Ian B.Dodd and J.Barry Egan
Improved detection of helix-turn-helix DNA-binding motifs in protein sequences
Nucleic Acids Research, 1990; 18(17):5019-5026

Improved detection of helix-turn-helix DNA-binding motifs in protein sequences

Ian B.Dodd and J.Barry Egan
Department of Biochemistry, University of Adelaide, Box 498 GPO, Adelaide SA 5001, Australia

Received July 3, 1990; Revised and Accepted August 8, 1990

ABSTRACT

We present an update of our method for systematic detection and evaluation of potential helix-turn-helix DNA-binding motifs in protein sequences [Dodd, I. and Egan, J. B. (1987) *J. Mol. Biol.* 194, 557 – 564]. The new method is considerably more powerful, detecting approximately 50% more likely helix-turn-helix sequences without an increase in false predictions. This improvement is due almost entirely to the use of a much larger reference set of 91 presumed helix-turn-helix sequences. The scoring matrix derived from this reference set has been calibrated against a large protein sequence database so that the score obtained by a sequence can be used to give a practical estimation of the probability that the sequence is a helix-turn-helix motif.

INTRODUCTION

X-ray crystallography of the bacteriophage λ Cro protein (1), the *Escherichia coli* CAP protein (2) and the λ CI protein (3) revealed the protein substructure, known as the helix-turn-helix (HTH) motif, which is responsible for the ability of these proteins to bind in a sequence-specific manner to DNA. Since then, the structures of other HTH motif-containing proteins and a number of the DNA complexes of these proteins have been solved. On the basis of sequence similarity it is evident that a large class of DNA-binding proteins use the HTH motif (4, 5).

Because of the functional importance of the motif, we previously developed a method for the systematic detection and evaluation of potential HTH motifs from protein sequences (6). The method works by measuring the amino acid sequence similarity between a protein segment and a reference set (master set) of aligned HTH motifs. A quantitative score for each segment of a protein under test is obtained using an amino acid *versus* position scoring matrix (weight matrix) derived from amino acid conservations in the master set. The score for a segment is simply the sum of the weights obtained by each amino acid at each position of the segment. Each overlapping segment of a protein under test is scored in this way to find the highest scoring segment, which is the best HTH motif candidate.

The first step in deriving our weight matrix from the master set was to make a frequency matrix, which contained the frequencies of occurrence of each amino acid at each position in the master set. These observed amino acid frequencies were then divided by the frequencies expected on the basis of average amino acid usage in all proteins. The final weight was the natural logarithm of this quotient. Thus, preferred amino acids (occurring more often than average) obtained positive weights, and avoided amino acids (occurring less often than average) obtained negative weights, and the magnitude of the weights was related to the strength of this preference or avoidance.

Our master set of 37 sequences was built up by a process of successive recruitment. We started with a weight matrix derived from a master set of three known HTH motifs. Any protein regions detected significantly by this weight matrix, and which were from known or very likely sequence-specific DNA-binding proteins, were assumed to be HTH motifs and were added to the master set. A new weight matrix was then derived and used to recruit new HTH motifs in the same way, and the process repeated. Each new master set was subjected to the following test: every sequence of the set had to be detected significantly by a weight matrix made from all the other members of the set; sequences failing this test were removed. This ensured reasonable statistical homogeneity of the master set.

The final weight matrix was then calibrated against a protein sequence database to allow a simple, practical estimation of the likelihood that the highest scoring segment of a test protein is a HTH motif. We intended that this estimate be used as a guide for further study of the protein.

Weight matrix methods for HTH motif detection have also been developed by Matthews and coworkers (7, 8, 4) and Drummond *et al.* (9). Some quite different approaches for HTH detection have been described by White (10) and Shestopalov (11). However, calibration against a large protein sequence database, which is necessary for an appreciation of sensitivity and selectivity, is unique to our method.

Firstly, we critically examine some aspects of our procedure of weight matrix development. Secondly, we present our latest and most powerful weight matrix, which is derived from a master set of 91 HTH sequences, and describe its use in detecting and evaluating potential HTH motifs.

WEIGHT MATRIX DEVELOPMENT

Routine significance testing

Full calibration of a weight matrix against a protein sequence database, although ideal for significance testing, is very time-consuming and was therefore only done for the final weight matrix. During development of the weight matrix, however, much statistical testing was required, particularly for deciding

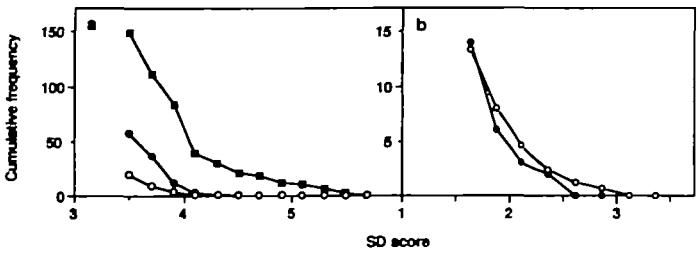

Figure 1. The fit of non-HTH segment scores and highest scores to the normal distribution. The distributions of segment scores (a) and highest scores (b) from 200 non-HTH proteins were examined for fit to the normal distribution. The x-axes give the SD scores (that is, scores expressed in standard deviation units relative to the appropriate mean). The y-axes give the cumulative frequency with summation from high to low. That is, each point indicates the total number of scores with a given SD score or higher. Only those regions of the distributions important for significance testing are shown. Open circles, score frequencies expected on the basis of the normal distribution; closed circles, scores obtained with the weight matrix of Dodd and Egan (6); closed squares, scores obtained with the weight matrix of Brennan *et al.* (4).

WEIGHT MATRIX VARIABLES				NUMBER OF DATABASE PROTEINS SCORING ≥ 3.1 SD			
Weight Matrix	Size of Master Set	Weights	Positions in Motif	HTH	?	non-HTH	Total
1	10	RAW	20	18	16	11	45
2	10	RAW	22	20	11	6	37
3	37	MOD	20	35	22	18	75
4	57	RAW	20	26	16	3	45
5	57	MOD	20	46	22	5	73
6	57	MOD	22	48	19	6	73
7	91	MOD	19	52	12	7	71
8	91	MOD	22	53	15	6	74

Table 1. The performance of different weight matrices. The PIR database – Annotated Release 20 (12) was examined with different weight matrices. The weight matrix variables are described in the text. Matrix 7 is the same as matrix 8 but with positions 7, 12 and 13 set to zero. Database proteins with SD scores of at least 3.1 were classified as HTH, non-HTH or unknown (?) as follows: proteins known or very likely to have sequence-specific DNA-binding activity were classified as HTH; proteins with a known function not involving sequence-specific DNA-binding were classified as non-HTH; proteins of unknown function, or which have functions that may involve sequence-specific DNA binding, were classified as unknown. The scores used for master set proteins in the database were the self-excluded sub-matrix scores. On the basis of the routine significance test we would expect 1/1000 non-HTH proteins to score at least 3.1 SD, that is, approximately 5.5 proteins. The observed numbers of non-HTH proteins for the eight matrices respectively were: 17.1, 8.5, 25.0, 4.7, 7.2, 8.1, 8 4 and 7.5 (the unknowns were distributed proportionately between the HTH and non-HTH classes). Thus, for most matrices, the significance estimates based on the routine test were in reasonable accordance with the results from the full database. The anomalous results may reflect poor representation of the non-HTH proteins in the database, with respect to the matrix used, by the 200 non-HTH proteins used for the routine test.

whether potential HTH sequences were eligible for the master set. For this routine testing, a more rapid method was used. We estimated for any score, the probability that a score at least as high would be obtained by a single non-HTH protein; the lower this probability, the more significant the score.

Previously we used a sequence randomization method for this testing. However, we found that the probability estimates using this method were seriously in error; the proportion of non-HTH proteins in the database scoring significantly was about 10-fold more than expected (6).

An alternative method for significance testing (8, 4) is to compare the score of the highest scoring segment from a test protein to a distribution of all segment scores from non-HTH proteins. To estimate the probability of a single non-HTH segment scoring at least as highly as the test segment, the assumption is made that the non-HTH segment scores are normally distributed. This probability is then used to calculate the probability of finding a segment scoring so highly in a non-HTH protein the size of the test protein. However, we found that the proportion of non-HTH segments with high scores was much greater than that expected from the normal distribution (Figure 1a), such that the use of the normal distribution to calculate probabilities led to a serious overestimation of the significance.

Our present method gave much better significance estimates. In this method, the highest score from the test protein was compared with the *highest* scores from non-HTH proteins. We chose a set of 200 presumed non-HTH proteins from the database in a pseudo-random manner: a larger set of proteins (each with at least 100 amino acids) was randomly chosen and was then pruned to 200 proteins by removing those that were likely to interact with DNA or were of unknown function. For any weight matrix then, the highest score of each of these proteins was obtained and the mean ($X_{non\text{-}HTH}$) and standard deviation ($s_{non\text{-}HTH}$) of these 200 scores calculated. Any score on the weight matrix (X_{test}) could thus be expressed as an SD score (SD score = $(X_{test} - X_{non\text{-}HTH})/s_{non\text{-}HTH}$). For each matrix tested, the highest scores of the non-HTH proteins fitted the normal distribution quite well (Figure 1b). Thus, we could use the normal distribution to get a reasonable estimate of the significance of any SD score. These significance estimates were generally fairly consistent with the results of scanning the full database (see legend to Table 1 and Table 4). This method is used for the rest of this paper.

Increasing the size of the master set

Our wish to increase the size of the master set stemmed from our belief that larger master sets would produce more powerful weight matrices. We tested this by examining the effect of master set size on the ability of the derived weight matrix to detect HTH proteins in the sequence database.

Table 1 shows the results of testing a number of different weight matrices on the protein sequence database. The database used for this paper was the Protein Identification Resource (PIR) *Protein Sequence Database – Annotated Release 20, March 1989* (12). This database contains 5980 sequences, however 488 of these contain sequence ambiguities and were not used, leaving 5492 sequences that were examined. Database proteins with SD scores (using the routine testing method described above) of at least 3.1 ($p = 0.001$) were classified as HTH proteins, non-HTH proteins and unknowns. This classification was based on the possession of sequence-specific DNA-binding activity (details given in Table 1 legend). Of course, one cannot deduce the presence of a HTH motif from a protein's DNA-binding activity; many proteins bind to DNA using non-HTH structures. Therefore, with our classification scheme, proteins that bind to DNA but are not HTH proteins could be wrongly placed in the HTH class. However, assuming that non-HTH DNA-binding proteins score no better than non-HTH proteins in general, there would have to be approximately 1000 non-HTH DNA-binding proteins among the 5492 database proteins (18%) for one to expect a single one scoring at least 3.1 SD.

Our method for master set enlargement was slightly different from that described previously (6). Firstly, as before, the sequences in the master set were from proteins known or very

Table 2. The latest master set of HTH sequences.

Bacterial transcriptional control proteins						
BirA	20	HS	GEQLGETLGM	SRAAINKHIQ	4.2	21
LacI	4	VT	LYDVAEYAGV	SYQTVSRVVN	5.6	20
CytR	10	AT	MKDVALKAKV	STATVSRALM	3.6	22
PurR	2	AT	IKDVAKRANV	STTTVSHVIN	5.2	23
Ka RbtR	4	IT	IYDLAELSGV	SASAVSAILN	5.2	24
DeoR	22	LH	LKDAAALLGV	SEMTIRRDLN	5.8	6
AsnC	23	TA	YAELAKQFGV	SPGTIHVRVE	4.2	6
TrpRAV77	66	MS	QRELKNELGV	GIATITRGSN	2.5	19
CAP	168	IT	RQEIGQIVGC	SRETVGRILK	6.0	20
AraC	195	FD	IASVAQHVCL	SPSRLSHLFR	4.1	20
Eca AraC	200	LR	IDEVARHVCL	SPSRLAHLFR	4.5	25
Fnr	195	MT	RGDIGNYLGL	TVETISRLLG	4.6	20
Ada	100	VT	LEALADQVAM	SPFHLHRLFK	4.0	26
DicA	21	HT	QRSLAKALKI	SHVSVSQWER	6.0	27
DicC	11	GS	KTKLAQAAGI	RLASLYSWKG	2.6	27
Fis	72	GN	QTRAALMMGI	NRGTLRKKLK	4.9	28
LysR	19	GS	LTEAAHLLHT	SQPTVSRELA	5.5	29
MetR	17	GS	LAAAAAVLHQ	TQSALSHQFS	3.6	29
IlvY	16	RH	FGRSARAMHV	SPSTLSRQIQ	5.0	29
Ecl AmpR	21	LS	FTHAAIELNV	THSAISQHVK	5.1	29
Pa TrpI	7	HS	ISLAAEELHV	THGAVSRQVR	4.6	30
Rm NodD1	24	RK	LTAAARRINL	SQPAMSAAIA	3.9	31
Rm NodD2	21	RK	LTAAARRVKL	SQPAMSAAIA	3.8	31
Pp XylS	228	IS	LERLAELAMM	SPRSLYNLFE	4.8	32
Bs XylR	27	IS	RAKLSEMTGL	NKSTVSSQVN	4.7	33
Kp NifA	494	WV	QAKAARLLGM	TPRQVAYRIQ	4.9	9
Rm NifA	511	WN	QAKAARILEK	TPRQVGYALR	3.8	9
Kp NtrC	443	GH	KQEAARLLGW	GRNTLTRKLK	4.5	9
Rm NtrC	450	GN	QIKAADLLGL	NRNTLRKKIR	5.5	34
BR MerR	3	FR	IGELADKCGV	NKETIRYYER	6.3	35
Sa MerR	3	MK	ISELAKACDV	NKETVRYYER	5.7	35
Tn21 MerR	8	LT	IGVFAKAAGV	NVETIRFYQR	4.0	35
NAH7 NahR	21	RR	VSITAENLGL	TQPAVSNALK	3.8	36
RK2 TrfB	35	KP	QATFATSLGL	TRGAVSQAVH	2.6	37
Tn10 TetR	25	LT	TRKLAQKLGV	EQPTLYWHVK	5.7	6
pSC101 TetR	25	LT	TRRLAERLGV	QQPALYWHFK	5.5	6
Recombination/transposition proteins						
Tn21 TnpR	162	EQ	KTKLAREFGI	SRETLYQYLR	6.3	6
Tn501 TnpR	162	EP	KAQLAREFNI	SRETLYQYLR	6.6	6
γδ TnpR	159	LG	ASHISKTMNI	ARSTVYKVIN	3.1	6
Tn3 TnpR	159	TG	ATEIAHQLSI	ARSTVYKILE	4.2	6
Tn2501 TnpR	168	IS	ISAIAREFNT	TRQTILRVKA	4.8	38
P inv.	159	TP	RQKVAIIYDV	GVSTLYKRFP	3.6	6
P1 C inv.	159	IP	RKQVAIIYDV	AVSTLYKKFP	3.7	6
St H inv.	160	HP	RQQLAIIFGI	GVSTLYRYFP	5.9	6
IS1 InsA	64	VG	CRATARIMGV	GLNTIFRHLK	3.5	39
λ Nu1	3	VN	KKQLADIFGA	SIRTIQNWQE	4.3	6
Mu B	19	TT	FKQIALESGL	STGTISSFIN	3.4	40
Bacteriophage transcriptional control proteins						
λ Cro	14	FG	QTKTAKDLGV	YQSAINKAIH	4.0	6
λ CI	32	LS	QESVADKMGM	GQSGVGALFN	4.6	6
λ CII	24	LG	TEKTAEAVGV	DKSQISRWKR	5.5	6
434 CI	16	LN	QAELAQKVGT	TQQSIEQLEN	5.6	6
434 Cro	17	MT	QTELATKAGV	KQQSIQLIEA	3.7	6
P22 C2	19	IR	QAALGKMVGV	SNVAISQWER	6.1	6
P22 C1	24	RG	QRKVADALGI	NESQISRWKG	7.2	6
P22 Cro	11	GT	QRAVAKALGI	SDAAVSQWKE	6.4	6
ϕ80 CI	16	LK	QRDLAEALST	SPQTVNNWIK	6.3	41
ϕ80 gp30	21	GS	HKVLAEKVGV	TPQQAINMLK	3.3	41
16-3 C	4	IT	QAELARRVGQ	SQQAINNLFA	6.7	42
ϕ105 repr.	19	LT	QVQLAEKANL	SRSYLADIER	4.3	43
P2 C	18	LS	RQQLADLTGV	PYGTLSYYES	4.7	6
P2 Cox	12	IP	YQEFAKLIGK	STGAVRRMID	3.1	44
186 Apl	15	VT	LQQFAELEGV	SERTAYRWTT	5.5	45
Mu Ner	24	LS	LSALSRQFGY	APTTLANALE	2.7	46
Mu C	98	RN	MEELRRQYRL	SQPQIYQIIA	4.2	47
D108 Ner	26	MS	LAELGRSNHL	SSSTLKNALD	2.6	46
RNA polymerase sigma factors						
RpoD	571	YT	LEEVGKQFDV	TRERIRQIEA	5.1	6
HtpR	251	ST	LQELADRYGV	SAERVRQLEK	5.7	6
Bs RpoD	330	RT	LEEVGKVFGV	TRERIRQIEA	5.6	6
Bs SpoIIAC	219	QT	QSEVAERLGI	SQVQVSRLEK	7.1	48
Bs SigB	222	KS	QKETGDILGI	SQMHVSRLQR	5.2	49
Bs SpoIIG	204	KT	QKDVADMMGI	SQSYISRLEK	7.4	6
Bs Spo0A	196	VL	YPDIAKKFNT	TASRVERAIR	2.7	50
Kp NtrA	365	MV	LADIAQAVEM	HESTISRVTT	4.6	51
Rm NtrA	6	LN	LRIVADAIKM	HESTVSRVTS	4.2	51
Av NtrA	390	LV	LHDIAEAVGM	HESTISRVTT	5.1	51
DNA replication/chromosome partition proteins						
R6K repl.	57	IR	AEDLAALAKI	TPSLAYRQLK	3.4	52
pSC101 repl.	50	FT	YNQYAQMMNI	SRENAYGVLA	2.5	53
P1 ParB	166	MS	QKDIAAKEGL	SQAKVTRALQ	5.1	54
Mini-F D	161	TG	RTEKARIWEV	TDRTVRTWIG	3.4	55
F SopB	9	GN	ISALADAENI	SRKIITRCIN	3.7	56
Eukaryotic homeobox proteins						
Dm prd	241	YT	REELAQRTNL	TEARIQVWFS	5.3	57
Dm eve	98	PR	RCELAAQLNL	PESTIKVWFQ	5.5	57
Dm en	482	RR	RQQLSSELGL	NEAQIKIWFQ	4.9	58
Dm zen1	29	TR	RIEIAQRLSL	CERQVKIWFQ	4.7	57
Dm Antp	325	RR	RIEIAHALCL	TERQIKIWFQ	5.7	58
Sc MATa1	98	KE	KEEVAKKCGI	TPLQVRVWFI	4.8	6
Sp Pi	134	SE	FYDLSAATGL	TRTQLRNWFS	3.2	59
Sc PHO2	105	VE	RKKISDLIGM	PEKNVRIWFQ	3.2	58
Human Oct-1	133	EE	ITMIAFQLNM	EKEVIRVWFC	3.8	60
Ce Unc-86	399	ER	IASIADRLDL	KKNVVRVWFC	3.5	60
Rat Pit-1	242	QE	IMRMAEELNL	EKEVVRVWFC	3.0	60

Table 2. The latest master set of HTH sequences. Each master set entry consists of the protein name, the position in the protein of the first residue of the motif, the HTH motif sequence, the SD score of the sequence using the self-excluded sub-matrix, and a reference. The references tend to be for papers which present evidence for DNA binding or for review articles, in which references for the sequences may be found. Unless otherwise indicated, proteins are from *E. coli*. Abbreviations: Av, *Azotobacterium vinelandii*; Bs, *Bacillus subtilis*; BR, *Bacillus* RC607; Ce, *Caenorhabditis elegans*; Dm, *Drosophila melanogaster*; Eca, *Erwinia carotovora*; Ecl, *Enterobacter cloacea*; inv., invertase; Ka, *Klebsiella aerogenes*; Kp, *Klebsiella pneumoniae*; Pa, *Pseudomonas aeruginosa*; Pp, *Pseudomonas putida*; repl., replication initiator; repr., repressor; Rm, *Rhizobium meliloti*; Sa, *Staphylococcus aureus*; Sc, *Saccharomyces cerevisiae*; Sp, *Saccharomyces pombe*; St, *Salmonella typhimurium*. Most sequences were extracted from the PIR Protein Database (12).

As the $\bar{X}_{non\text{-}HTH}$ and $s_{non\text{-}HTH}$ values varied little between different self-excluded sub-matrices, the averages of these statistics from five different sub-matrices were used in calculation of the SD scores in order to save computing time.

The similarity criterion for master set membership was as follows. A master set sequence accumulated similarity points for every other master set sequence with which it shared a direct amino acid homology of at least 50%: 50–60%, 1 point; 60–70%, 2 points; 70–80%, 3 points; 80–90%, 4 points; 90–95%, 5 points; 95–100%, 6 points. All the master set sequences were compared with each other and if any sequences obtained more than 5 points, then sequences were removed until all sequences scored 5 points or less.

a. Frequency matrix.

Amino Acid	Motif Position -2	-1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Totals
A	2	1	3	14	10	12	75	6	15	9	1	1	4	3	8	15	4	4	4	11	0	10	212
C	0	0	1	1	0	0	0	0	0	3	3	1	1	0	0	0	0	0	0	1	0	3	14
D	0	1	0	1	14	0	0	14	1	0	5	0	1	2	0	0	0	0	1	1	0	2	43
E	4	5	0	11	26	0	0	16	9	3	3	0	3	12	13	0	0	2	0	1	13	6	127
F	4	0	4	0	0	4	0	1	0	10	0	0	0	0	1	0	0	1	1	1	22	0	49
G	9	7	1	4	0	0	8	0	0	0	50	0	6	0	7	1	0	3	1	1	0	4	102
H	4	3	1	1	2	0	0	3	2	0	5	0	3	3	0	2	0	2	4	5	0	2	42
I	10	0	13	3	2	15	0	4	9	4	0	17	0	2	0	1	31	1	4	8	16	1	141
K	4	4	6	11	12	1	1	14	11	0	5	2	2	7	2	1	0	5	8	4	5	15	120
L	16	1	17	0	1	35	0	3	12	31	0	22	0	2	1	1	22	1	1	12	20	0	198
M	7	0	2	1	1	1	0	0	5	7	1	10	0	0	2	0	2	0	0	2	0	1	42
N	0	8	0	1	0	0	0	2	1	1	14	0	8	1	4	2	0	4	9	0	0	11	66
P	1	6	0	1	0	0	0	0	0	0	0	0	3	13	7	0	0	0	0	0	0	3	34
Q	2	1	21	9	11	0	0	9	8	0	0	2	1	17	7	12	0	3	12	5	3	9	132
R	9	10	14	9	5	0	1	16	10	0	1	0	1	17	8	7	0	17	28	3	0	16	172
S	2	17	0	8	4	1	6	1	2	2	3	0	37	1	25	5	0	29	3	0	1	5	152
T	6	24	3	12	1	5	0	2	2	4	0	5	20	4	3	39	0	4	1	0	4	3	142
V	7	3	1	1	2	16	0	0	2	12	0	29	0	5	3	3	32	0	7	8	7	0	138
W	2	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	2	21	0	0	27
Y	2	0	4	3	0	1	0	0	2	4	0	1	1	2	0	2	0	15	5	7	0	0	49
	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	91	2002

b. Weight matrix.

Amino Acid	Motif position -2	-1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	-125	-194	-84	70	36	54	238	-15	77	26	-194	-194	-56	-84	14	77	-56	-56	-56	46	-195	36
C	-64	-64	-63	-63	-64	-64	-64	-64	-64	47	47	-63	-63	-64	-64	-64	-64	-64	-64	-63	-64	47
D	-156	-154	-156	-154	109	-156	-156	109	-154	-156	6	-156	-154	-85	-156	-156	-156	-156	-154	-154	-156	-85
E	-31	-9	-171	70	156	-171	-171	107	50	-60	-60	-171	-60	78	86	-171	-171	-101	-171	-170	86	9
F	10	-130	10	-130	-130	10	-130	-129	-130	102	-130	-130	-130	-130	-129	-130	-130	-129	-129	-129	180	-130
G	30	5	-190	-51	-191	-191	18	-191	-191	-191	202	-191	-10	-191	5	-190	-191	-80	-190	-190	-191	-51
H	62	33	-76	-76	-7	-78	-78	33	-7	-78	84	-78	33	33	-78	-7	-78	-7	62	84	-78	-7
I	75	-156	101	-45	-86	116	-156	-16	65	-16	-156	128	-156	-86	-156	-155	188	-155	-16	53	122	-155
K	-31	-31	10	70	79	-170	-170	94	70	-171	-9	-100	-100	25	-100	-170	-171	-9	38	-31	-9	101
L	66	-212	72	-213	-212	144	-213	-102	37	132	-213	97	-213	-142	-212	-212	97	-212	-212	37	88	-213
M	122	-74	-3	-73	-73	-73	-74	-74	88	122	-73	158	-74	-74	-3	-74	-3	-74	-74	-3	-74	-73
N	-137	72	-137	-136	-137	-137	-137	-67	-136	-136	128	-137	72	-136	2	-67	-137	2	84	-137	-137	104
P	-156	23	-157	-156	-157	-157	-157	-157	-157	-157	-157	-157	-46	101	39	-157	-157	-157	-157	-157	-157	-46
Q	-60	-130	175	90	110	-131	-131	90	78	-131	-131	-60	-130	154	65	119	-131	-20	119	31	-20	90
R	65	76	110	65	7	-155	-154	123	76	-155	-154	-155	-154	129	54	40	-155	129	179	-45	-155	123
S	-118	96	-188	21	-48	-187	-8	-187	-118	-118	-77	-188	174	-187	135	-26	-188	150	-77	-188	-187	-26
T	11	149	-59	80	-169	-8	-170	-99	-99	-30	-170	-8	131	-30	-59	198	-170	-30	-169	-170	-30	-59
V	17	-67	-177	-177	-108	100	-178	-178	-108	71	-178	160	-178	-16	-67	-67	169	-178	17	31	17	-178
W	44	-26	-26	-26	-26	-26	-26	-26	-26	-25	-26	-25	-26	-26	-26	-26	-26	-26	44	279	-26	-26
Y	-40	-110	30	1	-110	-109	-110	-110	-40	30	-110	-109	-109	-40	-110	-40	-110	162	52	86	-110	-110

Table 3. Frequency and weight matrices derived from the latest master set. The numbering of the motif positions follows the convention of Pabo and Sauer (20). The method of calculation of weights from the frequencies is summarized in the text and has been described fully previously (6). The expected proportions of each amino acid were the proportions of the amino acids in the PIR database – Annotated release 15 (12): A, 0.07657; C, 0.02064; D, 0.05150; E, 0.06016; F, 0.03980; G, 0.07322; H, 0.02361; I, 0.05181; K, 0.05995; L, 0.09125; M, 0.02272, N, 0.04298; P, 0.05218; Q, 0.04024; R, 0.05146; S, 0.07121; T, 0.05933; V, 0.06460; W, 0.01413; Y, 0.03271. The mean and standard deviation of the highest scores of the 200 non-HTH proteins were 238.71 ($\bar{X}_{non\text{-}HTH}$) and 293.61 ($s_{non\text{-}HTH}$).

Table 4. Calibration of the latest weight matrix on the database

SCORE CLASS	DATABASE PROTEINS Classification HTH	?	non-HTH	Non-HTH frequency Observed	Expected	HTH %
7.0	2	0	0	0	0	100
6.5	1	0	0	0	0	100
6.0	7	2	0	0	0	100
5.5	11	2	0	0	0	100
5.0	5	1	0	0	0	100
4.5	8	1	0	0	0	100
4.0	9	1	1	1.1(1.1)	0.2(0.2)	90
3.5	5	4	2	3.1(4.2)	1.1(1.3)	71
3.0	6	4	6	8.0(12.2)	6.1(7.4)	50
2.5	5	7	15	20.3(32.5)	26.7(34.1)	25

Table 4. Calibration of the latest weight matrix on the protein sequence database. Proteins in the database scoring at least 2.5 SD with the latest weight matrix were grouped into 0.5 SD score classes (for example, the 2.5 SD score class is 2.500−2.999). The proteins were classified as HTH, non-HTH or unknown (?), as described in the legend to Table 1. Forty seven proteins of the master set are present in the database; the scores used for these proteins were the self-exclusion sub-matrix scores. The other proteins are listed in Table 5a.

The expected frequency of non-HTH proteins is the number expected for each SD score class on the basis of the normal distribution. The observed frequency was calculated by distributing the unknown proteins proportionately between the HTH and non-HTH classes. The numbers in parentheses are the cumulative totals. The % HTH column gives the percentage of HTH over the sum of HTH and non-HTH.

likely to be sequence-specific DNA-binding proteins, or from their close relatives. Secondly, each sequence was significantly detectable using a weight matrix made from the other members of the master set (self-excluded sub-matrix). The new routine significance test was used for this with a 2.5 SD significance threshold ($p = 0.006$). A third, new, criterion was that any sequence could not be too similar to other individual sequences in the master set (details given in the legend to Table 2). This was an attempt to reduce biases in the weight matrix caused by over-representation of some protein families, a criticism raised by Argos (13).

Three comparisons in Table 1 showed the effect of master set size. Increasing the size of the master set from 10 to 57, from 37 to 57 and from 57 to 91 (Table 1, matrices 1 *versus* 4, 3 *versus* 5, and 6 *versus* 8) increased the number of HTH proteins detected (true positives) without increasing, and in two cases decreasing, the number of non-HTH proteins detected (false positives). Thus, our procedure of master set enlargement seems to be justified.

Derivation of weights

We also examined our procedure of deriving weights from the master set. As described, we modified the raw amino acid frequencies by normalizing them against expected amino acid occurrence. An alternative is to simply use the raw frequencies as weights (7, 8, 4). From examination of the database, it appears that the modification step markedly improves the performance of the weight matrix. The modified weight matrix from a 57 member master set detected 20 extra true positives and only two extra false positives compared with the raw weight matrix (Table 1, matrices 5 *versus* 4).

The number of motif positions

The last variable we examined was the number of positions in the motif used for detection. We (6) originally used a motif length of 20 residues. However, we have noticed some amino acid preferences outside this region, which were potentially useful for detection purposes. The weight matrices used by Matthews and coworkers include two extra positions on the carboxy-terminal side (4, 7, 8). Weight matrices using this 22 position motif seemed slightly better at detecting HTH proteins (Table 1, matrices 1 *versus* 2, and 5 *versus* 6), and we have since adopted this motif length. We also tested the suggestion of Yudkin (14) that removal of positions 7, 12 and 13 from consideration might improve detection. We found that it did not (Table 1, matrices 7 *versus* 8), although the small size of the difference emphasizes that these positions are contributing little to detection. We have not tested other motif lengths.

The latest weight matrix and its use

Our latest master set, comprising 91 presumed HTH motif sequences, is shown in Table 2. Forty seven of the master set proteins are listed in the database. Note that the MetR protein is not the same as the Met repressor (MetJ) which was recently shown (15) not to contain a HTH motif (nor is it detected by our method−see Table 5b). Functional classes added since our previous master set (6) are the DNA replication/chromosome partition group and the eukaryotic homeobox group. Evidence for the existence of a HTH motif in the Drosophila Antp homeobox has been provided by NMR (16). Proteins in the master set tend, perhaps unduly, to assist detection of strongly related proteins. For example, once one homeobox protein entered the master set, other homeobox proteins followed readily. However, although the master set does contain some strongly related subgroups, there is significant similarity between the groups. For example, a weight matrix made from a master set from which all of the homeobox proteins was removed was able to detect 5 of the 11 homeobox proteins at the 2.5 SD level.

The frequency matrix and the weight matrix are shown in Table 3. In terms of raw frequencies, the most striking conservations are those at positions 5, 9 and 15. However, it is clear from the weight matrix that there are strong amino acid preferences and avoidances at most positions. This topic is discussed in more detail elsewhere (17).

According to other authors, residues other than glycine are unlikely to occur at position 9 for structural reasons (7, 11). However, 41 of our master set motifs do not have glycine at this position. In other respects our master set sequences are reasonably consistent with the structural criteria of these authors. If the glycine requirement is ignored, 87 of the sequences comply with the rules of Ohlendorf *et al.* (7) and 80 of the sequences fit the more stringent template of Shestopalov (11).

The data in Table 1 show that the new weight matrix (matrix 8) is more powerful than the matrix of Brennan and Matthews (4, matrix 2) and our previous (6) weight matrix (matrix 3), detecting considerably more true positives without detecting more false positives. Of the 19 true positives detected at the 3.1 SD level by the new weight matrix that were not detected by our previous weight matrix (one protein is detected by the old but not the new matrix), there are 5 proteins of the bacterial transcriptional protein class, 3 recombination/transpositon proteins, 1 sigma factor, 2 replication/partition proteins and 8 homeobox proteins.

The matrix detects all known HTH motifs: λ Cro, CAP, λ CI, 434 Cro, 434 CI, and LacI (5) except for TrpR (see Table 5b). This inability to detect the TrpR HTH sequence has been raised

as a failing of our method (4, 14). However, we believe that the TrpR HTH motif is a special case. The TrpR protein only functions as a repressor when complexed with tryptophan, and it appears that the TrpR HTH motif contains atypical residues that allow it to respond to the presence and absence of tryptophan. From a comparison of the TrpR structure with and without bound tryptophan it was concluded that the alanine at position 10 of the motif allows space for the motif to fall back towards the core

Table 5. Scores of selected proteins.

(a) Database proteins scoring ≥ 2.5 SD.

7.1	2	GalR (H)	3.2	224	Mouse polyoma virus VP2/3 (?)
6.6	25	Tn1721 TetR (H)	3.1	158	Mu G inv. (H)
6.5	24	T4 rIIB (?)	3.0	160	Rabbit α tropomyosin (N)
6.3	36	Xl MM3 homeo. (H)	3.0	211	MetL (N)
6.1	102	T7 hyp. gp7.7 (?)	3.0	575	RSV pol (?)
6.1	2	EbgR (H)	3.0	43	Glutaredoxin (N)
6.0	36	Xl AC1 homeo. (H)	3.0	3	Mp clplast. ribo. S3 (N)
5.9	285	Dm ftz (H)	2.9	96	Reovirus(2) haemagglutinin (N)
5.9	17	λ nin B-68 hyp. (?)	2.9	134	SBMV coat (N)
5.5	96	*leuABCD* hyp. 133 (?)	2.8	170	F TraJ (?)
5.5	47	Mouse m6 homeo. (H)	2.8	109	T7 gp2.8 hyp. (?)
5.2	36	Dm Ubx (H)	2.8	120	T7 internal virion B (?)
5.1	40	F SopA (?)	2.8	6	Mouse tumour antigen p53 (H)
4.7	365	DnaB (?)	2.8	17	Rat T kininogen LMW pc. (N)
4.6	36	Mouse Mo-10 homeo. (H)	2.8	82	Mastadenovirus h7 pIVa2 (?)
4.1	6	P22 Xis (?)	2.7	788	Ce myosin heavy chain (N)
4.0	178	α galactosidase (N)	2.7	156	Lc lactate dehydrogenase (N)
3.9	35	Protein A (?)	2.7	73	T3 adenosyl met. tfrase. (N)
3.9	40	T4 57A hyp. (?)	2.7	717	β galactosidase (N)
3.8	196	*uvrC* hyp. 28K (?)	2.6	137	Chicken α tropomyosin (N)
3.7	8	Mp clplast. ribo. L22 (N)	2.6	128	Pm γ fibrinogen pc. (N)
3.6	68	P22 Abc1 (?)	2.6	49	λ H (N)
3.6	279	Rat serum albumin (N)	2.6	334	VEEV structural (N)
3.4	305	RecN (?)	2.6	722	Simian 11 rotavirus VP3 (N)
3.4	153	M5 (AppY) (H)	2.6	100	PtsP (N)
3.4	259	Sc met-tRNA syn. (N)	2.5	411	T4 DNA polymerase (?)
3.4	545	Ma retrovirus-rel. pol (?)	2.5	374	Papilloma virus 8 hyp. E2 (?)
3.2	1658	Hs V. W. factor pc. (N)	2.5	254	Vaccinia virus WR h7 (?)
3.2	139	ATPase N (KdpC)	2.5	142	MS2 RNA/RNA polymerase (N)

Table 5. Scores of selected proteins.

(b) Other proteins

8.1	93	Bs SpoIIIC	2.6	266	Hs excision repair
7.0	87	Dm transposon mariner	2.5	478	RpoD
6.1	23	186 CP76 (CII)	2.4	5	SP01 TF1
5.6	21	Hs centromere CENP-B	2.2	66	TrpR
5.2	6	MalI	2.2	851	MalT
5.2	16	Sm hyp. ORF3	2.1	5	SP01 sigma gp28
5.1	15	Bs Sin hyp.	2.1	300	T7 RNA polymerase
4.5	205	Sg hyp. StrR	2.0	269	λ and 434 Int
4.4	27	Bs hyp. GerE	2.0	17	CysB
4.3	28	P4 orf88	1.9	160	Sc MATα2
4.1	267	Nc mito. hyp. URF1 intron	1.8	192	OmpR
3.7	153	UhpA	1.8	18	IHF-A
3.1	206	Vh LuxB	0.7	442	DnaA
3.1	280	Sc RNA polymerase 40K	0.7	1	IHF-B
3.0	237	Bs RpoD	0.4	24	186 CI
2.9	266	Mouse excision repair	0.3	87	MetJ
2.8	41	Hs plasminogen act. inhibitor-2	0.2	580	SV40 large T antigen
2.8	169	RK2 KorB	0.1	46	Mu repressor
2.7	77	pNE131 and pIM13 repl.	−1.0	26	LexA
2.7	185	Bl Spo0H	−1.3	27	P22 Mnt
2.7	180	Bs Spo0H	−1.9	14	P22 Arc

Table 5. Scores of selected proteins. Each entry consists of the SD score, the position in the protein of the highest scoring segment and the name of the protein. (a) Database proteins scoring ≥ 2.5 SD on the latest weight matrix. The HTH (H), non-HTH (N) and unknown (?) classification of these proteins, used for Table 4, is indicated. M5 (AppY) and mouse tumour antigen p53 are DNA-binding (61, 62). Master set proteins in the database are not listed. (b) Scores of other proteins. The sequences were taken from the PIR Annotated or Preliminary databases (12), or from references given in (6) except for MalI (63) and P4 orf88 (64). More than 20 homeobox proteins from the Preliminary database scored above 2.5 SD but are not listed. The entries for RpoD and Bs RpoD are for second potential HTH motifs. The SD score for TrpR on the matrix from the master set with TrpRAV77 removed is 1.3.

Proteins are from *E. coli* unless otherwise indicated. Abbreviations are as in Table 2, except for: act., activator; Bl, *Bacillus licheniformis*; cplast., chloroplast; homeo., homoebox; hyp., hypothetical; Hs, *Homo sapiens*; Lc, *Lactobacillus casei*; LMW, low molecular weight; Ma, *Mesocricetus auratus*; met., methionine; mito., mitochondrial; Mp, *Marchantia polymorpha*; Nc, *Neurospora crassa*; pc., precursor; Pm, *Petromyzon marinus*; rel., related; ribo., ribosomal; Sg, *Streptomyces griseus*; Sm, *Streptococcus mutans*; SBMV, Southern bean mosaic virus; syn., synthetase; tfrase., transferase; VEEV, Venezuelan equine encephalitis virus; Vh, *Vibrio harveyi*; V.W., Von Willebrand; Xl, *Xenopus laevis*.

of the protein in the absence of tryptophan, and that the glycine at position 18 provides room for a tryptophan binding pocket behind the recognition helix (18). Our weight matrix penalizes these amino acid occurrences very strongly, in fact they obtain the lowest scores of all the positions in the TrpR motif. A mutant of TrpR, TrpRAV77, in which the alanine at position 10 is replaced by valine, is able to act as a repressor in the absence of tryptophan (19). This mutant motif obtains a statistically significant score and is a member of the master set. The other known HTH protein requiring a cofactor for efficient DNA binding, CAP, is detected sigificantly with our weight matrix (6.0 SD, Table 2) but not with the matrix of Brennan and Matthews (4) when our statistical criteria are used (0.73 SD). It is not known whether residues in the CAP HTH motif are involved in responding to cyclic-AMP.

The calibration of the weight matrix against the database is shown in Table 4. Database proteins scoring above 2.5 SD have been grouped according to their SD score and classified as HTH, non-HTH or unknown, as described above. The number of non-HTH DNA-binding proteins scoring above 2.5 SD, and thus wrongly classified as HTH, is likely to be very small. This is because on the basis of the normal distribution, there would have to be 161 non-HTH sequence-specific DNA-binding proteins in the database (3%) for one of them to be expected to score at this level ($p = 0.006$). Below 2.5 SD, the likely increasing contribution of non-HTH DNA-binding proteins to the HTH class makes this classification risky. The proteins in each class are listed in Table 5a. There was a reasonable agreement between the number of false positives observed in each class and the number expected on the basis of the SD score (Table 4).

The most important information in Table 4 is the percentage of HTH proteins in each score class. This percentage can be used as a practical estimate, for any score above 2.5 SD, of the likelihood that a protein contains a HTH motif.

Thus, the procedure for testing a protein for a HTH motif with our method is as follows. Firstly, find the score of the segment of the protein that scores highest on the weight matrix. A computer is ideal for this, but using a calculator to obtain scores for a few candidate segments found by visual inspection of the sequence is a reasonable alternative. We are happy to analyse sequences sent to us (preferably in computer readable form). Our current INTERNET address is jegan@boffin.ua.oz.au . Secondly, convert the score to an SD score [SD score = (score−238.71)/293.61]. Thirdly, use the HTH% column of Table 4 to obtain the likelihood that the segment is a HTH motif.

The probability estimate gained using our method is based solely on the amino acid sequence; functional information about the protein can be used to subjectively adjust the estimate. For example, if a protein is extracellular, then the HTH estimate should probably be adjusted downwards; if a protein is known to regulate gene expression, then the estimate should probably be adjusted upwards. As we have argued, if a protein is known to be a sequence-specific DNA-binding protein, then a score of at least 2.5 SD indicates that it is almost certainly a HTH protein.

The scores for some selected proteins are shown in Table 5. Many of these are sequence-specific DNA-binding proteins that do not score highly with our method.

ACKNOWLEDGEMENTS

We are grateful to Brian Matthews for comments on the manuscript and to our funding sources. I.B.D held a Commonwealth Postgraduate Research Award and J.B.E is supported by a Program Grant from the Australian Research Council.

REFERENCES

1. Anderson, W. F., Ohlendorf, D. H., Takeda, Y. and Matthews, B. W. (1981) *Nature (London)*, **290**, 754–758.
2. McKay, D. B. and Steitz, T. A (1981) *Nature (London)*, **290**, 744–749.
3. Pabo, C. O. and Lewis, M. (1982) *Nature (London)*, **298**, 443–447.

4. Brennan, R. G. and Matthews, B. W. (1989) *J. Biol. Chem.* **264**, 1903–1906.
5. Brennan, R. G. and Matthews, B. W. (1989) *Trends Biochem. Sci.* **14**, 286–300.
6. Dodd, I. B. and Egan, J. B. (1987) *J. Mol. Biol.* **194**, 557–564.
7. Ohlendorf, D. H., Anderson, W. F. and Matthews, B. W. (1983) *J. Mol. Evol.* **19**, 109–114.
8. Brennan, R. G., Weaver, L. H. and Matthews, B. W. (1986) *Chemica Scripta*, **26B**, 251–255.
9. Drummond, M., Whitty, P. and Wootton, J. (1986) *EMBO J.* **5**, 441–447.
10. White, S. W. (1987) *Protein Eng.* **1**, 373–376.
11. Shestopalov, B. V. (1988) *FEBS Lett.* **233**, 105–108.
12. George, D. S., Barker, W. C. and Hunt, L. T. (1986) *Nucleic. Acids Res.* **14**, 11–16.
13. Argos, P. (1989) In Creighton, T. E. (ed.), Protein Structure–A Practical Approach. IRL Press, Oxford, pp. 49–78.
14. Yudkin, M. D. (1987) *Protein Eng.* **1**, 371–372.
15. Rafferty, J. B., Somers, W. S., Saint-Girons, I. and Phillips, S. E. V. (1989) *Nature (London)*, **341**, 705–710.
16. Otting, G., Qian, Y., Müller, M., Affolter, M., Gehring, W. and Wüthrich, K. (1988) *EMBO J.* **7**, 4305–4309.
17. Dodd, I. B. and Egan, J. B. (1988) *Protein Eng.* **2**, 174–175.
18. Zhang, R. -g., Joachimiak, A., Lawson, C. L., Schevitz, R. W., Otinowski, Z. and Sigler, P. B. (1987) *Nature (London)*, **327**, 591–597.
19. Kelley, R. L. and Yanofsky, C. (1985) *Proc. Nat. Acad. Sci., U.S.A.* **82**, 483–487.
20. Pabo, C. O. and Sauer, R. T. (1984) *Ann. Rev. Biochem.* **53**, 293–321.
21. Buoncristiani, M. R., Howard, P. K. and Otsuka A. J. (1986) *Gene* **44**, 255–261.
22. Valentin-Hansen, P., Larsen, J. E. L., Højrup, P. H., Short, S. A. and Barbier, C. S. (1986) *Nucl. Acids Res.* **14**, 2215–2228.
23. Rolfes, R. J. and Zalkin, H. (1988) *J. Biol. Chem.* **263**, 19653–19661.
24. Wu, J., Anderton-Loviny, T., Smith, C. A. and Hartley, B. S. (1985) *EMBO J.* **4**, 1339–1344.
25. Lei, S-P., Lin, H-C., Heffernan, L. and Wilcox, G. (1985) *J. Bacteriol.* **164**, 717–722.
26. Teo, I., Sedgewick, B., Kilpatrick, M. W., McCarthy, T. V. and Lindahl, T. (1986) *Cell*, **45**, 315–324.
27. Béjar, S., Bouché, F. and Bouché, J-P. (1988) *Mol. Gen. Genet.* **212**, 11–19.
28. Koch, C., Vandekerckhove, J. and Kahmann, R. (1988) *Proc. Nat. Acad. Sci., U.S.A.* **85**, 4237–4241.
29. Henikoff, S., Haughn, G. W., Calvo, J. M. and Wallace, J. C. (1988) *Proc. Nat. Acad. Sci., U.S.A.* **85**, 6602–6606.
30. Chang, M., Hadero, A. and Crawford, I. P. (1989) *J. Bacteriol.* **171**, 172–183.
31. Fisher, R. F., Egelhoff, T. T., Mulligan, J. T. and Long, S. R. (1988) *Genes Dev.* **2**, 282–293.
32. Inouye, S., Nakazawa, A. and Nakazawa, T. (1986) *Gene* **44**, 235–242.
33. Kreuzer, P., Gärtner, D., Allmansberger, R. and Hillen, W. (1989) *J. Bacteriol.* **171**, 3840–3845.
34. Szeto, W. W., Nixon, B. T., Ronson, C. W. and Ausubel, F. M. (1987) *J. Bacteriol.* **169**, 1423–1432.
35. Helmann, J. D., Wang, Y., Mahler, I. and Walsh, C. T. (1989) *J. Bacteriol.* **171**, 222–229.
36. You, I. -S., Ghosal, D. and Gunsalas, I. C. (1988) *J. Bacteriol.* **170**, 5409–5415.
37. Thomas, C. M. and Smith, C. M. (1986) *Nucl. Acids Res.* **14**, 4453–4469.
38. Michiels, T., Cornelis, G., Ellis, K. and Grinsted, J. (1987) *J. Bacteriol.* **169**, 624–631.
39. Zerbib, T., Jakowec, M., Prentki, P., Galas, D. J. and Chandler, M. (1987) *EMBO J.* **6**, 3163–3169.
40. Miller, J. L., Anderson, S. K., Fujita, D. J., Chaconas, G., Baldwin, D. L. and Harshey, R. M. (1984) *Nucl. Acids Res.* **12**, 8627–8636.
41. Ogawa, T., Ogawa, H. and Tomizawa, J. (1988) *J. Mol. Biol.* **202**, 537–550.
42. Dallmann, G., Papp, P. and Orosz, L. (1987) *Nature (London)*, **330**, 398–401.
43. Van Kaer, L., Gansemans, Y., Van Montagu, M. and Dhaese, P. (1988) *EMBO J.* **7**, 859–866.
44. Saha, S., Haggård-Ljungquist, E. and Nordström, K. (1989) *Proc. Nat. Acad. Sci., U.S.A.* **86**, 3973–3977.
45. Dodd, I. B., Kalionis, B. and Egan, J. B. (1990) *J. Mol. Biol.* (in press).
46. Kukolj, G., Tolias, P. P., Autexier, C. and DuBow, M. S. (1989) *EMBO J.* **8**, 3141–3148.
47. Bölker, M., Wulczyn, F. G. and Kahmann, R. (1989) *J. Bacteriol.* **171**, 2019–2027.
48. Yudkin, M. D. (1987a) *J. Gen. Microbiol.* **133**, 475–481.
49. Binnie, C., Lampe, M. and Losick, R. (1986) *Proc. Nat. Acad. Sci., U.S.A.* **83**, 5943–5947.
50. Ferrari, F. A., Trach, K., LeCoq, D., Spence, J. and Ferrari, E. (1985) *Proc. Nat. Acad. Sci., U.S.A.* **82**, 2647–2651.
51. Merrick, M. J., Gibbins, J. and Toukdarian, A. (1987) *Mol. Gen. Genet.* **210**, 323–330.
52. Mukherjee, S., Erikson, H. and Bastia, D. (1988) *Cell*, **52**, 375–383.
53. Vocke, C. and Bastia, D. (1983) *Cell*, **35**, 495–502.
54. Davis, M. A. and Austin, S. J. (1988) *EMBO J.* **7**, 1881–1888.
55. Lane, D., de Feyter, R., Kennedy, M., Phua, S-H. and Semon, D. (1986) *Nucl. Acids Res.* **14**, 9713–9728.
56. Mori, H., Oshima, A., Ogura, T. and Hiraga, S. (1986) *J. Mol. Biol.* **192**, 1–15.
57. Hoey, T. and Levine, M. (1988) *Nature (London)*, **332**, 858–861.
58. Bürglin, T. R. (1988) *Cell*, **53**, 339–340.
59. Kelly, M., Burke, J., Smith, M., Klar, A. and Beach, D. (1988) *EMBO J.* **7**, 1537–1547.
60. Herr, W., Sturm, R. A., Clerc, R. G., Corcoran, L. M., Baltimore, D. M., Sharp, P. A., Ingraham, H. A., Rosenfeld, M. G., Finney, M., Ruvkun, G. and Horvitz, H. R. (1988) *Genes Dev.* **2**, 1513–1516.
61. Atlung, T., Nielsen, A. and Hansen, F. G. (1989) *J. Bacteriol.* **171**, 1683–1691.
62. Zakut-Houri, R., Oren, M., Bienz, B., Lavie, V., Hazum, S. and Givel, D. (1983) *Nature (London)*, **306**, 594–597.
63. Reidl, J., Römisch, K., Ehrmann, M. and Boos, W. (1989) *J. Bacteriol.* 171, 4888–4899
64. Halling, C., Calendar, R., Christie, G. E., Dale, E. C., Dehò, G., Finkel, S., Flensburg, J., Ghisotti, D., Kahn, M. L., Lane, K. B., Lin, C. -s., Lindqvist, B. H., Pierson, L. H. (III), Six, E. W., Sunshine, M. G and Ziermann, R. (1990) *Nucleic Acids Res.* 18, 1649.