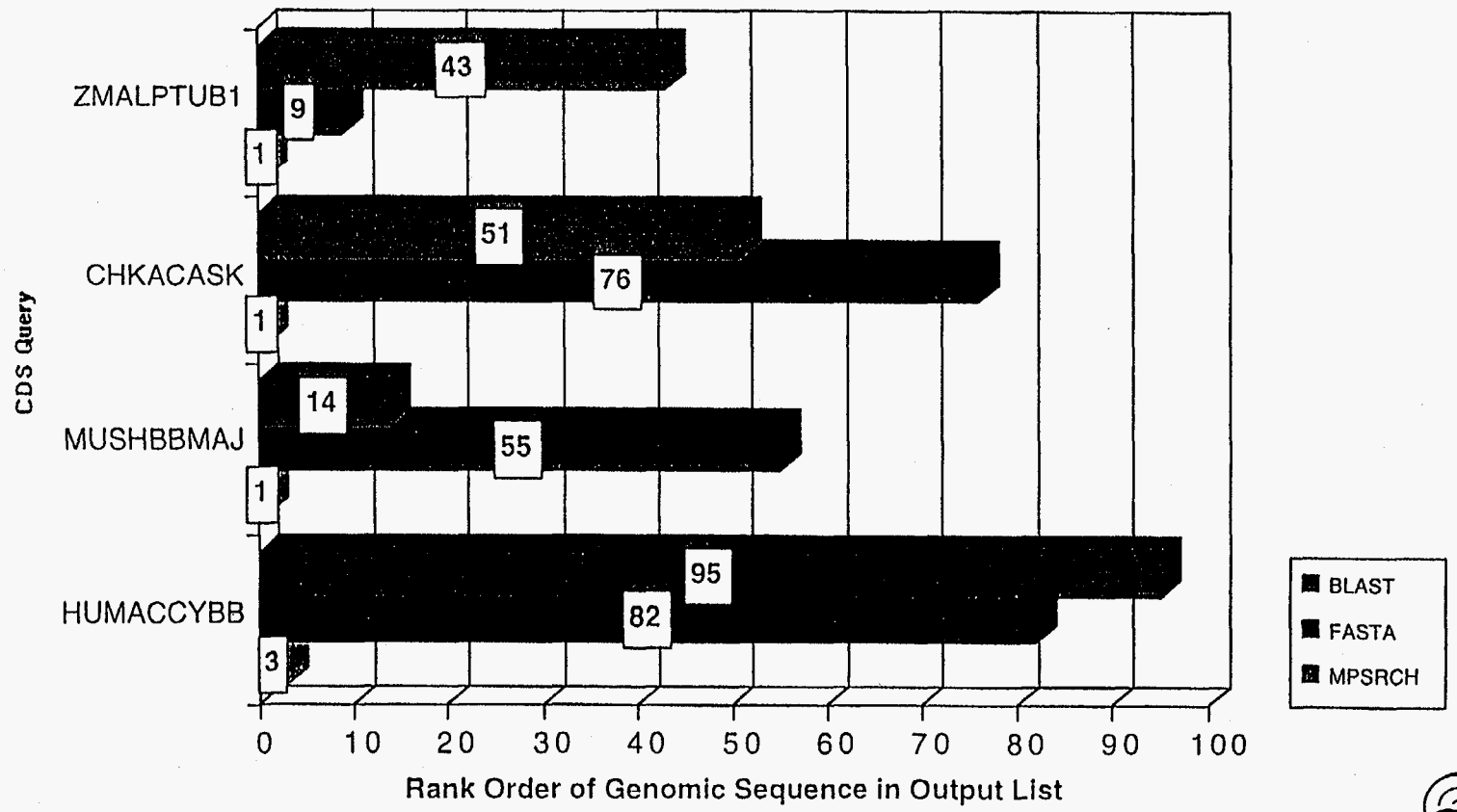


JP = Joining penalty

cDNA Queries Require Affine Gap Penalties



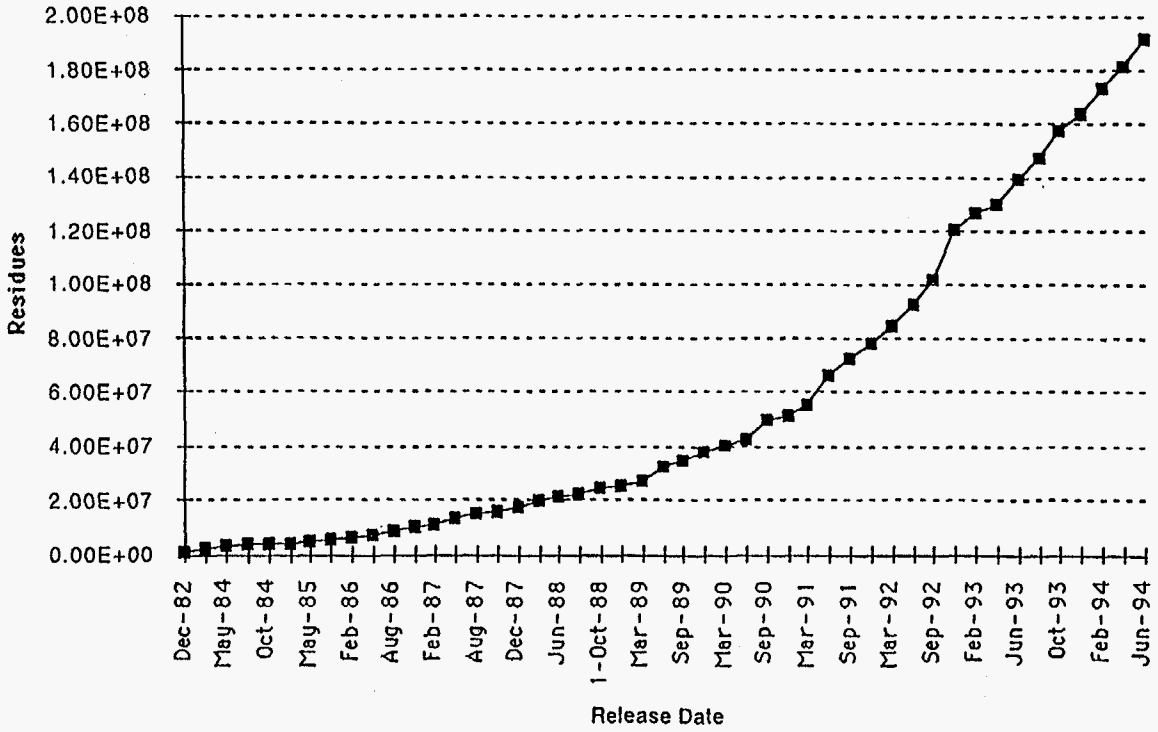
Genomic Sequences Missed by Blast/FastA



MetR Search Versus Swiss-Prot 28

Search	PAM	Gap	Gap Size	1% Expectation Number True +	1% Expectation Number False +	5% expectation Number True +	5% expectation Number False +	# Missed = False -
MPsrch_PPA	1	27	6	3	3	3	4	57
MPsrch_PPA	50	20	6	24	0	27	1	33
MPsrch_PPA	100	20	6	41	0	42	1	18
MPsrch_PPA	150	20	6	48	0	51	0	9
MPsrch_PPA	200	20	6	52	0	53	0	7
MPsrch_PPA	250	20	4	47	0	50	0	10
MPsrch_PPA	150	6	6	3	0	9	1	51
MPsrch_PPA	150	10	6	47	0	49	1	11
MPsrch_PPA	150	20	6	48	0	51	0	9
MPsrch_PPA	150	40	6	49	0	52	0	8
MPsrch_PPA	150	80	6	49	0	52	0	8
MPsrch_PPA	200	5	5	2	0	2	0	58
MPsrch_PPA	200	10	6	51	0	53	2	7
MPsrch_PPA	200	20	6	52	0	53	0	7
MPsrch_PPA	200	40	6	51	0	53	0	7
MPsrch_PPA	200	80	6	50	0	51	0	9
MPsrch_PPA	200	20	2	50	0	53	0	7
MPsrch_PPA	200	20	4	51	0	53	0	7
MPsrch_PPA	200	20	6	52	0	53	0	7
MPsrch_PPA	200	20	8	52	0	53	0	7
MPsrch_PPA	200	20	10	52	0	53	0	7
BLAST	2	∞	∞	2	0	2	0	58
BLAST	50	∞	∞	22	0	23	0	37
BLAST	100	∞	∞	30	0	32	0	28
BLAST	150	∞	∞	32	0	35	0	25
BLAST	200	∞	∞	36	0	40	0	20
BLAST	250	∞	∞	32	0	35	0	25

GenBank Growth



The Significance of Similarity Scores Decreases with Database Growth

- The score of any sequence alignment is constant
- The number of database entries grows exponentially
- The number of nonhomologous entries >> homologous entries
- Greater sensitivity is required to detect homologies

How Search Programs Handle Sequence Redundancy

Program	Search Both DNA Strands	IUPAC Code Matching	PAM matrix substitutions	Affine Gap Penalties	Genetic Code Redundancy	Rapid N-mer Search
FASTA	User Specifies	No	Hard Wired	No	6 Frame trans.	Possible
FASTDB	User Prompted	Yes	User Selected	Yes	6 Frame Trans & Genetic Code Matrix	Possible
BLAST	Automatic	No	User Provided	No	6 Frame Trans	Preferred
BLAZE	User Specifies	Yes	User Selected	Yes	No	Possible
MPSRCH	Automatic	Yes	User Selected	Yes	6 Frame Trans	Possible
QUEST	User Choice	Yes	N/A	N/A	Encoded in Patterns	Possible
PROFILE	N/A	N/A	User Provided	Yes	N/A	N/A

Comparison of Rapid Database Search Program Features

Program	Query Format	# Seqs/ Query File	Multiple Database Search	Variable PAMs	Variable Gap Penalty?	Variable Gap Size Penalty?	Output Limit-ation	Score Optimiz ation	Align-ments?	Standard Devia-tion?	Expect-ations?
FASTA	FASTA format	1	yes	120/250	Fixed	Fixed	# scores # align.	Yes	Yes	No	No
FASTDB	IG Format	N	yes	Any PAM	Variable	Variable	#scores	Yes	Yes	Yes	No
BLAST	FASTA	1	if indexed	Yes, via PAM file	No Gaps Allowed	No Gaps Allowed	V & B para-meters	No	Yes	Yes	Yes
BLAZE	FASTA or IG	1	No	Select-able	Variable	Variable	#scores % max score	Always	Yes	Yes	Yes
MPSrch	FASTA or IG	N	Yes	Any PAM	Variable	Variable	#scores % max score	Always	Yes	Yes	Yes

ISMB - 1995
Intelligent Systems for Molecular Biologists
Doug Brutlag

Near-Optimal Sequence Alignments

Altschul, S. F. and Erickson, B. W. (1986). Optimal sequence alignment using affine gap costs. *Bull Math Biol*, 48, 603-16.

Bellman, R. and Kalaba, R. (1960). On Kth best policies. *J. SIAM*, 8 (4), 582-588.

Clarke, S., Krikorian, A. and Rausen, J. (1963). Computing the N best loopless paths in a network. *J. SIAM.*, 11 (4), 1096-1102.

Gusfield, D., Balasubramanian, K. and Naor, D. (January 1992). Parametric Optimization of Sequence Alignment. Proceedings of the third annual ACM-SIAM Joint Symposium Discrete Algorithms. Orlando Florida,

Hoffman, W. and Pavley, R. (1959). A Method for the Solution of the Nth Best Path Problem. *J. ACM.*, 6, 506-514.

Naor, D. and Brutlag, D. L. (1994). On Near-Optimal Alignments of Biological Sequences. *J. Computational Biology* 1 (4), 349-366.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443-453.

Perko, A. (1986). Implementation of Algorithms for K shortest loopless paths. *Networks*, 16, 149-160.

Saqi, M. A. S. and Sternberg, M. (1991). A Simple Method to Generate Non-trivial Alternate Alignments of Protein Sequences. *J. Mol. Biol.*, 219, 727-732.

Saqi, M. A. S., Bates, P. and Sternberg, M. J. E. (1992). Towards an automatic method of predicting protein structure by homology: an evaluation of suboptimal sequence alignments. *Protein Engineering*, 5 (4), 305-311.

Schwartz, R.M. and Dayhoff, M. O. (1978). Matrices for Detecting Distant Relationships. *Atlas of Protein Structure*, 353-358.

Shier, D. R. (1979). On Algorithms for finding the K shortest paths in a Network. *Networks*, 9, 195-214.

Smith, T. F. and Waterman, M. (1981). Identification of common molecular subsequences. *J. Mol. Biol. 147*, 195-197.

6

Near-Optimal Sequence Alignments

Vingron, M. and Argos, P. (1990). Determination of reliable regions in protein sequence alignments. *Protein Engineering*, 3, 565-569.

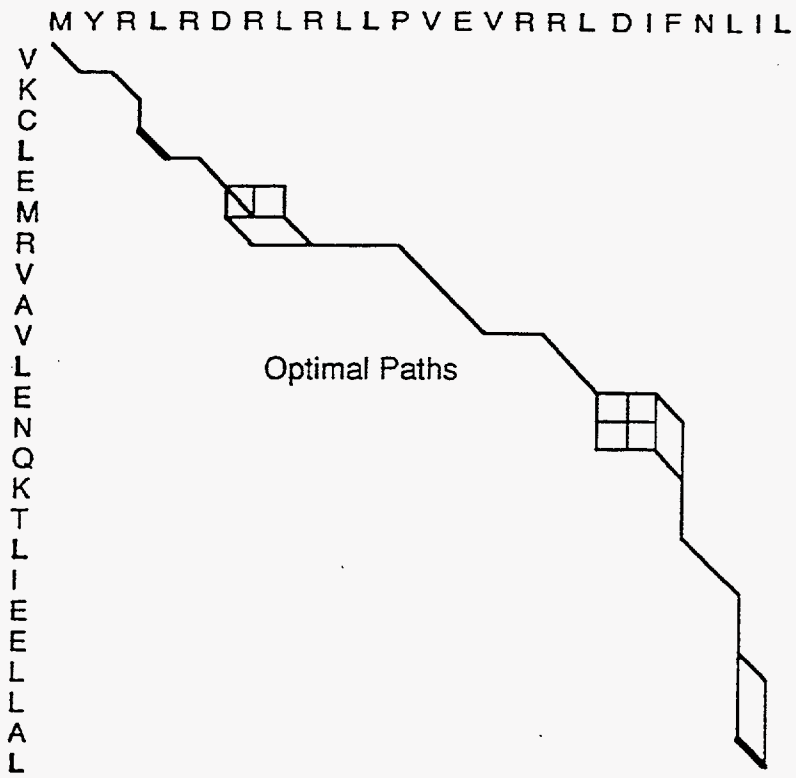
Waterman, M. S. (1983). Sequence alignments in the neighborhood of the optimum with general application to dynamic programming. *Proc. Nat. Acad. Sci.*, 80, 3123-3124.

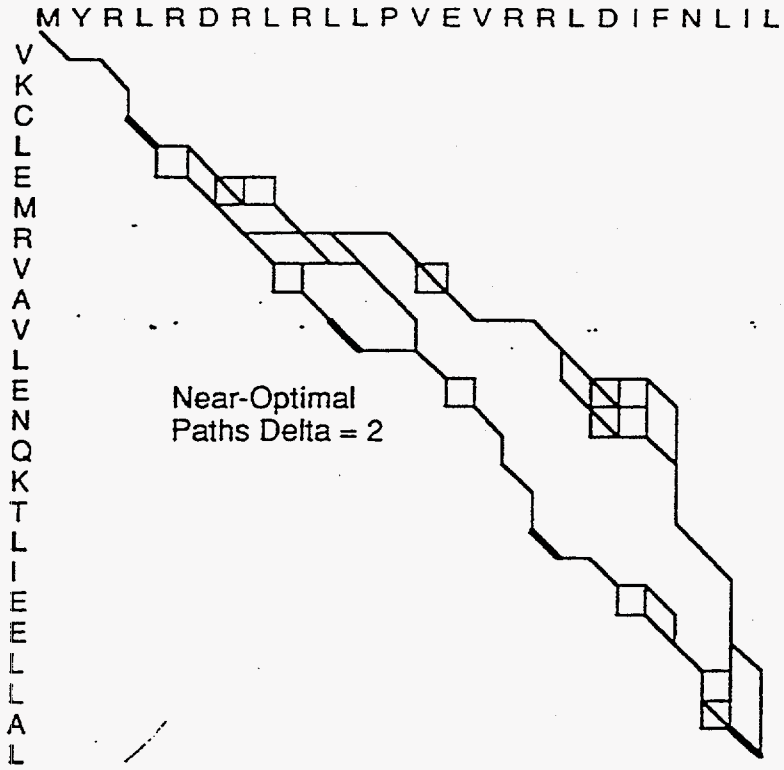
Waterman, M. S. and Byers, T. H. (1985). A dynamic programming algorithm to find all solutions in a neighborhood of the optimum. *Math. Biosci.*, 77, 179-188.

Waterman, M. S., Eggert, M. and Lander, E. (1992). Parametric sequence comparisons. *Proc Natl Acad Sci U S A*, 89(13), 6090-3.

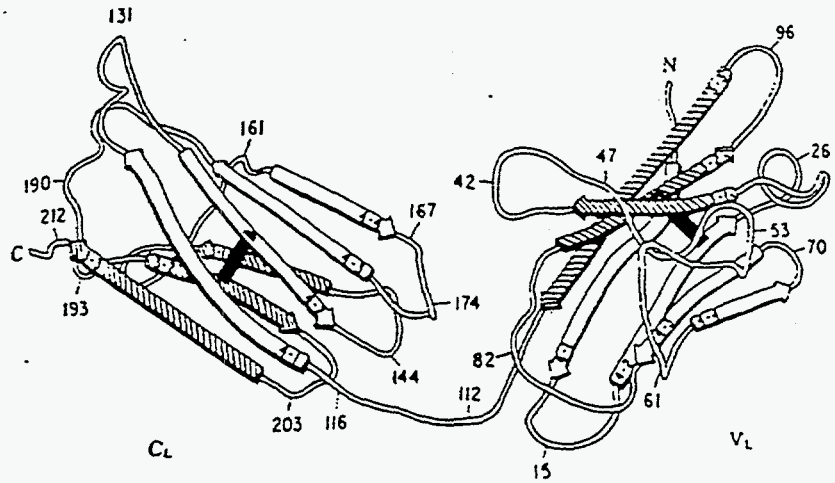
7

Optimal and Near-Optimal Alignments of Two Leucine Zippers

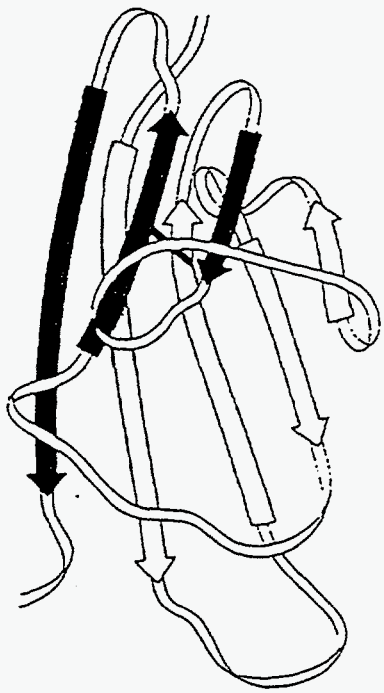




Immunoglobuylin
3-D Domain and Hypervariability

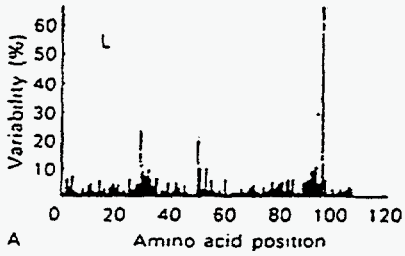


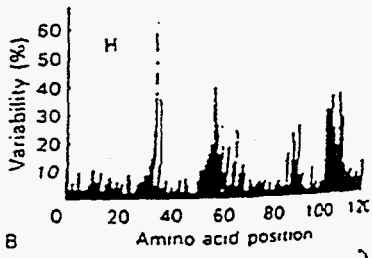
λ chain



basic structure.

Figure 35-13
Hypervariable regions in (A) light and (B) heavy chains. The degree of variability in amino acid sequence is plotted versus amino acid position. [After J. D. Capra and A. B. Edmundson. The antibody combining site. Copyright © 1976 by Scientific American, Inc. All rights reserved.]





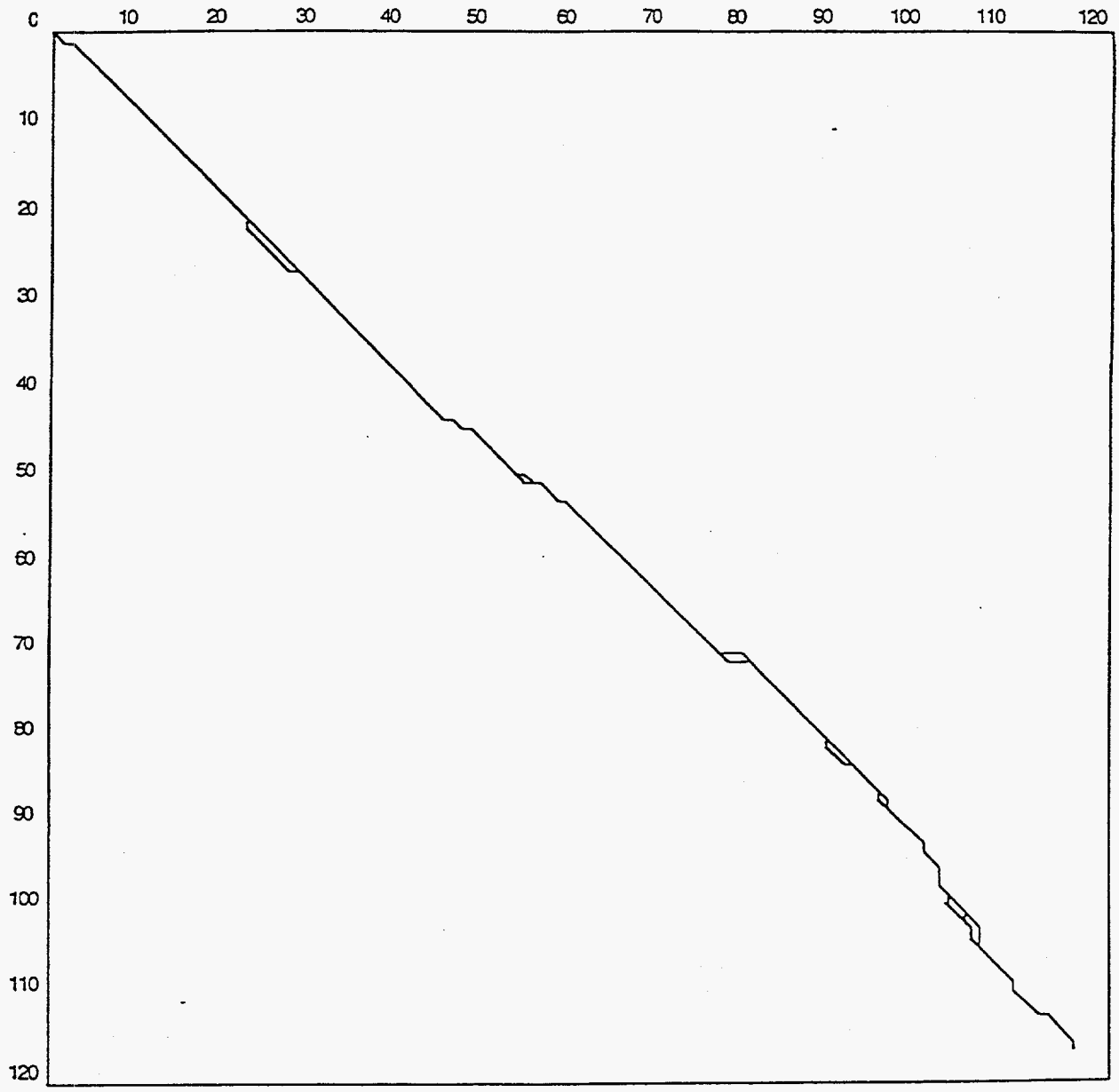

0
10
20
30
40
50
60
70
80
90
100
110
120
10
20
30
40
50
60
70
80
90
100
110
120

Figure 2

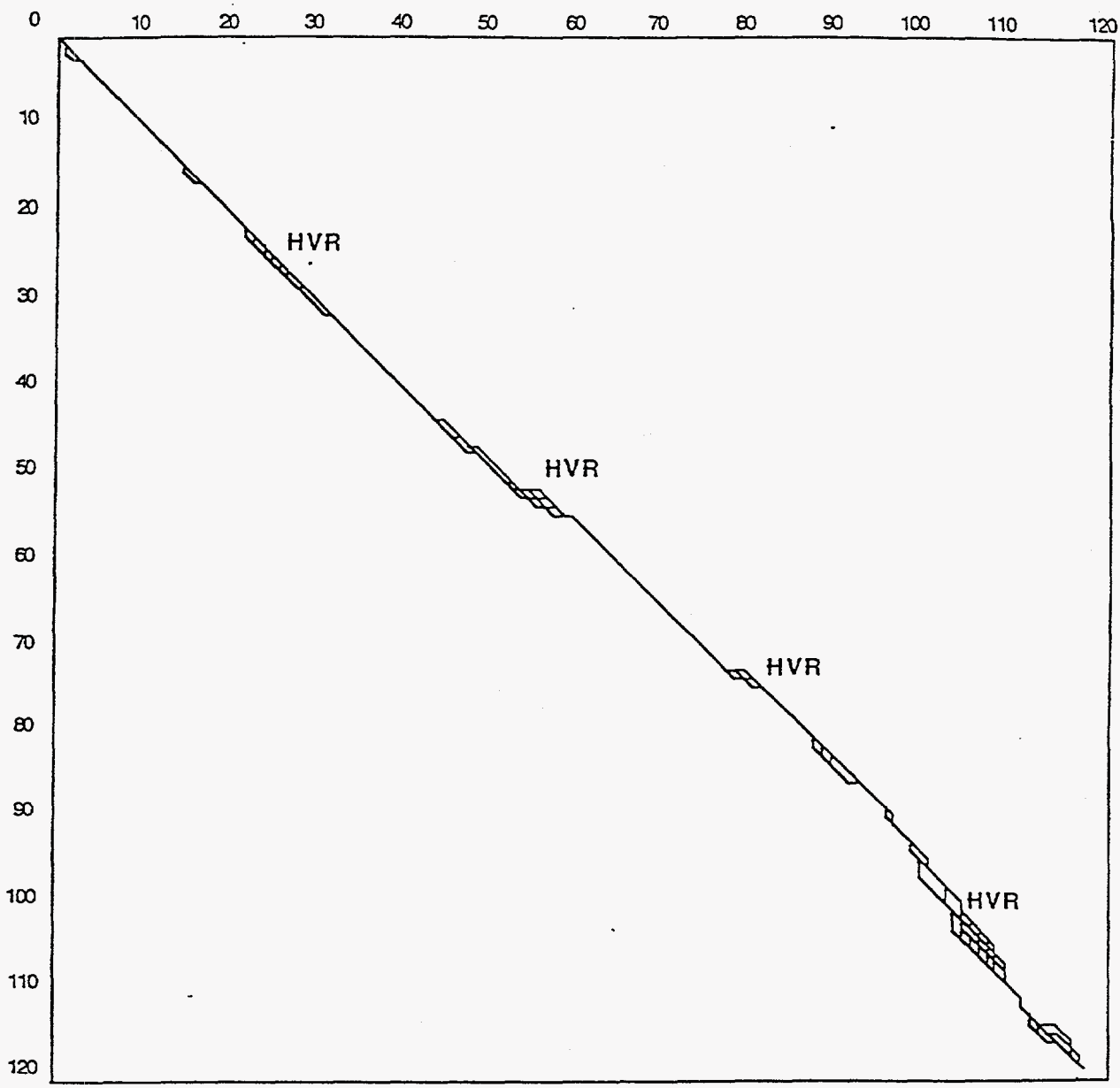
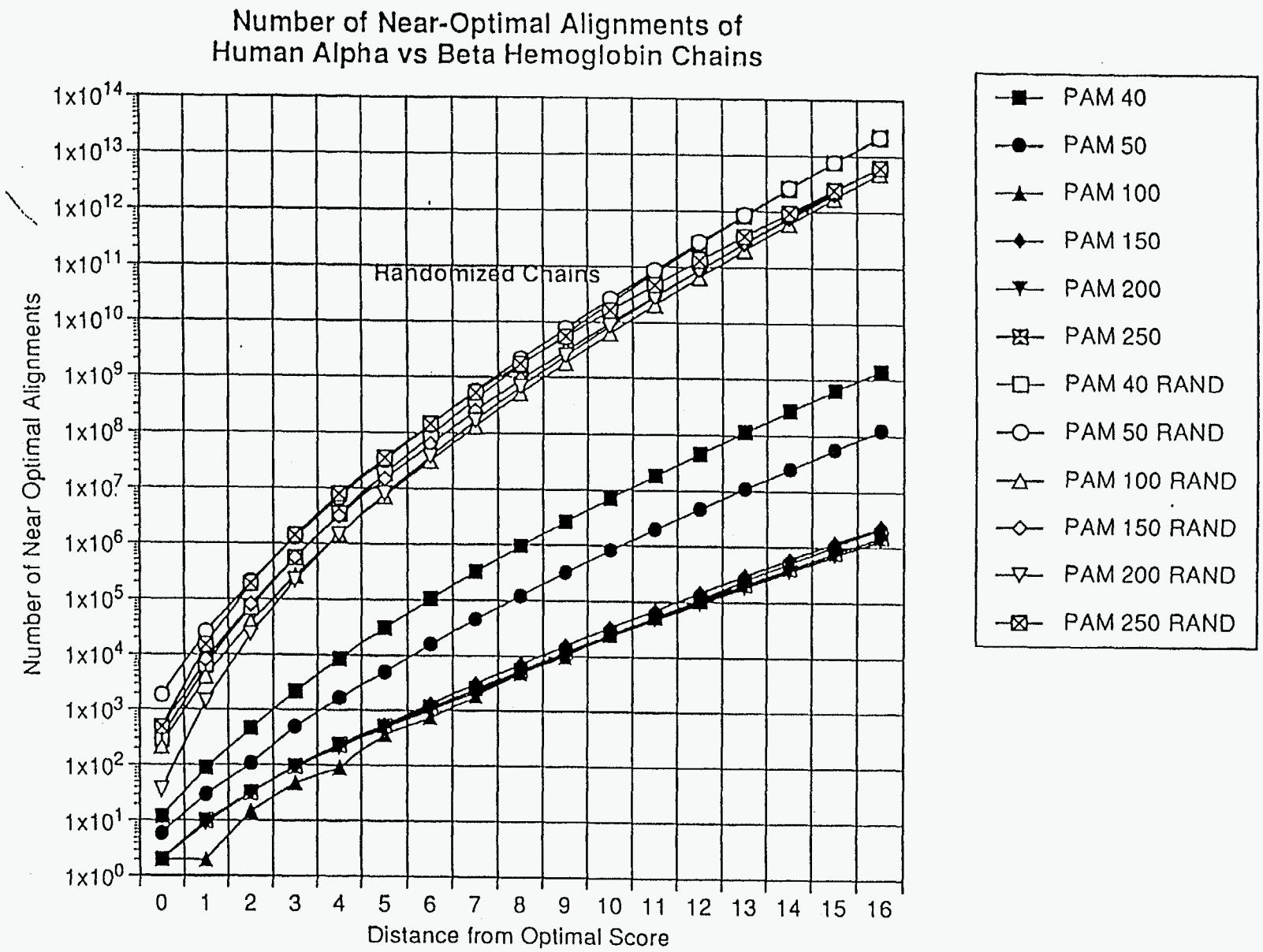
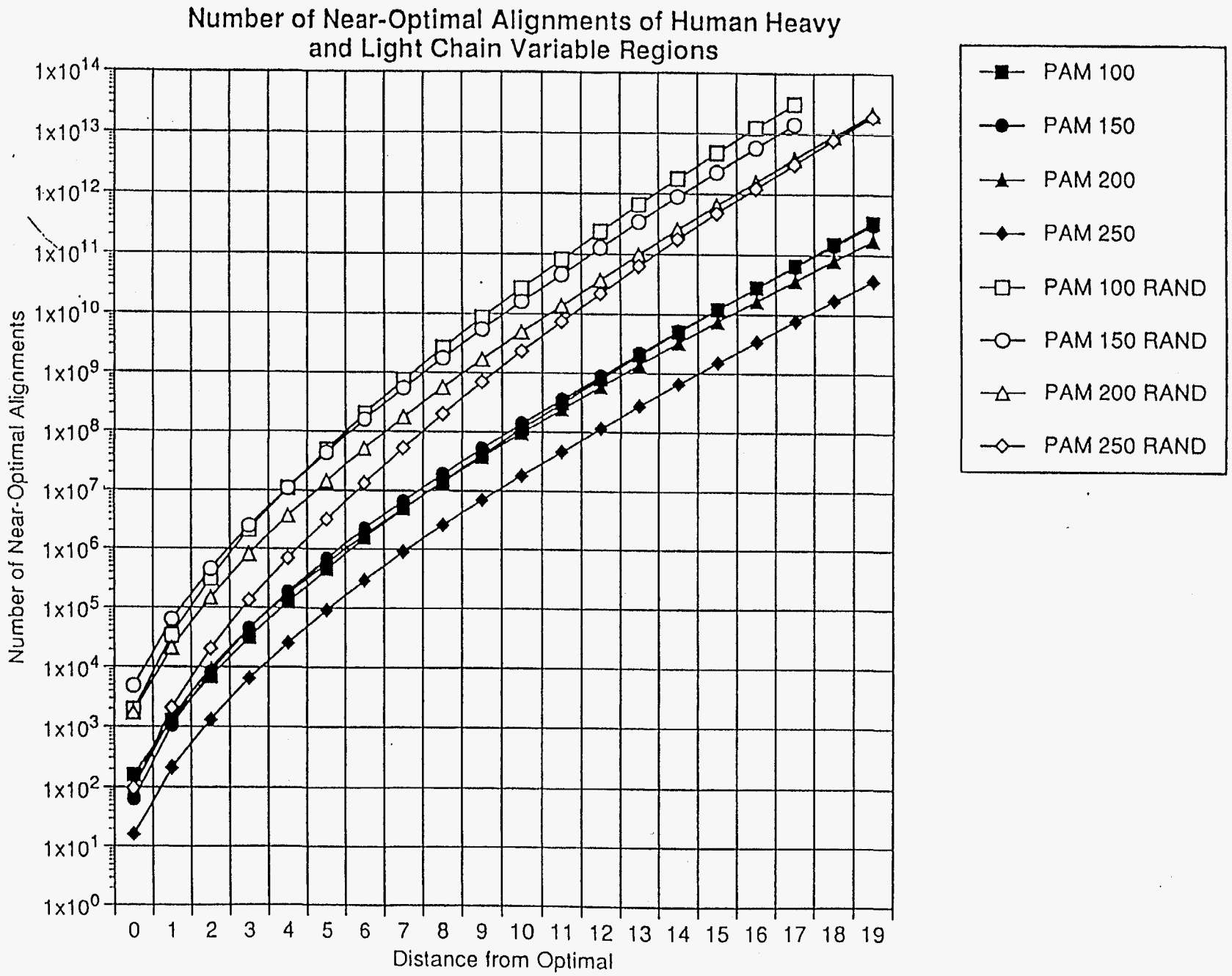


Figure 3

7


Number of Near-Optimal Alignments of
Human Alpha vs Beta Hemoglobin Chains
Randomized Chains
Number of Near Optimal Alignments
1x10^14
1x10^13
1x10^12
1x10^11
1x10^10
1x10^9
1x10^8
1x10^7
1x10^6
1x10^5
1x10^4
1x10^3
1x10^2
1x10^1
1x10^0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Distance from Optimal Score
PAM 40
PAM 50
PAM 100
PAM 150
PAM 200
PAM 250
PAM 40 RAND
PAM 50 RAND
PAM 100 RAND
PAM 150 RAND
PAM 200 RAND
PAM 250 RAND

74A


Number of Near-Optimal Alignments of Human Heavy and Light Chain Variable Regions
Number of Near-Optimal Alignments
1x10^14
1x10^13
1x10^12
1x10^11
1x10^10
1x10^9
1x10^8
1x10^7
1x10^6
1x10^5
1x10^4
1x10^3
1x10^2
1x10^1
1x10^0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Distance from Optimal
PAM 100
PAM 150
PAM 200
PAM 250
PAM 100 RAND
PAM 150 RAND
PAM 200 RAND
PAM 250 RAND

ISMB - 1995
Intelligent Systems for Molecular Biologists
Doug Brutlag

Protein Motifs

Gribskov, M., Homyak, M., Edenfield, J., & Eisenberg, D. (1988). Profile scanning for three-dimensional structural patterns in protein sequences. *Comput Appl Biosci*, *4*, 61-6.

Hodgman, T. C. (1989). The elucidation of protein function by sequence motif analysis. *CABIOS*, 5, 1-14.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. and Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262 (5131), 208-14.

Moore, J. F., Engelberg, A., & Bairoch, A. (1988). Using PC/Gene for protein and nucleic acid analysis. *Biotechniques*, 6, 566-572.

Patthy, L. (1987). Detecting homology of distantly related proteins with consensus sequences. *J. Mol. Biol.*, 198, 567-577.

Rooman, M. J., & Wodak, S. (1988). Identification of predictive sequence motifs limited by protein structure data base size. *Nature*, 335, 45-49.

Rooman, M. J., Wodak, S. J. and Thornton, J. M. (1989). Amino acid sequence templates derived from recurrent turn motifs in proteins: critical evaluation of their predictive power. *Protein Eng*, 3 (1), 23-7.

Saqi, M. A. S. and Sternberg, M. J. E. (1994). Identification of sequence motifs from a set of proteins with related function. *Protein Engineering* 7 (2), 165-171.

Staden, R. (1988). Methods to define and locate patterns of motifs in sequences. *Comput Appl Biosci*,-4 (1), 53-60.

Tatusov, R. L., Altschul, S. F. and Koonin, E. V. (1994). Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci U S A* 91 (25), 12091-5.

Taylor, W. R. (1986). Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.*, 188, 233-258.

Thornton, J. M. and Gardner, S. P. (1989). Protein motifs and data-base searching. *Trends Biochem Sci*, 14 (7), 300-4.

Position	Probe 8 - 37	38 - 67	68 - 98	130 - 159	160 - 188	189 - 214	216 - 246	247 - 276	Consensus	Profile A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Gap Opening	Gap Extension
1	V	T	T	T	.	.	K	E	T	15	-3	12	13	-8	11	4	10	12	2	6	11	9	7	6	14	32	12	-22	-8	25	25
2	V	G	G	Q	.	.	E	E	E	16	-8	22	24	-14	22	8	4	8	-1	2	15	8	16	3	15	13	9	-30	-12	25	25
3	Y	E	E	Q	.	Q	R	Q	Q	11	-16	23	27	-16	7	22	0	15	-1	3	17	8	35	14	8	7	0	-19	-11	25	25
4	K	K	K	L	A	D	T	R	K	12	-16	13	13	-15	5	8	4	31	0	11	16	8	15	21	12	15	5	-8	-12	100	100
5	R	P	N	P	G	L	V	P	P	17	-6	9	8	-12	11	9	6	8	3	7	12	28	11	12	17	12	13	-21	-15	100	100
6	Y	F	F	Y	Y	A	Y	F	Y	-5	20	-23	-19	57	-21	10	20	-20	29	10	-3	-24	-20	-18	-5	-5	8	38	58	100	100
7	I	P	T	E	P	V	L	V	V	15	-3	6	8	-2	8	4	20	3	14	14	5	17	6	4	12	16	25	-27	-8	100	100
8	C	C	C	C	C	C	C	C	C	32	142	-48	-49	-10	19	-7	29	-51	-73	-53	-25	9	-53	-27	78	28	30	-129	101	100	100
9	S	K	D	P	K	D	P	E	D	16	-13	25	22	-24	15	12	0	23	-8	1	19	21	20	16	20	15	2	-26	-18	100	100
10	F	E	S	H	K	.	R	H	H	6	-7	9	11	-4	2	21	3	11	2	3	13	6	13	15	11	7	2	-3	1	24	24
11	A	E	D	E	D	.	D	A	D	27	-15	40	37	-26	25	14	1	12	-8	-3	24	12	24	4	16	16	3	-42	-15	24	24
12	D	G	G	G	D	.	G	G	G	32	-3	43	30	-31	62	3	-7	6	-16	-9	26	14	17	-6	29	22	10	-53	-28	24	24
13	.	.	.	.	S	V	.	.	S	9	-1	5	5	0	6	4	9	5	5	6	7	5	4	5	12	8	10	-9	-1	24	24
14	C	C	C	C	C	C	C	C	C	32	142	-48	-49	-10	19	-7	29	-51	-73	-53	-25	9	-53	-27	78	28	30	-129	101	100	100
15	G	E	D	D	.	N	D	G	D	19	-13	40	31	-25	31	12	-1	11	-10	-5	27	9	21	3	18	15	4	-38	-16	28	28
16	A	K	L	K	.	R	R	K	K	9	-17	9	9	-15	0	11	2	35	-1	13	14	8	16	30	11	10	3	5	-14	28	28
17	A	G	R	R	S	K	S	C	S	15	3	9	8	-13	12	7	2	16	-7	2	13	13	8	20	27	12	4	-1	-10	100	100
18	Y	F	F	F	F	F	Y	F	F	-19	1	-46	-34	83	-33	3	41	-35	62	27	-19	-36	-36	-25	-10	-10	16	63	81	100	100
19	N	T	T	S	V	R	T	A	T	19	1	11	10	-9	14	4	10	11	1	5	15	12	6	8	19	33	12	-17	-9	100	100
20	K	S	T	L	G	H	T	M	T	13	-5	9	9	-3	10	7	9	11	7	10	11	8	8	7	15	22	10	-14	-6	100	100
23	N	L	K	P	K	K	A	K	K	13	-20	13	13	-20	5	9	1	42	-2	13	21	11	18	25	14	13	3	-10	-15	100	100
24	W	H	A	S	T	D	F	K	S	10	-7	7	6	-1	5	10	4	9	4	3	11	4	7	11	15	11	3	-6	2	100	100
25	K	H	N	R	W	Y	N	S	N	5	-8	9	7	-2	2	16	0	14	-1	1	23	3	9	17	15	6	-1	2	3	100	100
26	.	.	.	.	T	.	.	.	T	8	-3	5	5	-1	5	5	6	5	4	5	8	4	5	5	9	11	7	-9	0	25	25
27	.	.	.	.	L	.	.	.	L	7	-6	3	4	3	2	5	8	4	9	8	6	3	5	5	7	6	8	-7	1	25	25
28	.	.	.	.	Y	.	.	.	S	6	-1	3	3	4	2	7	6	3	5	5	7	2	4	4	7	6	6	-5	5	25	25
29	L	L	M	L	L	L	L	L	L	-1	-58	-32	-17	77	-28	-10	59	-17	107	92	-19	-16	-4	-19	-20	-3	61	23	18	100	100
30	Q	T	K	K	K	R	R	E	K	7	-18	15	16	-22	3	14	-1	39	-6	9	18	9	21	33	13	14	0	0	-18	100	100
31	A	R	K	R	H	D	S	R	R	9	-10	12	11	-16	3	18	-1	24	-6	5	15	12	17	33	15	8	0	10	-13	100	100
32	H	H	H	H	V	H	H	H	H	0	-12	28	28	-7	-7	104	-13	11	-8	-15	37	16	50	35	-4	-2	-11	-15	17	100	100
33	.	.	F	.	.	.	I	.	I	6	-3	1	2	6	1	5	12	2	10	8	5	1	3	3	7	6	10	-6	5	30	30
34	L	S	N	E	A	Q	Q	S	S	17	-7	17	18	-11	13	12	4	11	1	4	18	11	21	8	22	11	5	-15	-10	100	100
35	C	L	R	K	E	K	S	V	K	9	-7	7	9	-8	5	7	8	19	2	10	11	6	9	15	15	11	10	-10	-5	100	100
36	K	T	F	V	C	T	F	V	V	9	2	-4	-2	10	3	1	23	1	17	13	2	1	-4	-1	10	20	23	-12	10	100	100
37	H	H	H	H	H	H	H	H	H	-4	-17	40	39	-12	-13	152	-27	15	-17	-27	53	21	73	50	-9	-7	-26	-14	25	100	100
38									*	11	19	13	14	14	12	21	2	25	16	2	7	8	7	16	13	15	11	2	9		

FIG. 2. Profile of the *Xenopus laevis* transcription factor TFIIIA zinc finger. The eight repeats of the zinc finger sequence that form the probe are shown descending vertically at the left, labeled with the positions where they occur in the complete sequence. Insertions made to align the sequences are shown as periods. The profile calculated by **PROFILEMAKE** is shown in the box. The rows correspond to the positions in the aligned sequences, and the columns contain the score for each possible amino acid residue when aligned at that position. The position-specific gap penalties are given in the two right-hand columns. The consensus sequence is shown immediately to the left of the box, and represents the highest scoring column at each row in the profile. In other words, the consensus residue is the amino acid that would receive the highest score when aligned with that position in the aligned probe sequences.

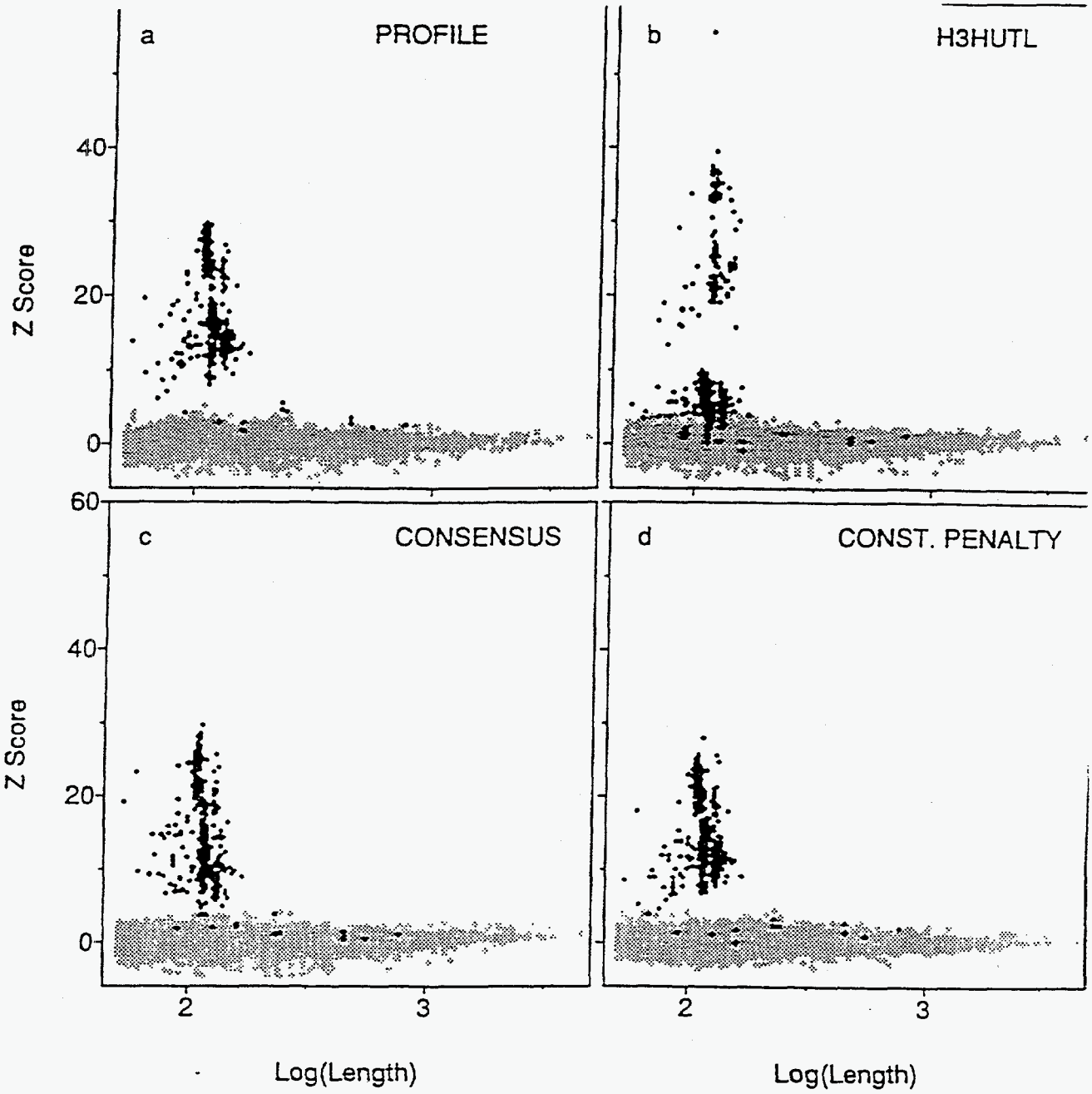
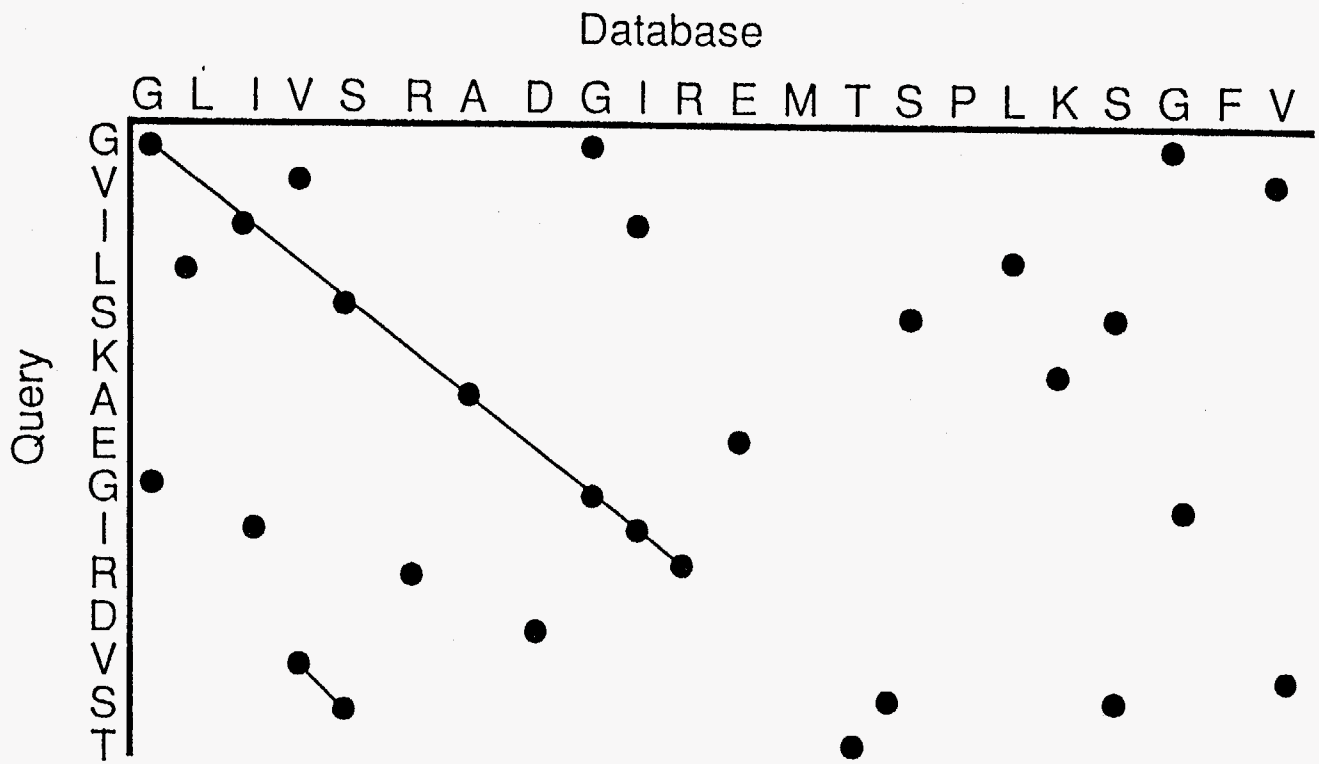


FIG. 3. Selectivity of a profile, compared to simpler alternatives. Each panel shows tl scores from comparison of the specified profile or sequence to each sequence in the PSQ ar New databases. The scores have been normalized by **PROFILENORMAL** and are shown c a Z scale, i.e., they have been scaled such that unrelated sequences have a mean of 0.00 and standard deviation of 1.0. The Z scores are plotted against the log of the length of the databa sequence. Sequences related to the immunoglobulin variable region motif are shown as blac circles; unrelated sequences are lightly shaded. (a) Profile derived from 20 human and mou κ, λ, and heavy chain variable regions. (b) Heavy chain variable region, PSQ entry H3HUT (c) Consensus sequence derived from 20 sequences in (a). (d) Profile (a) with insertion/del tion penalties set to a constant value at all positions.

Dynamic Programming



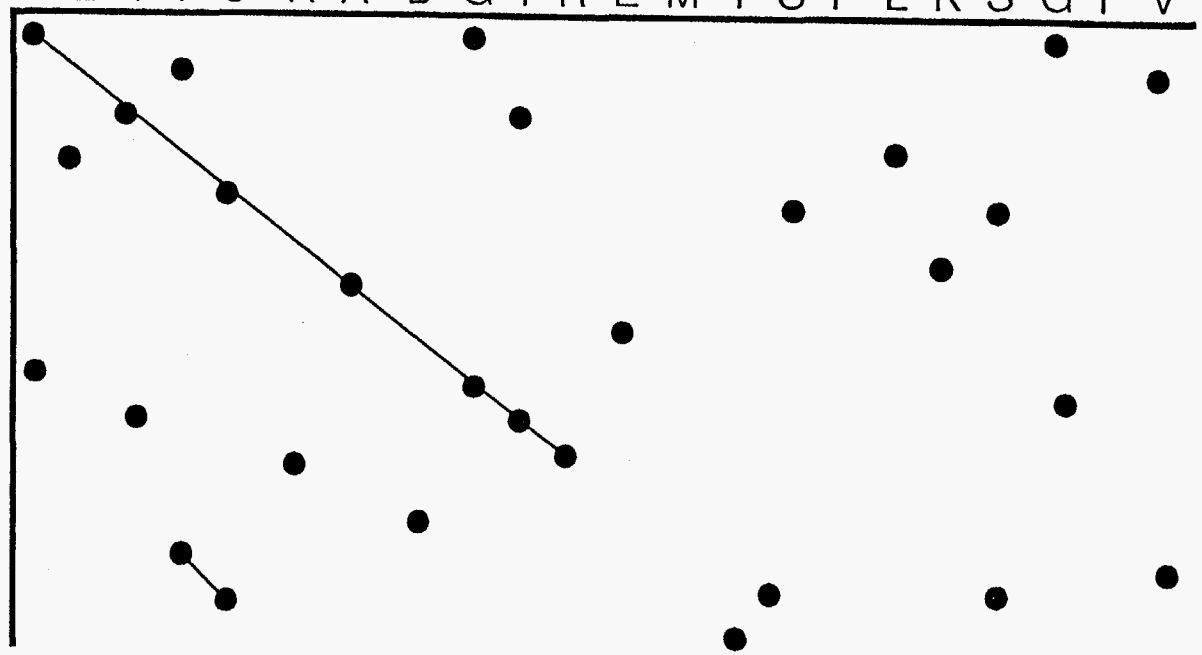
Generalized Dynamic Programming

Query

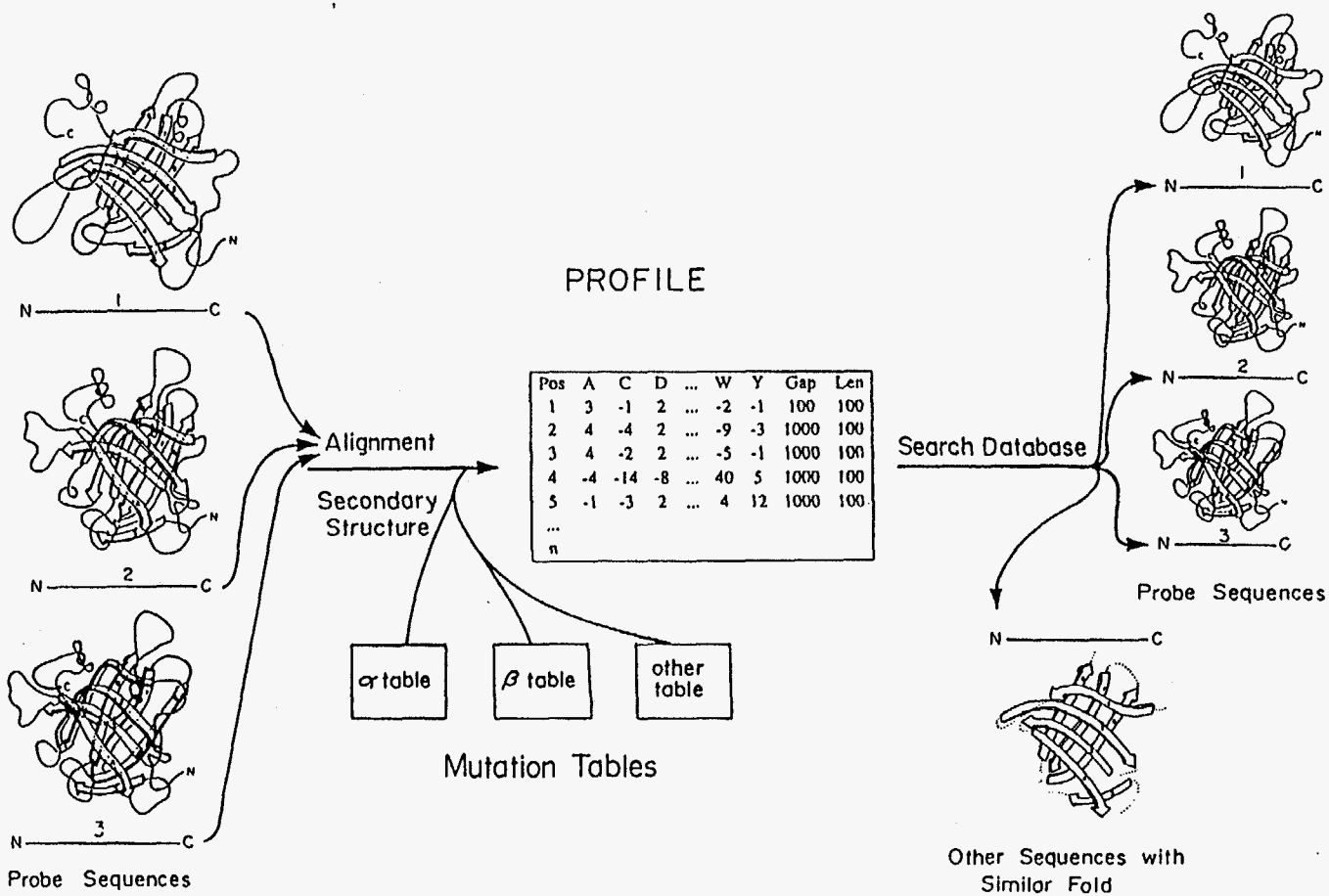
```
A R N D C Q E G H I L K M F P S T W Y V
(3 4 5 1-5 2 7 1 9-3 5 0-6 1 2 5 6-7 3 4)
(3 0 5 9-5-3 2 2-3-3 2 0-2 1 1-5 6-7 3 4)
(1 3 5 2-5-3 2 2-3 2 2 0-2 1 1-5 6-7 3 4)
(6 4-3 0 2-1-3-1 4 3-5 1-3 3 4-5 2-3 2-1)
(1 3 5 2-5-3 2 2-3 2 2 0-2 1 1-5 6-7 3 4)
(3 0 5 9-5-3 2 2-3-3 2 0-2 1 1-5 6-7 3 4)
(2-3 4-2 5 2-3 1-1 0 2 5-4 2-3 4 5-1 0 4)
(6 4-3 0 2-1-3-1 4 3-5 1-3 3 4-5 2-3 2-1)
(2-3 4-2 5 2-3 1-1 0 2 5-4 2-3 4 5-1 0 4)
(1 3 5 2-5-3 2 2-3 2 2 0-2 1 1-5 6-7 3 4)
(3 4 5 1-5 2 7 1 9-3 5 0-6 1 2 5 6-7 3 4)
(1 3 5 2-5-3 2 2-3 2 2 0-2 1 1-5 6-7 3 4)
(6 4-3 0 2-1-3-1 4 3-5 1-3 3 4-5 2-3 2-1)
(2-3 4-2 5 2-3 1-1 0 2 5-4 2-3 4 5-1 0 4)
(2-3 4-2 5 2-3 1-1 0 2 5-4 2-3 4 5-1 0 4)
(6 4-3 0 2-1-3-1 4 3-5 1-3 3 4-5 2-3 2-1)
(3 0 5 9-5-3 2 2-3-3 2 0-2 1 1-5 6-7 3 4)
(1 3 5 2-5-3 2 2-3 2 2 0-2 1 1-5 6-7 3 4)
```

Database

G L I V S R A D G I R E M T S P L K S G F V



2D Structure Compatibility Search



SAM Scoring Matrices

Amino Acid Replacements Observed in Secondary Structures

SAM250 (α-helix)

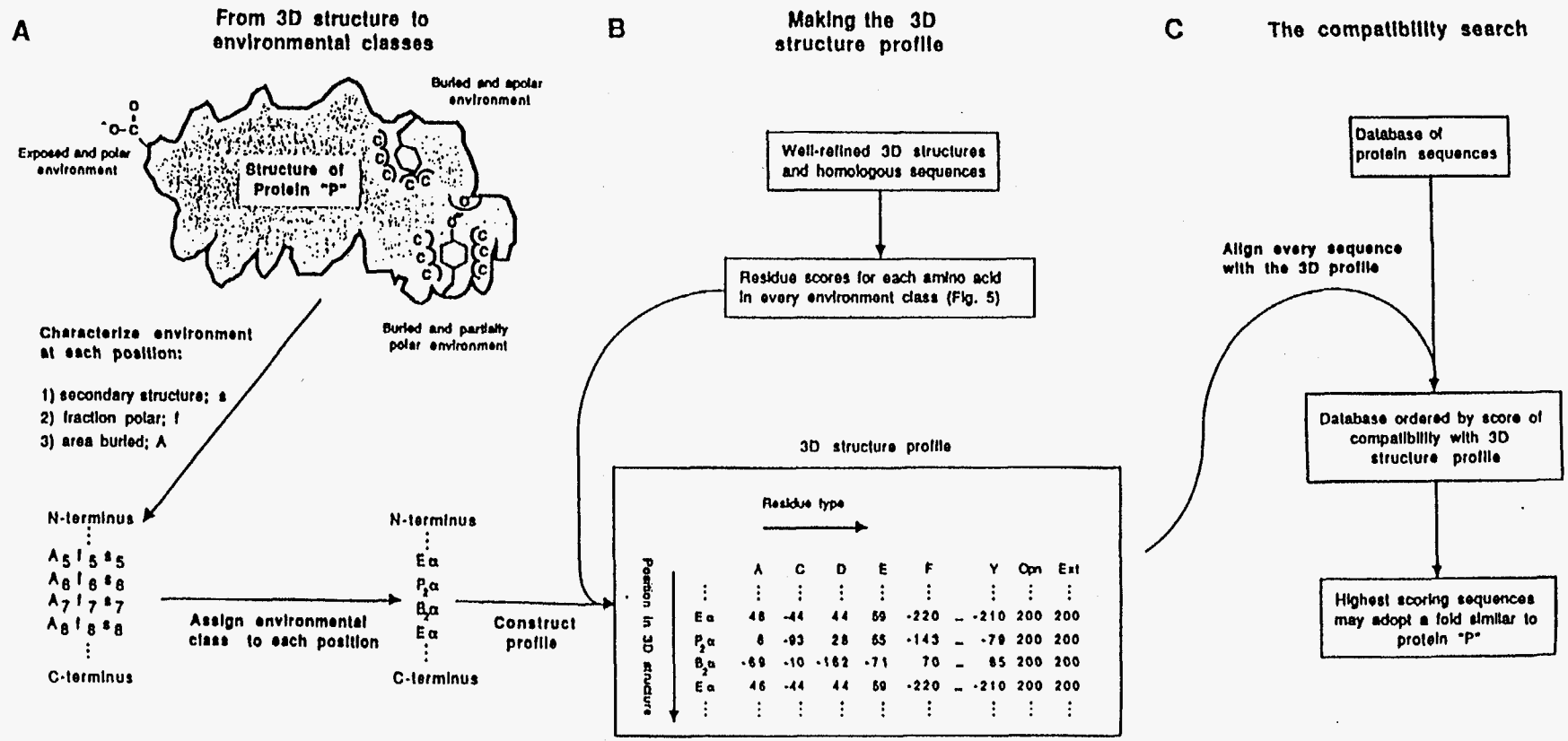
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	50																			
C	17	9																		
D	23	5	16																	
E	24	7	16	28																
F	11	5	4	4	56															
G	27	7	13	14	4	31														
H	17	5	8	9	7	9	34													
I	15	4	5	6	5	6	4	9												
K	21	8	13	16	5	13	12	7	55											
L	15	11	8	9	16	8	9	20	9	100										
M	13	2	3	5	6	4	3	5	7	25	16									
N	21	4	10	11	4	12	8	4	13	9	3	8								
P	20	2	7	9	1	10	4	2	7	6	1	5	30							
Q	19	4	9	12	4	9	8	4	14	10	3	7	4	8						
R	15	3	6	8	3	7	7	3	19	9	2	5	2	5	17					
S	25	8	12	13	5	15	9	7	14	11	5	10	8	9	8	15				
T	23	6	9	11	5	12	8	5	13	11	4	8	5	7	6	11	9			
V	22	12	10	11	10	12	9	22	11	24	16	10	8	10	9	14	13	62		
W	6	1	1	2	14	1	1	1	2	6	1	1	1	1	1	2	1	3	92	
Y	7	1	2	2	18	2	8	2	3	7	1	2	1	1	1	2	2	5	10	74

SAM250 (β-turn)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	11																			
C	3	68																		
D	12	3	21																	
E	9	2	13	11																
F	5	2	5	3	30															
G	21	5	21	18	7	68														
H	9	3	9	7	11	10	58													
I	5	2	6	4	6	9	7	3												
K	16	5	17	16	7	17	13	11	61											
L	7	8	8	7	12	8	7	14	10	72										
M	4	2	5	3	3	7	5	2	10	21	2									
N	10	3	13	9	5	18	11	5	16	8	4	10								
P	12	2	7	7	3	10	8	4	8	5	3	8	100							
Q	7	1	8	6	3	13	10	3	16	6	2	7	8	4						
R	6	3	7	5	2	13	6	2	20	6	1	6	7	4	7					
S	13	8	15	12	7	22	11	8	18	9	6	13	12	9	9	19				
T	9	3	11	7	4	15	8	5	15	8	3	9	8	5	5	12	10			
V	7	4	8	5	7	11	7	6	11	16	5	7	5	4	4	9	6	12		
W	2	1	4	1	1	7	5	1	6	4	1	4	1	1	1	5	1	1	58	
Y	4	5	5	3	15	6	13	3	6	9	3	5	4	3	2	7	4	4	6	29

- Similarity between amino acids depends on *secondary structure context*, in addition to the amino acids themselves.
- SAM matricies have been used in database search (Eisenberg, 1991); we plan to attempt using them for structure prediction.
- SAM ≡ Acceptable Structural Mutation

3D Environment Compatibility Search



Environment class	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R
$B_1\ \alpha$	1.00	1.32	0.18	1.27	1.17	0.66	1.26	-0.66	-2.53	-1.16	-0.73	-1.29	-2.73	-1.08	-1.93	-1.74	-1.97	-0.34	-1.82	-1.67
$B_1\ \beta$	1.17	0.85	0.07	1.13	1.47	1.09	0.55	-0.79	-2.02	-0.94	-0.22	-1.12	-2.91	-1.67	-1.42	-1.93	-2.56	-1.91	-2.69	-1.18
B_1	1.05	1.45	0.17	1.10	1.11	1.02	0.98	-0.91	-1.92	0.26	-1.22	-1.53	-2.81	-1.17	-2.42	-2.52	-1.76	-1.12	-2.59	-2.16
$B_2\ \alpha$	0.50	0.90	0.85	1.01	0.63	0.68	1.12	-0.69	-1.49	-2.21	-0.10	-1.50	-1.47	-0.23	-0.61	-0.71	-1.62	0.23	-0.78	0.06
$B_2\ \beta$	0.01	1.18	1.06	0.76	1.31	1.06	0.64	-1.55	-2.26	-0.49	-0.87	-2.27	-1.77	-1.22	-2.07	-1.07	-1.41	-0.77	-1.14	-0.20
B_2	1.02	1.05	1.12	0.84	0.81	0.60	0.90	-0.66	-1.66	0.19	-0.05	-0.76	-1.17	-0.76	-0.66	-1.35	-1.28	0.46	-2.34	-0.80
$B_3\ \alpha$	0.92	-0.03	0.58	0.15	0.04	-0.02	0.89	-0.57	-1.86	-0.68	-1.56	-0.57	-0.96	0.22	-0.06	0.08	-0.50	0.73	0.43	0.96
$B_3\ \beta$	0.75	0.81	1.30	0.18	0.54	0.56	-0.57	-0.93	-1.93	-0.34	-0.54	-0.44	-0.74	0.21	-0.24	-0.14	-0.86	0.82	-0.53	0.13
B_3	1.07	0.70	1.13	0.35	-0.17	-0.03	0.23	-0.96	-0.98	-0.13	-1.20	-0.53	-0.54	0.05	0.04	-0.36	-1.05	1.01	0.10	0.66
$P_1\ \alpha$	-1.35	-0.82	-0.59	-0.52	-0.24	0.10	-0.03	0.73	-0.49	-0.25	0.95	0.31	0.34	-0.14	-0.54	-0.17	-0.25	-0.52	-0.21	-0.28
$P_1\ \beta$	0.36	-0.49	0.17	-1.03	0.20	0.46	-0.27	0.64	-0.82	-0.55	1.49	0.93	0.33	-2.27	-1.32	-0.73	-1.07	-0.42	-1.21	-0.77
P_1	-1.26	-1.20	-1.31	-0.62	-0.23	-0.01	-1.19	0.46	-0.24	0.66	1.35	0.56	0.49	-0.63	-0.13	-0.61	0.38	-1.12	-0.74	-1.29
$P_2\ \alpha$	-1.14	-1.43	-0.79	-0.35	-0.54	-0.48	-0.45	0.06	-0.50	-0.26	-0.93	-0.05	-0.18	0.55	-0.05	0.56	0.28	0.06	0.61	0.50
$P_2\ \beta$	-0.79	-0.54	-0.84	-1.30	-0.33	0.13	-0.72	-0.55	-0.98	-1.29	-0.57	0.84	0.59	-0.08	-0.16	0.32	0.19	-0.87	0.59	0.10
P_2	-0.82	-0.86	-0.51	-0.70	-1.09	-0.88	-0.89	-0.15	-0.40	0.44	-0.60	0.06	0.26	0.27	0.50	0.27	0.49	0.13	0.44	0.30
$E\ \alpha$	-1.35	-2.20	-2.10	-1.58	-2.76	-1.10	-0.72	0.46	0.68	0.04	-0.44	-0.17	0.15	0.36	0.28	0.59	0.44	-0.19	0.13	-0.34
$E\ \beta$	0.64	-0.90	0.30	-1.66	-1.47	-1.74	-0.68	0.06	1.46	-0.96	-0.24	0.14	0.65	-0.19	-0.06	-0.16	-0.78	-0.83	-0.52	-0.49
E	-2.14	-1.90	-0.94	-1.19	-1.61	-0.91	-1.67	0.12	1.13	0.20	-0.46	0.12	0.32	-0.03	0.41	0.03	0.22	-0.25	-0.14	-0.32

Fig. 5. The 3D-1D scoring table. The scores for pairing a residue i with an environment j is given by the information value (*61*),

$$\text{3D-1D score } ij = \ln\left(\frac{P(i{:}j)}{Pi}\right)$$

where $P(i{:}j)$ is the probability of finding residue i in environment j and Pi is the overall probability of finding residue i in any environment. These probabilities were determined from a database of 16 known protein structures and sets of homologous sequences aligned to the sequence of known structure as described in Lüthy *et al.* (*28*). For each position in the aligned set of sequences, we determined the environment category of the position from the known structure and counted the number of each residue type found at the position within the set of aligned sequences. A residue type was counted only once per position. For example, if there were ten aspartates and one glycine found at a position in a set of aligned sequences, then both the Asp and Gly counters were both incremented by only one. The total number of residue replacements in our database was 8273. If the number of residues i in an environment j was found to be zero, the number was increased to one so that $P(i{:}j)$ was never zero. Boundaries for the environment categories (shown in Fig. 3) were adjusted iteratively to maximize the total 3D-1D score summed over all residues in our database:

$$\text{Total 3D-1D score} = \sum_{ij} N_{ij} \ln\left(\frac{P(i{:}j)}{Pi}\right)$$

where N_{ij} is the number of residues i in environment j. In this case, if N_{ij} was zero, the number was not increased to one. Instead, that term in the sum was treated as zero.

		Amino acid type													Gap penalty	
Position in fold	Environment class	A	C	D	E	F	G	...	R	S	T	V	W	Y	Opn	Ext
1	E	12	-46	22	3	-190	113	...	-32	32	12	-91	-214	-94	2	0.02
2	B_2	-68	-5	-128	-135	105	-168	...	-80	-117	-76	60	102	112	2	0.02
3	E α	46	-44	44	59	-220	68	...	-34	15	-17	-110	-135	-210	200	200
4	P_2α	6	-93	28	56	-143	-50	...	50	-18	-5	-48	-114	-79	200	200
5	E α	46	-44	44	59	-220	68	...	-34	15	-17	-110	-135	-210	200	200
6	P_2α	6	-93	28	56	-143	-50	...	50	-18	-5	-48	-114	-79	200	200
7	B_2α	-69	-10	-162	-71	90	-149	...	6	-147	-150	68	50	85	200	200
8	E α	46	-44	44	59	-220	68	...	-34	15	-17	-110	-135	-210	200	200
9	P_2α	6	-93	28	56	-143	-50	...	50	-18	-5	-48	-114	-79	200	200
10	B_1α	-68	-73	-197	-174	132	-253	...	-167	-273	-129	66	100	18	200	200
.	.	.	.	.	.	.	.		.	.	.		.	.	.	.
.	.	.	.	.	.	.	.		.	.	.		.	.	.	.
.	.	.	.	.	.	.	.		.	.	.		.	.	.	.

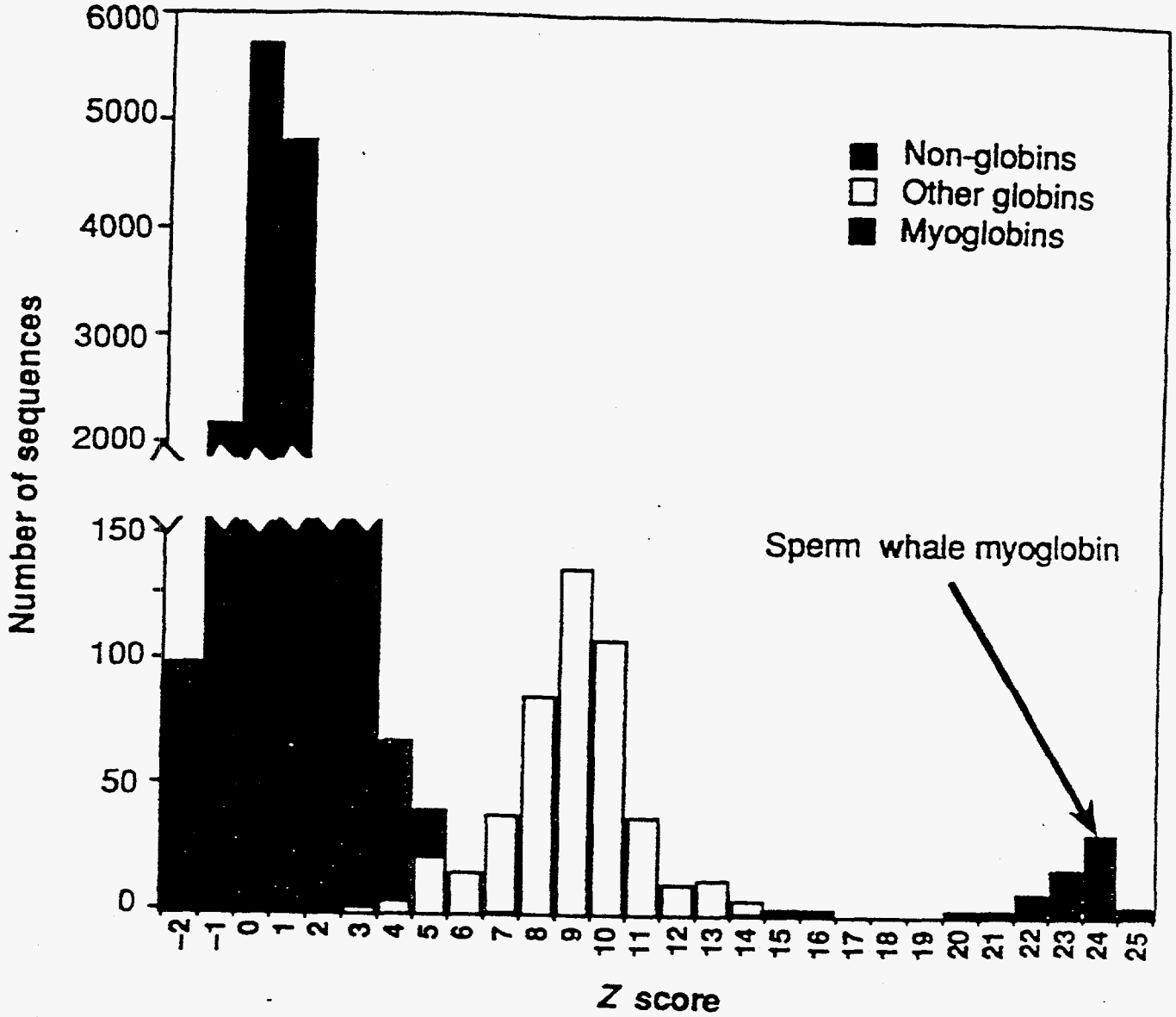


Fig. 6. Results of a compatibility search for the structure of sperm whale myoglobin. Myoglobin sequences are represented by black bars, other globin sequences are represented by white bars, and all other sequences are shown in gray bars. Sperm whale myoglobin is the eighth highest scoring protein (Z score = 23.7). Gaps were not allowed in helical regions (as defined in the protein data bank file). In nonhelical regions, a gap-opening penalty of 2.0 and a gap-extension penalty of 0.02 was used.

81

FIG. 2 *a*, Stereo view of the α-carbon backbone of the HSC70 ATPase fragment, along with the ATP molecule. The different colours correspond to the different structural domains: domain IA, green; IB, cyan; IIA, orange; IIB, magenta. *b*, Schematic drawing of the structure. Picture produced by a program written by A. M. Lesk and K. D. Hardman[41,42].

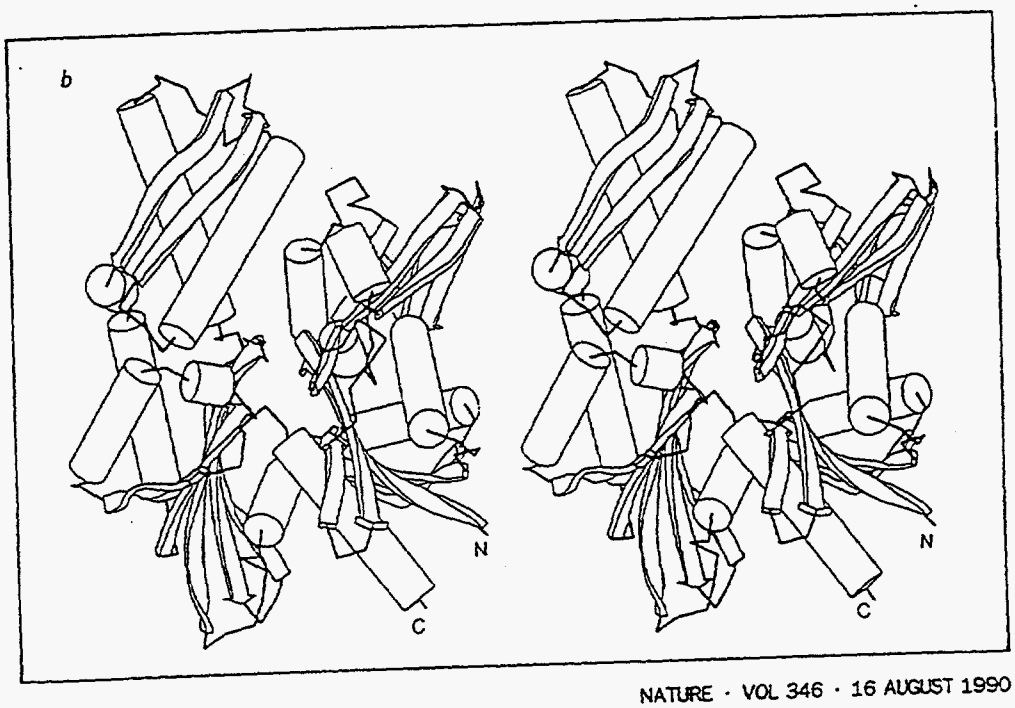
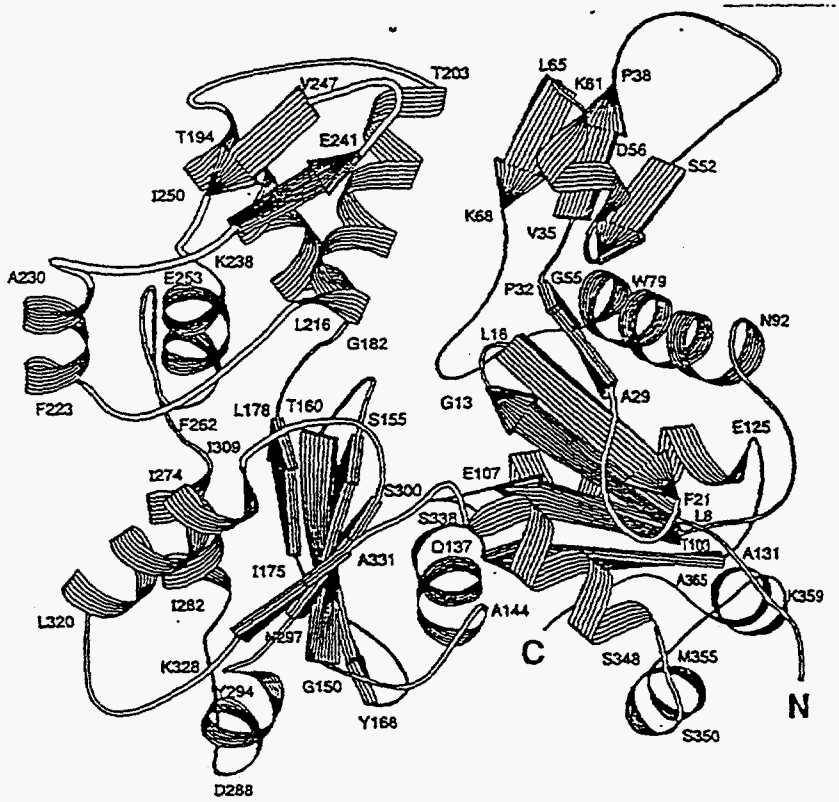


FIG. 1 Structure of ATP-actin:DNase I. *a*, C_α-stereo plot. Open circles mark C_α-positions of actin residues. The C_α-atoms of DNase I are connected by thin lines. Residues 102 and 103 are omitted as their positions have not been assigned. Three Ca^{2+} in the DNase I region and a single Ca^{2+} near the phosphates of ATP in actin are shown by filled circles. *b*, Schematic representation[64] of the three-dimensional structure of actin shown in the same orientation as *a*. First and last amino-acid residues in the helices and sheet strands are specified. The assignment of secondary structure is based on the automatic procedure of Kabsch and Sander[25]. However, in the drawing some of the helices and sheet strands have been extended by one or two residues beyond the strict assignments where geometry indicates that the secondary structure was likely to become more extended when refinement is complete.



86

Protein	Z score
↑	88.11
69 of 71 Actin Sequences	↕
↓	21.22
Kinase-related transforming protein (fgr)- feline sarcoma virus	17.47
Actin 5C - fruit fly	9.29
68-kD Heat shock protein - mouse	8.12
70-kD Heat shock protein - frog	7.95
70-kD Major heat shock - fruit fly	7.03
70-kD Heat shock cognate protein-bovine	**6.99**
HNRNP complex, protein C - frog	6.74
70-kD Heat shock cognate protein - human	6.31

Fig. 8. Sequence compatibility search with a 3D structure profile for actin (*47*). All sequences that received a Z score of 6.0 or greater are listed. A gap-opening penalty of 5.0 and a gap-extension penalty of 0.2 were used. The fgr protein is the result of a gene fusion between actin and a tyrosine-specific protein kinase (*63*). The bovine HSC70 protein, known to have a similar structure to actin, received a Z score of 6.99 and is shown in bold type.

```
============================
============================                    B L A Z E   (tm)
                  =======
               =======       A  High-Performance High-Sensitivity Biological
            =======       Sequence    Similarity    Searching     Program
         =======       Utilizing a Massively  Parallel  Implementation
      =======       of  the   Dynamic  Programming  Algorithm    of
   =======       Smith  and  Waterman
 =======
 ==========================================================================
 ========================== Release 1.0 - July 1992 =======================
```


INPUT PARAMETERS

```
DATALIB          SWISS-PROT 25
MATRIX           PAM150
GAPPEN           2.0
GAPSIZPEN        0.2
SCORES           1000
MINPERCENTMATCH  0

; ID   ACTS_HUMAN     STANDARD;      PRT;   377 AA.
; AC   P02568;
```

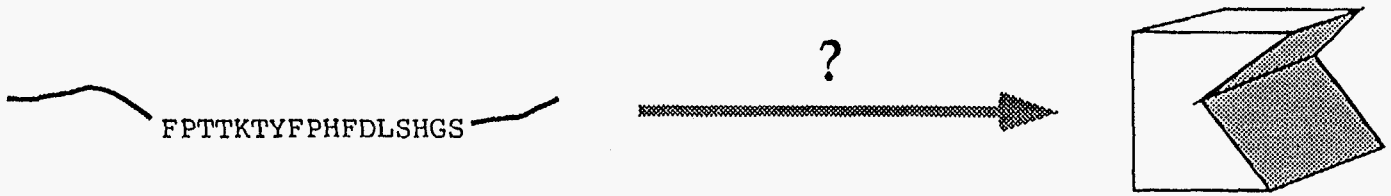
SEARCH STATISTICS

Query sequence length:	377
Score of query vs. itself:	1793
Number of sequences searched:	29955
Number of residues:	10214020
Mean score:	41
Standard Deviation:	89.75
Time:	0:01:01.109
Millions of residues compared per second:	63.013

	Sequence Name	Description	Length	Score	%Match	Exp
1.	ACTS_HUMAN	ACTIN, ALPHA SKELETAL MUSCLE	377	1793	100	0.000
2.	ACT2_XENTR	ACTIN, ALPHA SARCOMERIC/CARD	377	1787	100	0.000
3.	ACT2_XENLA	ACTIN, ALPHA SARCOMERIC/CARD	377	1785	100	0.000
4.	ACTC_HUMAN	ACTIN, ALPHA CARDIAC.	377	1784	99	0.000
95.	ACT_EUPCR	ACTIN.	379	1190	66	0.000
96.	ACT_OXYFA	ACTIN.	357	1171	65	0.000
97.	ACT4_SOLTU	ACTIN 85C (FRAGMENT).	195	888	50	0.000
98.	ACT_PINCO	ACTIN (FRAGMENT).	161	663	37	0.003
99.	KFGR_FSVGR	FGR TYROSINE KINASE TRANSFOR	545	619	35	0.011
100.	ACTS_PLEWA	ACTIN, ALPHA SKELETAL MUSCLE	125	609	34	0.014
101.	ACT1_ABSGL	ACTIN 1 (FRAGMENT).	140	606	34	0.015
102.	ACT1_DROME	ACTIN-5C (FRAGMENT).	137	393	22	3.543
103.	VCAP_VZVD	MAJOR CAPSID PROTEIN (MCP).	1396	81	5	99.999
104.	POLG_HCVA	GENOME POLYPROTEIN.	3898	79	4	99.999
105.	SYI_METTH	ISOLEUCYL-TRNA SYNTHETASE (E	1045	78	4	99.999
106.	POLG_HCVB	GENOME POLYPROTEIN.	3898	77	4	99.999
107.	RRPL_TSWVB	RNA-DIRECTED RNA POLYMERASE	2875	77	4	99.999
108.	GTFC_STRMU	GLUCOSYLTRANSFERASE-SI PRECU	1375	77	4	99.999
109.	CARB_BACSU	CARBAMOYL-PHOSPHATE SYNTHASE	1071	76	4	99.999
110.	VCAP_HSVSA	MAJOR CAPSID PROTEIN (MCP).	1371	76	4	99.999
111.	MCM2_YEAST	MINICHROMOSOME MAINTENANCE P	890	75	4	99.999
302.	HS70_MYCLE	HEAT SHOCK 70 KD PROTEIN (70	621	63	4	99.999
455.	HS70_MYCPA	HEAT SHOCK 70 KD PROTEIN (70	623	60	3	99.999
699.	HS71_DROME	MAJOR HEAT SHOCK 70 KD PROTE	643	57	3	99.999

The Structure Problem

"Given a protein sequence, compute it's structure."

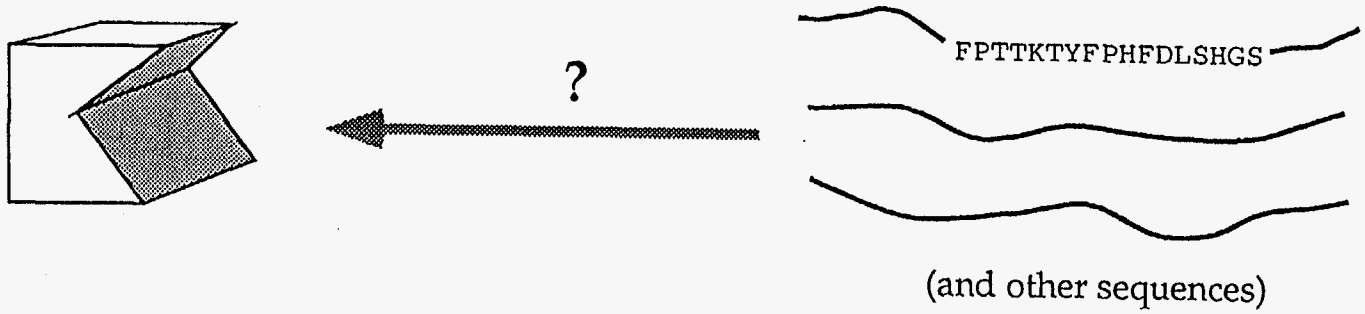


(Disclaimer: sample sequence and structure do not necessarily correspond to each other)

- Theoretically possible
- Astronomical, highly underconstrained search space
- Biophysics complex and incomplete
- Practically, next to impossible

Inverse Structure Problem

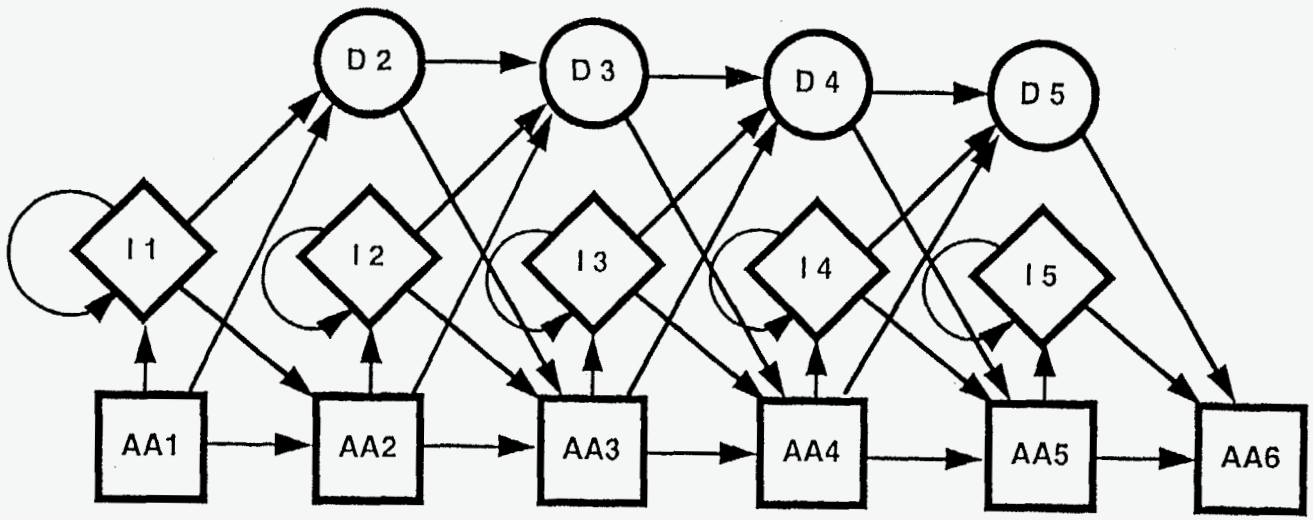
"Given a structure, what sequences fold into it?"



- Finite number of <u>structural motifs</u> (supersecondary structures)?
- Match sequences to known structures.
- Match <u>structural features</u> of amino acid positions.

Hidden Markov Model

(after Haussler)



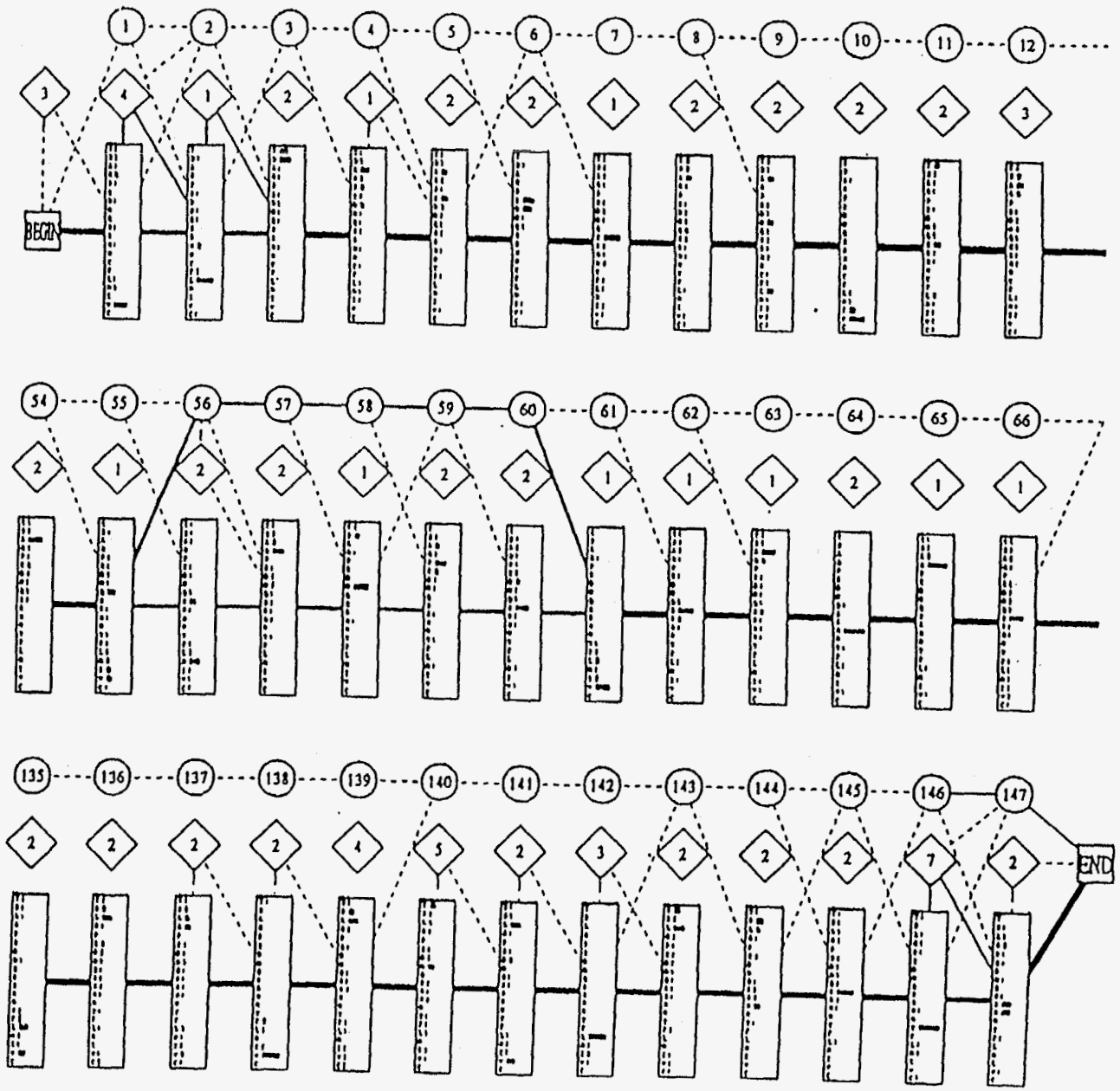


Figure 9: Parts of the final globin model. The position numbers are shown in the delete states.

ISMB - 1995
Intelligent Systems for Molecular Biologists
Doug Brutlag

Positional Correlations in Biological Sequences

Cooper, G. F. (1989). Current Research Directions in the Development of Expert Systems Based on Belief Networks. *App. Stoch. Models and Data Anal.*, 5, 39-52.

Gutell, R. R., Power, A., Hertz, G. Z., Putz, E. J. and Stormo, G. D. (1992). Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucl. Acids. Res.*, 20 (21), 5785-5795.

Klingler, T. and Brutlag, D. L. (1993). Detection of Correlations in tRNA with structural implications. in First International Conference on Intelligent Systems for Molecular Biology. Washington D.C. Eds. AAAI Press, Menlo Park, CA., pp. 225-233.

Klingler, T. M. and L., B. D. (1994). Discovering Side-Chain Correlations in α-helices. in Second International Conference on Intelligent Systems for Molecular Biology. Stanford University. Eds. Altman, R., Brutlag, D. L., Karp, P., Lathrop, R. and Searls, D. AAAI Press, pp. 236-243.

Klingler, T. M. and Brutlag, D. L. (1994). Discovering Structural Correlations in α-Helices. *Protein Science* 3 , 1847-1857.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. Royal Stat. Soc.*, 50, 157-224.

Neapolitan, R.E. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*, John Wiley and Sons, New York, NY, 1990.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, Inc., San Mateo, CA, 1988.

Richards, F. M. and Kundrot, C. E. (1988). Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins*, 3 (2), 71-84.

Rozkot, F., Sazelova, P. and Pivec, L. (1989). A novel method for promoter search enhanced by function-specific subgrouping of promoters-developed and tested on e.coli system. *Nucl. Acids Res.*, 17, 4799-4815.

Shoemaker, K.R., Fairman, R., Schultz, D.A., Robertson, A.D., York, E.J., Stewart, J.M. and Baldwin, R.L. (1990). *Biopolymers*, 29:1-11.

9.

Multiple Representations of Sequences

Weight Matrices and Profiles

Position

	1	2	3	4	5	6	7	8	9	10	11	12
A	2	1	3	13	10	12	67	4	13	9	1	2
R	7	5	8	9	4	0	1	16	7	0	1	0
N	0	8	0	1	0	0	0	2	1	1	10	0
D	0	1	0	1	13	0	0	12	1	0	4	0
C	0	0	1	0	0	0	0	0	0	2	2	1
Q	1	1	21	8	10	0	0	7	6	0	0	2
E	2	0	0	9	21	0	0	15	7	3	3	0
G	9	7	1	4	0	0	8	0	0	0	46	0
H	4	3	1	1	2	0	0	2	2	0	5	0
I	10	0	11	1	2	10	0	4	9	3	0	16
L	16	1	17	0	1	31	0	3	11	24	0	14
K	3	4	5	10	11	1	1	13	10	0	5	2
M	7	1	1	0	0	0	0	0	5	7	1	8
F	4	0	3	0	0	4	0	0	0	10	0	0
P	0	6	0	1	0	0	0	0	0	0	0	0
S	1	17	0	8	3	1	3	0	2	2	2	0
T	5	22	3	11	1	5	0	2	2	2	0	5
W	2	0	0	0	0	0	0	0	0	1	0	1
Y	1	0	4	2	0	1	0	0	2	4	0	1
V	6	3	1	1	2	15	0	0	2	12	0	28

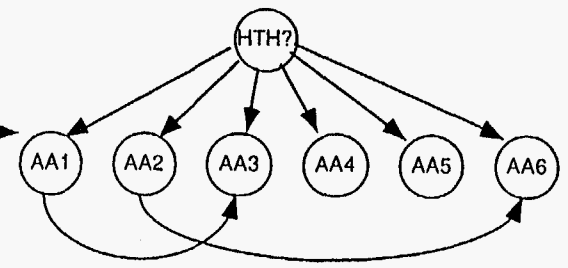
Consensus Sequences

Zinc Finger (C2H2 type)

CX{2,4}CX{12}HX{3,5}H

Sequences of Common Structure or Function

Bayesian Networks

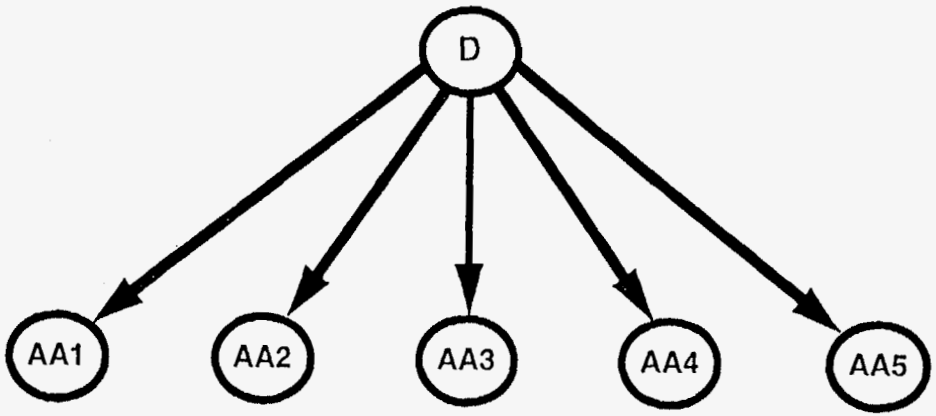


Sequence Alignments

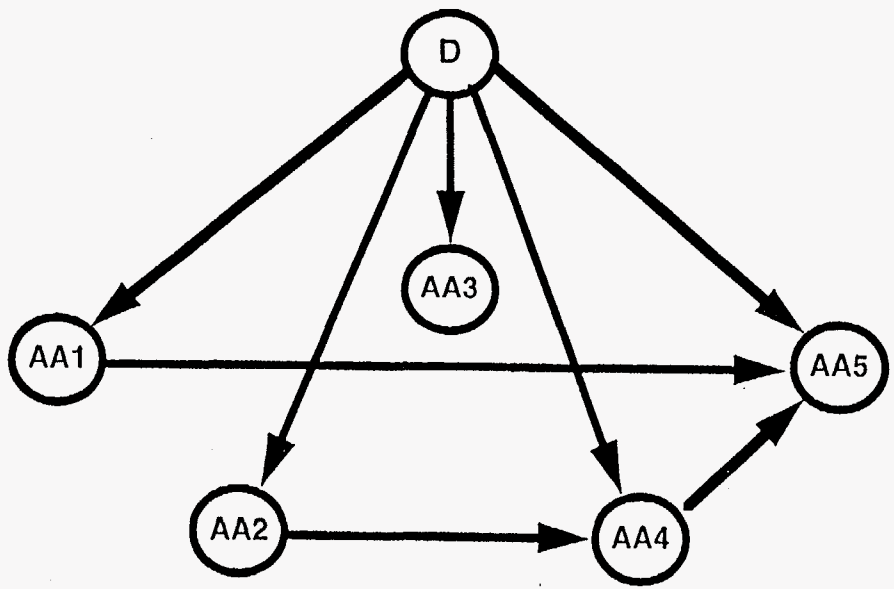
```
                    10        20        30        40        50
1     VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
       |:| :|: | |:||||     | |:||| |: : :|:|  :|  |      |:  |
2     HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
                    10        20        30        40        50
```

Initial Score	= 63	Optimized Score	= 98	Significance	= 5.51
Residue Identity	= 14%	Matches	= 21	Mismatches	= 22
Gaps	= 2	Conservative Substitutions			= 11

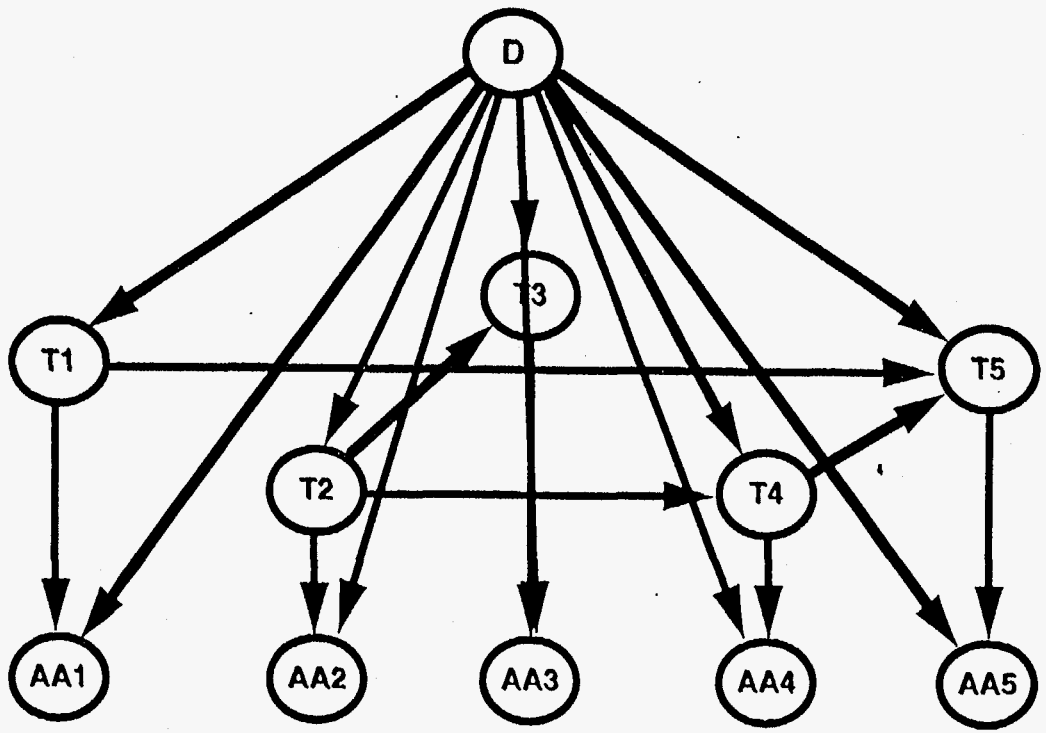
Simple Bayesian Network



Bayesian Network with Positional Correlations



Bayesian Network with Amino Acid Type Classes



Testing for Correlations in Sequences

- Chi-square statistics

  Test against null hypothesis that two positions have independent sequence distributions.

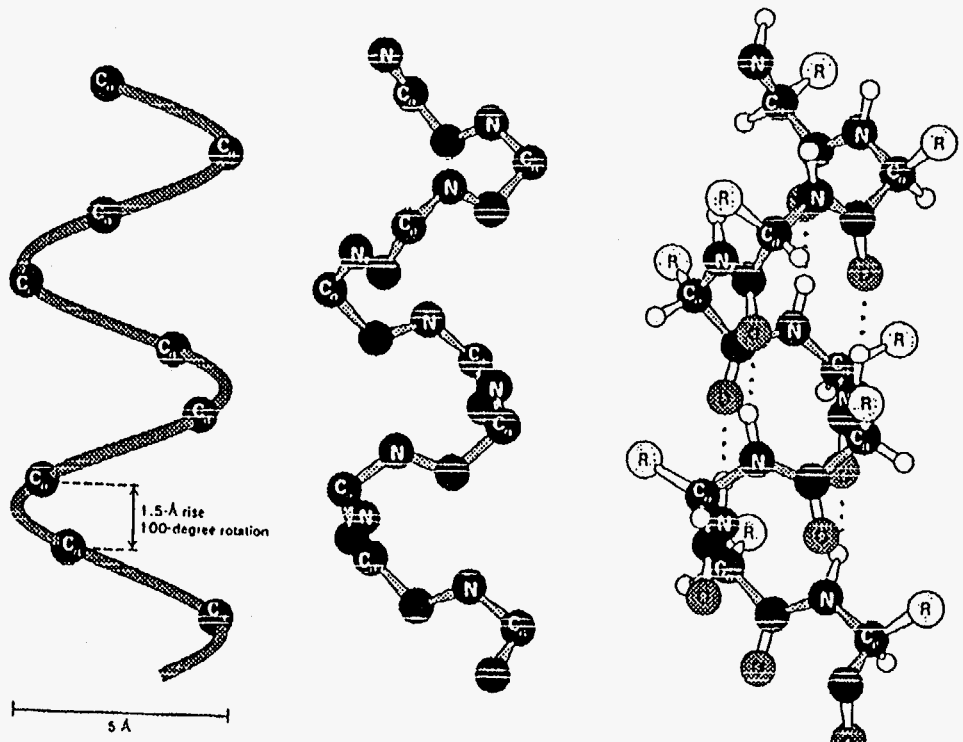
- Mutual information

  Based on entropies of sequence distributions.

- Monte Carlo simulation

  Repeated testing of simulated data sets.

α-Helix



Adapted from Biochemistry
by Stryer

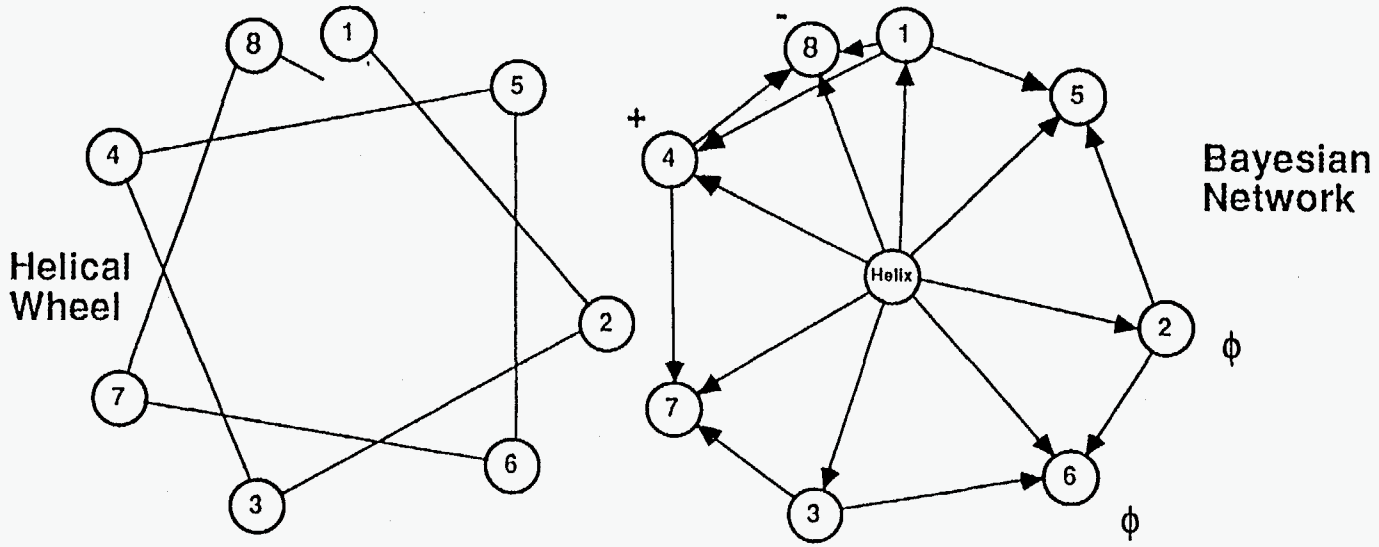
Sequence Data

- 3157 overlapping helical segments 8 long from 181 separate α-helices.
- 2349 overlapping strand segments 4 long from 316 β-sheets.

- 181 N-cap and C-cap sequences

- 7124 helical (i, i+1) residue pairs
- 6405 helical (i, i+2) residue pairs
- 5686 helical (i, i+3) residue pairs
- 4967 helical (i, i+4) residue pairs

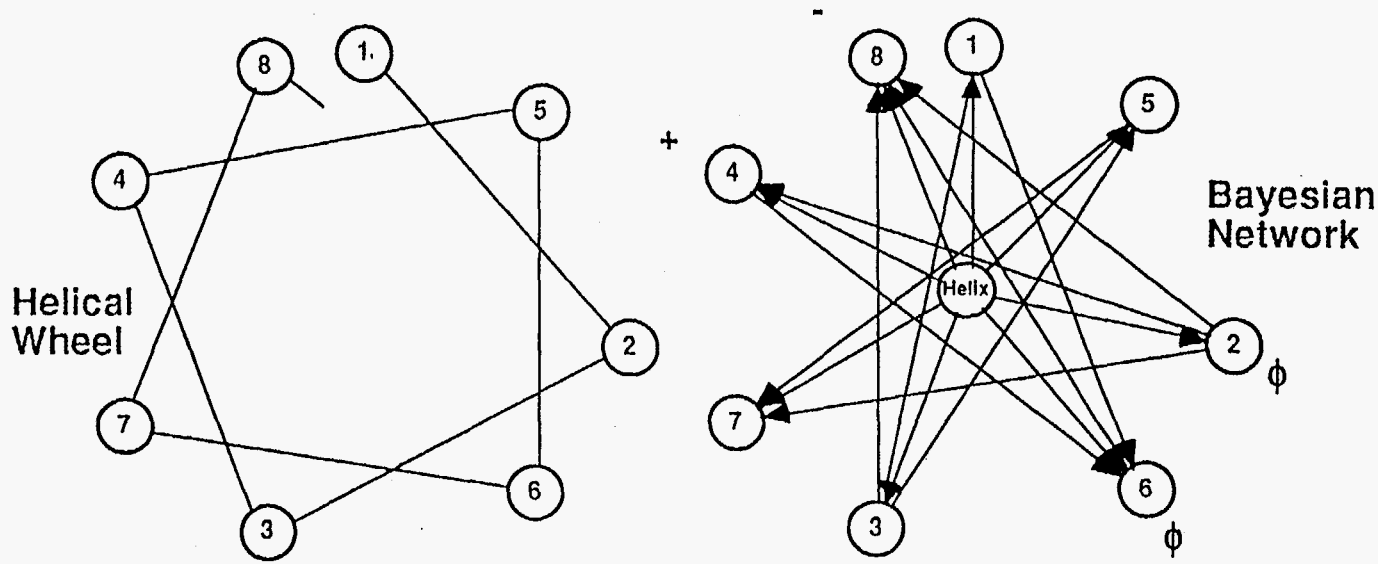
Hydrophobic Patch



Dependences between positions *i* and *i+4*:

Position *i* Hydrophobic	⇒	Position *i+4* P(Hydrophobic) ↑
	⇒	P(Hydrophilic) ↓
Hydrophilic	⇒	P(Hydrophilic) ↑
	⇒	P(Hydrophobic) ↓

Amphipathic α-Helix



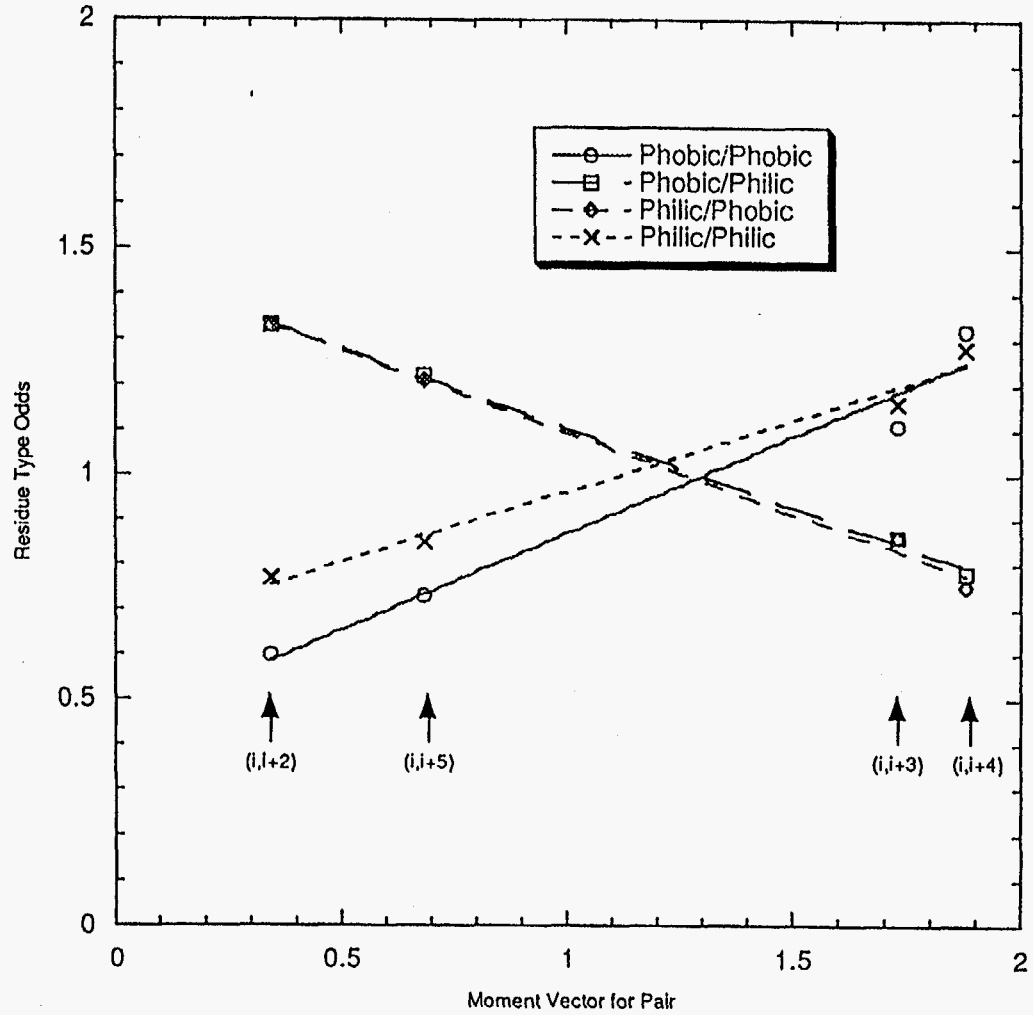
Dependence between positions *i* and *i+2*:

Position *i*		Position *i+2*	
Hydrophobic	⇒	P(Hydrophilic)	↑
	⇒	P(Hydrophobic)	↓
Hydrophilic	⇒	P(Hydrophobic)	↑
	⇒	P(Hydrophilic)	↓

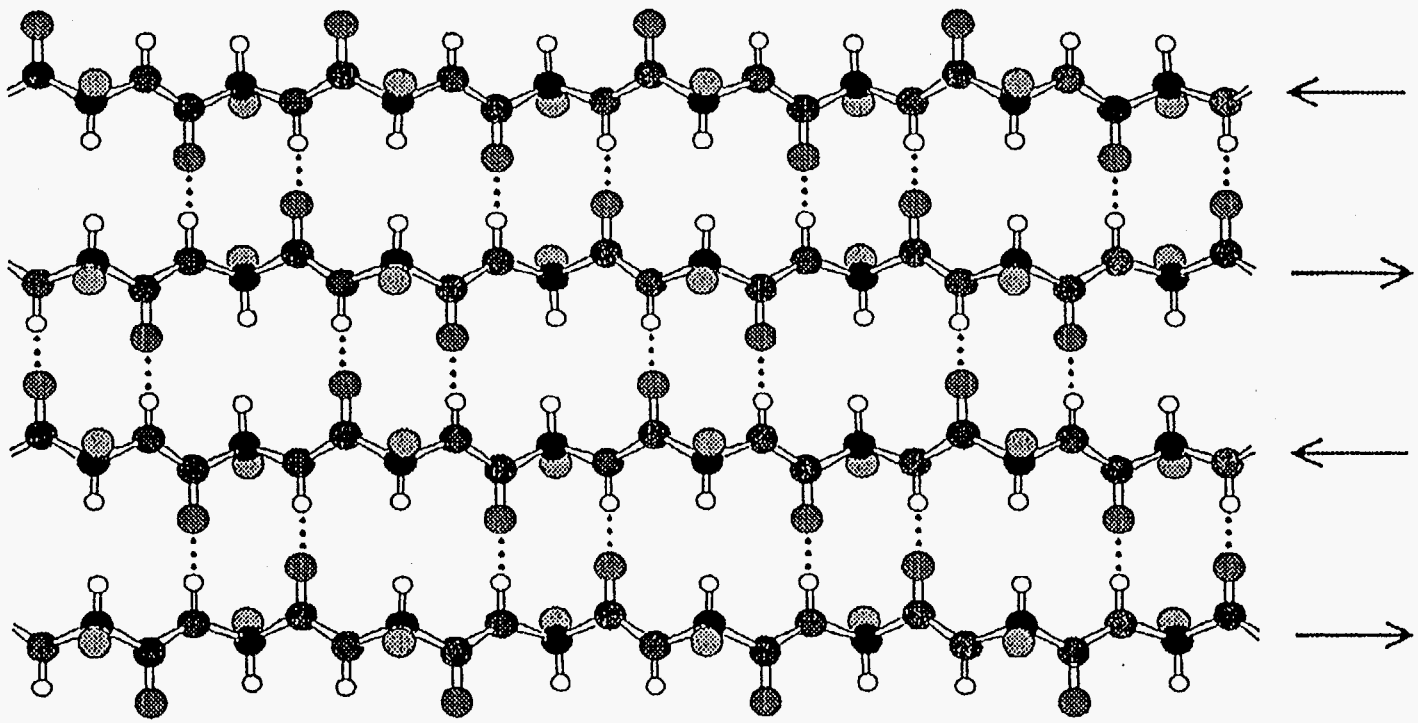
Hydropathy Correlations in α-Helices

Position	Phobic-Phobic	Odds	Phobic-Philic	Odds	Philic-Phobic	Odds	Philic-Philic	Odds
(i, i+2)	342(568)	0.60	866(650)	1.33	903(677)	1.33	595(773)	0.77
(i, i+3)	580(520)	1.11	473(553)	0.86	542(627)	0.86	772(666)	1.16
(i, i+4)	569(431)	1.32	388(495)	0.78	397(528)	0.75	776(607)	1.28
(i, i+5)	270(370)	0.73	512(418)	1.22	556(461)	1.21	439(519)	0.85

Residue Odds vs. Hydropathic Moment Vector

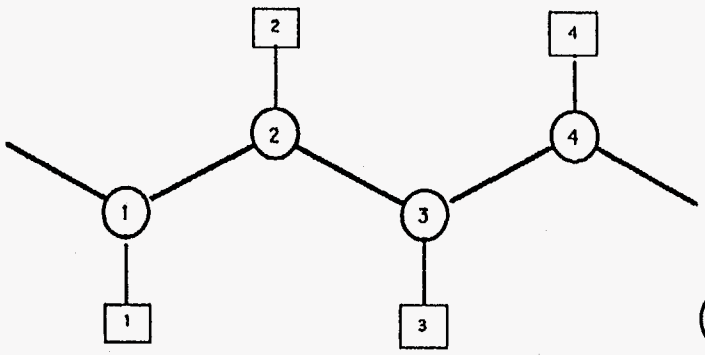


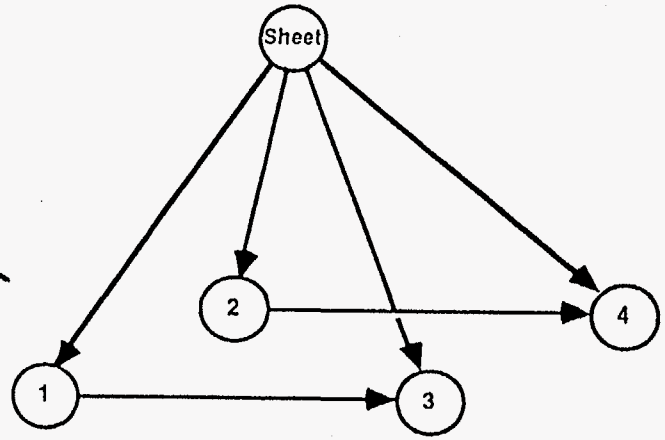
β-Sheet



Adapted from Biochemistry
by Stryer

Amphipathic β-Sheet

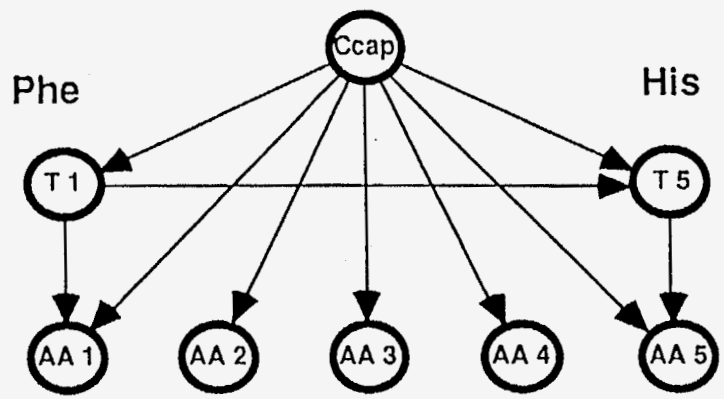




Dependence between positions *i* and *i+2*:

Position *i* Hydrophobic	⇒	Position *i+2* P(Hydrophobic) ↑
	⇒	P(Hydrophilic) ↓
Hydrophilic	⇒	P(Hydrophilic) ↑
	⇒	P(Hydrophobic) ↓

Phe-His Bridge in C-termini of α-Helices

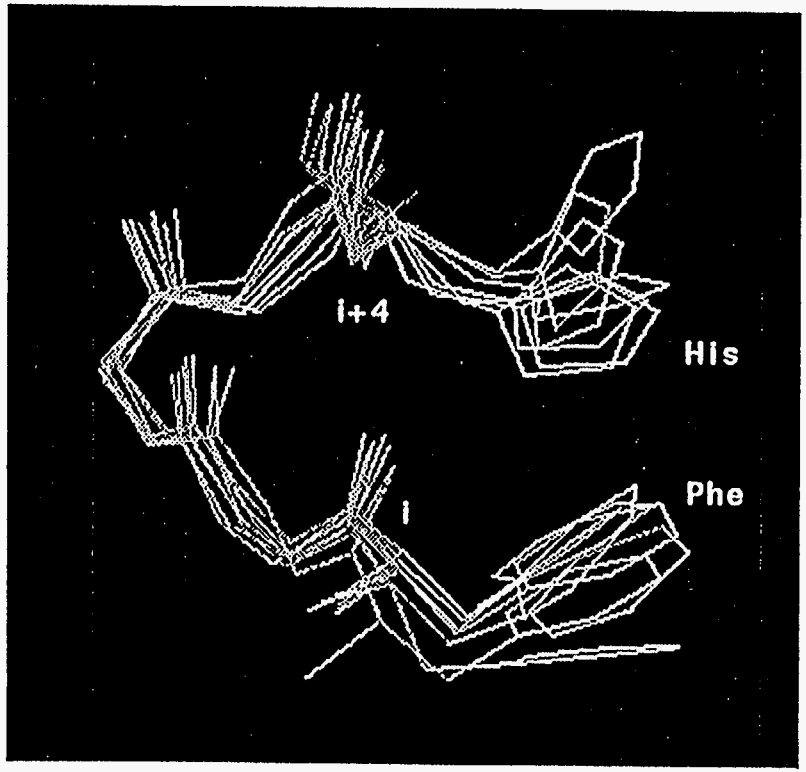


Dependence between positions *end-4* and *end*:

	Other	PheTyr	His	Row Sum
Other:	189:182(0.3)	19:22(0.3)	2: 6(2.9)	210
PheTyr:	13: 19(1.9)	4: 2(1.3)	5> 1(28.6)	22
His:	1: 2(0.3)	1: 0(3.1)	0: 0(0.1)	2
Col Sum	203	24	7	234

$p = 0.0001$, $\chi^2 = 38.86$, $df = 4$

Phenylalanine-Histidine Bridges



Example Residue Correlation in (i, i+4) Positions

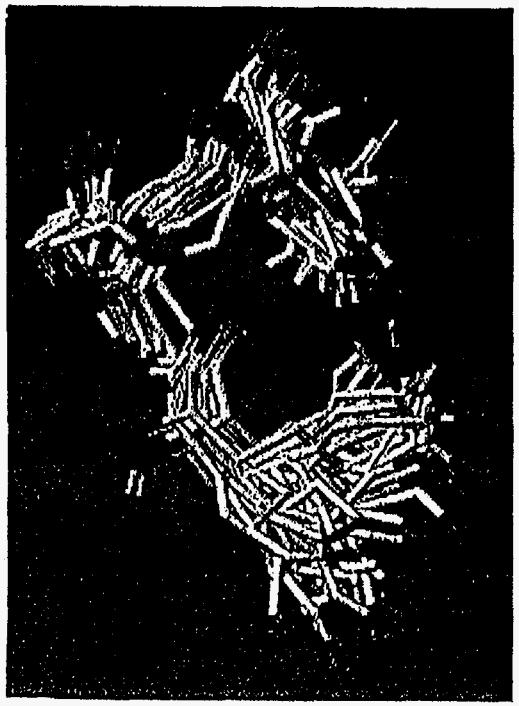
i vs. i+4	Aspartate	Not Asp
Lysine	33 (11.8)	250 (271)
Not Lys	172 (193)	4456 (4435)

$\chi^2 = 42.2$, $p < 0.0001$, Odds = 2.79

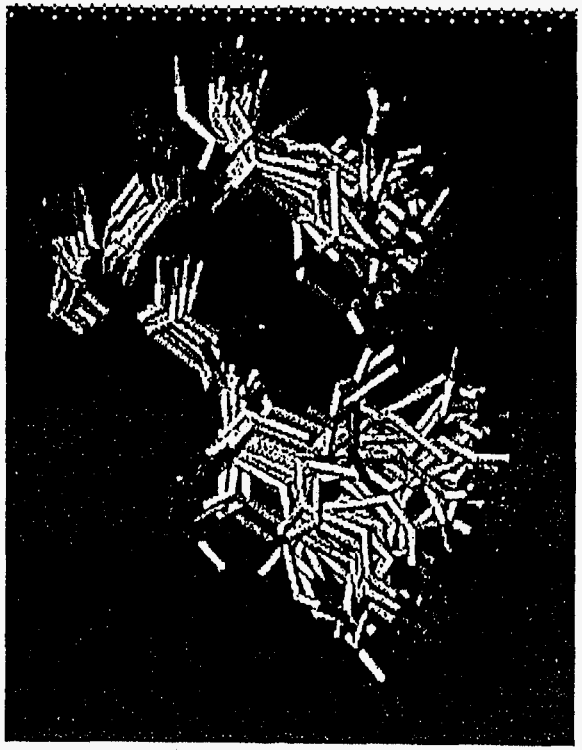
Residue Correlations at (i, i+4)

	obs.	exp.	$\chi 2$	Odds
KD	33	11.8	42.1	2.79
KE	42	20	27.6	2.10
LL	97	62.1	25.0	1.56
EK	55	30.4	23.4	1.81
FM	17	6.15	20.6	2.76
IL	60	37.9	15.8	1.58
QE	32	17.3	14.1	1.85
KL	16	36.1	13.6	0.44
SA	47	29.3	13.0	1.61
GA	43	27.8	10.1	1.55
PF	13	5.68	10.1	2.29

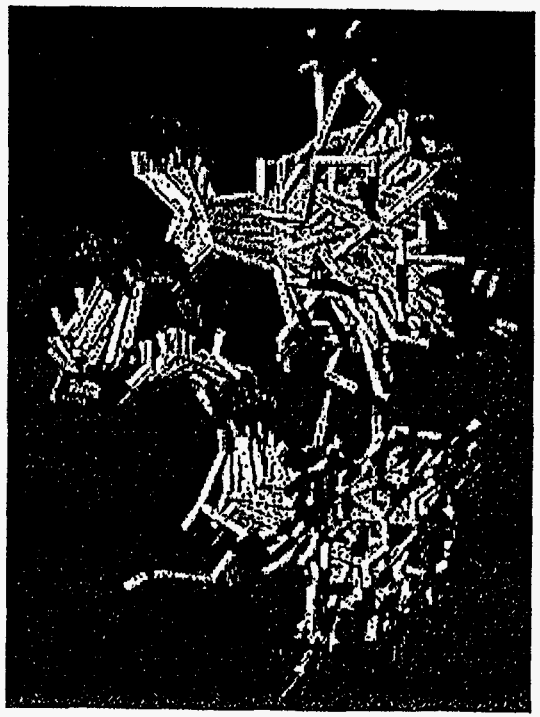
(i, i+4) helical pairs



KD

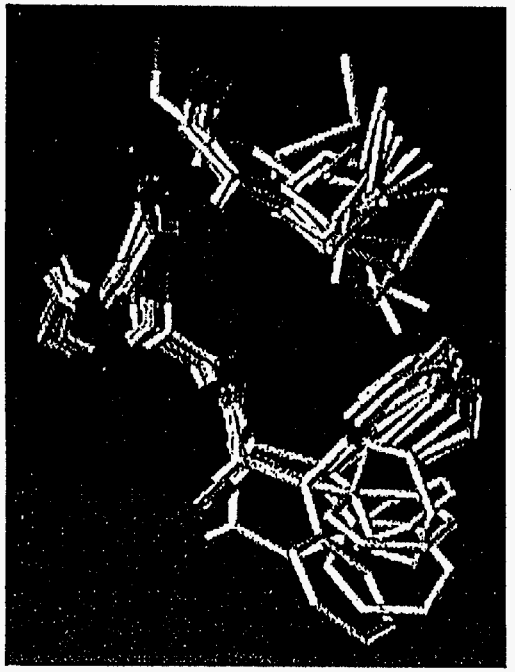


KE

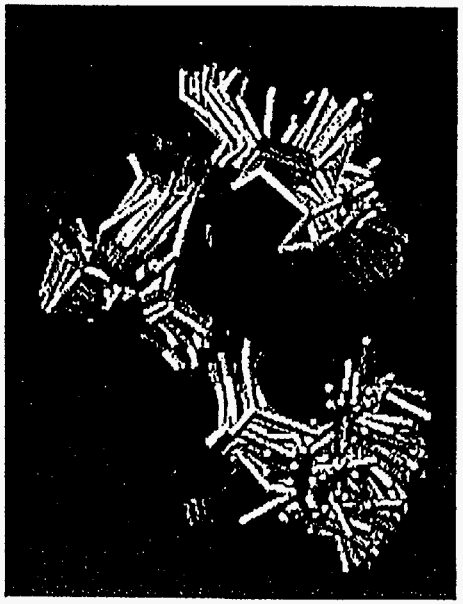


EK

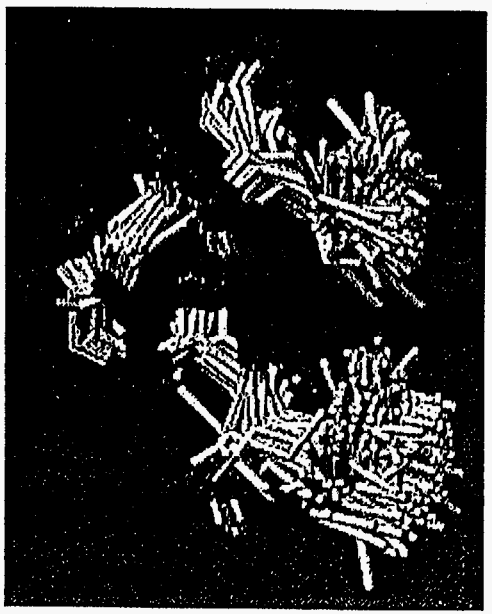
Still more (i, i+4) helical pairs



FM



IL



LL