2488–2497 *Nucleic Acids Research, 1996, Vol. 24, No. 13*

The helix–hairpin–helix DNA-binding motif: a structural basis for non-sequence-specific recognition of DNA

Aidan J. Doherty, Louise C. Serpell and Christopher P. Ponting[1,*]

Laboratory of Molecular Biophysics, Rex Richards Building, University of Oxford, South Parks Road, Oxford OX1 3QU, UK and [1]Fibrinolysis Research Unit, University of Oxford, The Old Observatory, South Parks Road, Oxford OX1 3RH, UK

Received April 11, 1996; Revised and Accepted May 23, 1996

ABSTRACT

One, two or four copies of the 'helix–hairpin–helix' (HhH) DNA-binding motif are predicted to occur in 14 homologous families of proteins. The predicted DNA-binding function of this motif is shown to be consistent with the crystallographic structure of rat polymerase β, complexed with DNA template–primer [Pelletier, H., Sawaya, M.R., Kumar, A., Wilson, S.H. and Kraut, J. (1994) *Science* 264, 1891–1903] and with biochemical data. Five crystal structures of predicted HhH motifs are currently known: two from rat pol β and one each in endonuclease III, AlkA and the 5′ nuclease domain of *Taq* pol I. These motifs are more structurally similar to each other than to any other structure in current databases, including helix–turn–helix motifs. The clustering of the five HhH structures separately from other bi-helical structures in searches indicates that all members of the 14 families of proteins described herein possess similar HhH structures. By analogy with the rat pol β structure, it is suggested that each of these HhH motifs bind DNA in a non-sequence-specific manner, via the formation of hydrogen bonds between protein backbone nitrogens and DNA phosphate groups. This type of interaction contrasts with the sequence-specific interactions of other motifs, including helix–turn–helix structures. Additional evidence is provided that alphaherpesvirus virion host shutoff proteins are members of the polymerase I 5′-nuclease and FEN1-like endonuclease gene family, and that a novel HhH-containing DNA-binding domain occurs in the kinesin-like molecule nod, and in other proteins such as cnjB, emb-5 and SPT6.

INTRODUCTION

The interaction of proteins with DNA, in a sequence-dependent manner, is fundamental to DNA synthesis, repair and degradation, and to the regulation of gene transcription. Many of these proteins contain small, discrete structural motifs that utilize either α-helices or β-strands to bind the phosphate backbone or the grooves of DNA. Among these are the helix–turn–helix, zinc finger, leucine zipper and helix–loop–helix motifs (1). These may arise in different molecular contexts which has been interpreted to be due either to divergent evolution via gene duplication and insertion or to structural convergence via the effects of selective pressures on protein function. Thus the helix–turn–helix (HtH) motif is found both in molecules with similar folds (e.g. homeodomain proteins) and in others with different folds (e.g. lambda repressor and cro proteins). Unlike gene regulatory proteins, other molecules that bind DNA do so in a manner that is non-sequence-specific. These proteins include nucleases, *N*-glycosylases, ligases, helicases, topoisomerases and polymerases that are essential for the protein-mediated synthesis and repair of DNA structure. Much less is known about how such proteins bind DNA or how this sequence independent binding is coupled to their function (2). These proteins, unlike sequence-dependent DNA binding proteins, have not previously been shown to possess a common structural motif.

Recently, the N-terminal region of an open reading frame (ORF) of the cyanobacterium *Synechocystis* sp. has been shown to be a member of a family of phosphodiesterases that includes phospholipases D and endonucleases (3). Further investigation of this ORF indicated that the sequence of its C-terminal region is similar to a previously proposed family of DNA-binding proteins (4) that includes *Bacillus subtilis* comE ORF A and human OriP-binding protein (OriP-BP). This suggests that the *Synechocystis* sp. ORF encodes a nuclease with a C-terminal DNA-binding domain. The family of DNA-binding domains was suggested (4) to include regions of *Escherichia coli* uvrC (a subunit of the uvrABC DNA repair enzyme) and *Haloarcula marisortui* ribosomal protein HL5.

Further studies indicated that this family of DNA-binding domains also was significantly similar in sequence to a variety of other molecules, each of which possessed a DNA-binding function, over regions of ~20 amino acids. It was considered that these sequence similarities either arose as results of evolutionary divergence from a common ancestor (i.e. homology) or by localised convergence of sequence due to adaptive replacements that were positively selected (cf. 5). Here evidence from similarity searches of sequence and structural databases is presented that suggests that a 'helix–hairpin–helix' motif occurs in 14 families of proteins and that this mediates a non-sequence-dependent interaction with DNA.

* To whom correspondence should be addressed

METHODS

Searches of sequence databases

Sequence searches were undertaken using a local similarity method of Barton (6) and Searchwise (7); a generalised profile method. Additional searches for homologues used BLAST (8), as implemented at the National Center for Biotechnology Information (NCBI) USA. Estimation of *p*-values for ungapped blocks within multiple alignments was provided by the program MACAW (9). Calculations of *p*-values were overestimations due to the use of a maximal search space (9) equal to the product of proteins' sequence lengths. The BLOSUM62 amino acid substitution matrix (10) was used for each of these computational methods. Secondary structure predictions from multiple alignments were provided by the neural network method (PHD) of Rost and Sander (11).

Searches of structure databases

Comparison of individual HhH structures with current structure databases was provided by the dynamic programming algorithm encoded within STAMP (12). The algorithm was used to calculate length-independent degrees of similarity (S_c-values) between a query structure and all other known structures where maximum values of S_c = 9.8 represent a comparison of any structure with itself. The putative HhH motifs in *E.coli* endonuclease III (residues 111–126), rat polymerase β (residues 59–74 and 100–115) and *E.coli* AlkA (residues 209–224) were subjected to database searches each as the query structure.

Comparison of HhH structures

The four bi-helical structures were superimposed using matrices generated by STAMP and examined using RASMOL (13). Examination of the hierarchical lists of scores from searches and visualisation of superpositions using RASMOL demonstrated that superpositions with S_c scores >8.00 represented highly similar structures. Multiple alignment of the four similar structures was performed using STAMP and values of P_{ij}, were calculated, these represent the confidence in the alignment at each C-α position. Crossing angles between the two helices of the HhH structures were determined using the structural analysis program ACTIVE (Hyeon Son, in preparation); helical regions were defined using Quanta (Molecular Simulations).

RESULTS AND DISCUSSION

Sequence-based searches

Regions of *comE* ORF A, OriP-BP, HL5 and uvrC have been suggested elsewhere to form a family of DNA-binding domain homologues (4). Results of preliminary database searches indicated that the sequences of several other DNA-binding proteins were apparently significantly similar to a 16 residue motif of *Synechocystis* sp. ORF, *comE* ORF A, OriP-BP, HL5 and uvrC: for example, regions similar to the motif in *Synechocystis* sp. ORF, rat polymerase β, *E.coli* ruvA and *E.coli* recR, when aligned in pairs, yielded probabilities of being aligned by chance (*p*-values) of $<5 \times 10^{-2}$, calculated using MACAW (9). In the absence of known structures for each of these proteins it was considered prudent to search, using both a local similarity algorithm (6) and a generalised profile method (7), for sequences that possess statistically significant similarities to a 16 residue profile calculated from regions of *Synechocystis* sp. ORF, *comE* ORF A, OriP-BP, HL5 and uvrC sequences.

The initial search (6) provided evidence that such sequences exist predominantly in DNA-interacting proteins. The 11 sequences, that were most similar to the profile (i.e. the top scoring sequences), were three NAD^+-dependent DNA ligases, *E.coli* recR and ruvA (both involved in DNA repair processes), rat DNA polymerase β, *Pseudomonas fluorescens* uvrC, *Drosophila melanogaster* nod (a DNA-binding kinesin homologue), human flap endonuclease-1 (FEN1), *Thermus aquaticus* DNA-binding protein 1a (DNAB1a), and an endonuclease from *Saccharomyces cerevisiae* (ORF YKL113c). Using MACAW (9), comparisons of these sequences in pairs and in groups demonstrated that the probabilities that these similarities in sequence arose by chance were small. For example, aligning 16 residue regions of *Rhodothermus marinus* DNA ligase, nod, FEN1 and DNAB1a yielded a *p*-value of 1×10^{-7}.

A new profile, containing both the original five sequences and 11 that were the next most similar, was calculated and compared (6) once more with databases. The top-scoring 53 sequences in this iteration contained the 16 residue motif; significantly the motif was conserved also for their close homologues, as defined by conservation at five or more of eight positions of hxxhxGhGxxxAxxhh, where h is a hydrophobic residue (VILMWFYA). Four exceptions to this, where homologues did not conserve the motif, were discarded; these four were: *Streptococcus pneumoniae* penicillin-binding protein, trout cellular tumour antigen p53, varicella-zoster virus gene 53 protein and *S.pombe* sexual differentiation process protein ISP7. A slight reduction of the acceptance threshold suggested the presence of multiple copies of the motif in NAD^+-dependent ligases (four copies), ruvA, polymerase β and HL5 (two copies each).

Comparison of a newly-generated profile with databases using Searchwise, a profile method (7), corroborated and extended these findings. In this iteration, the top scoring 68 sequences were all represented among previously-identified sequences, together with radC homologues, that also scored highly. Collating these top-scoring sequences with their homologues yielded a total of 107 sequences. In order to decrease redundancy, one of each pair of close homologues whose motifs showed >60% sequence identity was removed from the list. Unexpectedly, each of these 107 was found to represent a protein known to interact with DNA, except for four ORFs whose functions are unknown; no protein whose known function is unrelated to DNA was present among the 107. This procedure was unlikely to have identified by chance such a set of proteins with functions so disproportionately related to DNA (as an illustration of this it is noted that only 2% of SwissProt database entries have the string 'DNA' in their title); therefore, this motif is likely to have arisen in each of these sequences as a result of a common DNA-related function. From literature data (discussed below), this function is likely to be a DNA-binding function. Given the stringent selection criteria for this motif it is likely that many more examples, with slightly differing sequences, remain to be identified in current sequence databases. Seven of the 107 sequences could not be identified as close homologues of any other proteins (Table 1).

The 107 sequences could be clustered into 14 homologous families (Table 1). Some care was taken to ensure that sequences were demonstrably homologous by ensuring that significant sequence similarity existed outside of the motifs. This precaution

Table 1. HnH sequence alignment

CLASS	Residue	1 10 20	Copy	Accession Code	Species	NAME/FUNCTION
1	56	AEAKKLPGVGTKIAEKIDEF	1	Dpob_Rat	*Rno*	DNA Polymerase β
	97	NFLTRVTGIGPSAARKLVDE	2	Dpob_Rat	*Rno*	DNA Polymerase β
	223	EQLEKMPGCGPKIVHLWKEF	1	Ya26_Schpo	*Spo*	DNA Polymerase β homologue
	264	LLFYNIFGVGASHAAEWYQK	2	Ya26_Schpo	*Spo*	DNA Polymerase β homologue
	57	QDLKAFSGFGAKLLKKAEEI	1	A55874	*Cfa*	DNA Polymerase β homologue
	97	QELTQVHGFGPPRRAALFDR	2	A55874	*Cfa*	DNA Polymerase β homologue
	232	EAHTKLRNIGPSIAKKIQVI	1	Dpo4_Yeast	*Sce*	DNA Polymerase IV
	271	KYFKNCYGIGSEIAKRWNLL	2	Dpo4_Yeast	*Sce*	DNA Polymerase IV
	209	KDTEGIPCLGSKVKGIIEEI	1	Tdt_Human	*Hsa*	DNA nucleotidylexotransferase
	250	KLFTSVPGVGLKTSEKWFRM	2	Tdt_Human	*Hsa*	DNA nucleotidylexotransferase
2	191	DNLPGVKGIGEKTARKLLEE		Dpol_Theaq	*Taq*	DNA polymerase I (5′ nuclease domain)
	188	DNIPGVPGVGEKTAQALLQG		Dpol_Ecoli	*Eco*	DNA polymerase I (5′ nuclease domain)
	202	DNLPGIPGVGEKTAAKWIAE		Dpol_Myctu	*Mtu*	DNA polymerase I (5′ nuclease domain)
	192	DNLAGIKGIGPIKGIELLQQ		Mgu39705_2	*Mge*	DNA polymerase I (5′ nuclease domain)
	177	SKVPGVAGIGPKSATQLLVE		Exo_Ecoli	*Eco*	5′-3′ exonuclease
	204	DNIRGVEGIGAKRGYNIIRE		Exo5_Bpt5	*BT5*	5′ exonuclease
	233	DYCESIRGIGPKRAVDLIQK		Fen1_Human	*Hsa*	5′ FLAP endonuclease
	850	DYTMGLSRVGPVLALEILHE		Ral3_Schpo	*Spo*	RAD13, DNA repair protein
	861	DYTEGIPTVGCVTAMEILNE		Xpg_Human	*Hsa*	ERCC-5, DNA excision repair protein
	864	DYTNGLKGMGPVSSIEVIAE		Rad2_Yeast	*Sce*	rad2, DNA repair protein
	225	DYTDGVAGMGLKTALRYLQK		Yspexo1a_1	*Spo*	Exonuclease I
	375	DFTSRIPKIGPVRALKLIRY		Ya31_Schpo	*Spo*	ORF (SPAC18B11.01C)
	288	DYNRGVKGLGKNKSLQLAQC		Yen1_Yeast	*Sce*	ORF (YER041W)
	213	YFQRGVQNIGIVSVFDILGE		Celt12a2_6	*Cel*	ORF (T12A2.8)
3	109	AALEALPGVGRKTANVVLNT		End3_Ecoli	*Eco*	Endonuclease and DNA *N*-glycosylase (Endo III)
	108	DGLCALPGVGPKMANLVMQI		Cer10e4_6	*Cel*	Endo III homologue
	232	NELLGLPGVGPKMAYLTLQK		Yab5_Yeast	*Sce*	Endo III homologue
	109	EEVAALPGVGRSTAGAILSL		Muty_Ecoli	*Eco*	MutY, A/G specific adenine glycosylase
	143	EWAKGIPGVGPYTAGAVLSI		Spac26a3_2	*Spo*	MutY homologue
	115	KAILDLPGVGKYTCAAVMCL		Gtmr_Mettf	*Mth*	MutY homologue
4	178	SLVTNVKGIGPWSAKMFLIS		Mag_Yeast	*Sce*	MAG, DNA-3-methyladenine glycosidase
	68	EQLTAIKGIGPWTAQLFLLF		Sycslrb_123	*Ssp*	MAG homologue
	207	KTLQTFPGIGRWTANYFALR		3mg2_Ecoli	*Eco*	AlkA, DNA-3-methyladenine glycosidase II
	227	KNLIKIRGIGPWTANYVLMR		3mga_Bacsu	*Bsu*	DNA-3-methyladenine glycosidase
5	502	EELESLPGIGPSTAEKIIST		Sycslrd_69	*Ssp*	Phospholipase D homologue, nuclease?
	152	EELQGISGVGPSKAEAIIAY		Cmel_Bacsu	*Bsu*	comE ORF A, DNA-binding protein
	30	EELVLLKGIGTVKAQAIVDY		Dnu26684_1	*Dno*	Cmel orthologue
	107	RDLRSLQRIGPKKAQLIVGW		Humobp2a_1	*Hsa*	oriP-binding protein
	612	KQLQEIPGIGPKSAFSLALH		Nod_Droma	*Dme*	DNA-binding kinesin
	?	NLLQRINGLNEKTAKEIVQY		Celdyneinh	*Cel*	ORF near dynein locus
	816	DCLQFIPGFGPRKSHPFISQ		Tetcnjba_1	*Tth*	cnjB, gene active during meiosis
	353	PLLTRVAGLTRMMAQNIVAW		Ecouw67_332	*Eco*	ORF (o622)
6	1372	PPFNVIDGLGETLAQKIVDS		Mpu06833_1	*Mpu*	DNA polymerase III (polC)
	1374	PPFNSIPGLGTNAALNIVKA		Dp3a_Bacsu	*Bsu*	DNA polymerase III, α-chain
	833	YGIGAIKGVGEGPIEAIIEA		Dp3a_Ecoli	*Eco*	DNA polymerase III, α-chain
7	446	RRAMDVDGMGDKIIDQLVEK	1	Dnlj_Ecoli	*Eco*	NAD$^+$-dependent DNA ligase
	480	GKLTGLERMGPKSAQNVVNA	2	Dnlj_Ecoli	*Eco*	NAD$^+$-dependent DNA ligase
	512	LYALGIREVGEATAAGLAAY	3	Dnlj_Ecoli	*Eco*	NAD$^+$-dependent DNA ligase
	544	EELQKVPDVGIVVASHVHNF	4	Dnlj_Ecoli	*Eco*	NAD$^+$-dependent DNA ligase
	447	RKAMDIQGLGEKLIERLLEK	1	Dnlj_Theth	*Taq*	NAD$^+$-dependent DNA ligase
	481	EDLVGLERMGEKSAQNLLRQ	2	Dnlj_Theth	*Taq*	NAD$^+$-dependent DNA ligase
	513	LYALGLPGVGEVLARNLAAR	3	Dnlj_Theth	*Taq*	NAD$^+$-dependent DNA ligase
	545	EELLEVEEVGELTARAILET	4	Dnlj_Theth	*Taq*	NAD$^+$-dependent DNA ligase
	487	RDAMDIEGMGSQVARQLAES	1	Rmu10483_1	*Rhm*	NAD$^+$-dependent DNA ligase
	521	EDLLKLEGFAETRARNLLRA	2	Rmu10483_1	*Rhm*	NAD$^+$-dependent DNA ligase

Table 1. (*cont.***)**

	553	LFGLGIRHVGKTTAELLVQR	3	Rmu10483_1	*Rhm*	NAD⁺-dependent DNA ligase
	585	DELAALEGVGPITAESIANW	4	Rmu10483_1	*Rhm*	NAD⁺-dependent DNA ligase
	472	RTAFEIDGLGKSHIESFFAD	1	Dnlj_Zymmo	*Zym*	NAD⁺-dependent DNA ligase
	508	QLLIERERGWGELSVDNLISA	2	Dnlj_Zymmo	*Zym*	NAD⁺-dependent DNA ligase
	540	LFALGIRHVGAVTARDLAKS	3	Dnlj_Zymmo	*Zym*	NAD⁺-dependent DNA ligase
	600	ISFFHIPNMGGKIIRSLLDF	4	Dnlj_Zymmo	*Zym*	NAD⁺-dependent DNA ligase
	435	KTAMDINGLNINTITKLYEH	1	Mgu39703_12	*Mge*	NAD⁺-dependent DNA ligase
	472	QVLKLDLKIGDKLFNKLVDN	2	Mgu39703_12	*Mge*	NAD⁺-dependent DNA ligase
	504	LTGLGIKHVGNVLAKNLANH	3	Mgu39703_12	*Mge*	NAD⁺-dependent DNA ligase
	536	ENLISLNDVGITVAESLYNW	4	Mgu39703_12	*Mge*	NAD⁺-dependent DNA ligase
	327	EELDEVEGIGEVRAQKIKKG		Yack_Bacsu	*Bsu*	ORF (yack)
8	556	SSLETIEGVGPKRRQMLLKY		Uvrc_Ecoli	*Eco*	uvrC, exonuclease ABC subunit C
	537	SVLDDIPGIGEKRKKMLLKH		Uvrc_Bacsu	*Bsu*	uvrC, exonuclease ABC subunit C
	73	QELIAFPGIGPAKATTILAA		Sycslrb_14	*Ssp*	uvrC, exonuclease ABC subunit C
	73	TPLDEIPGVGAARKRALLAH		Rsu29587_1	*Rsp*	uvrC, exonuclease ABC subunit C
	377	NPLLQIPGVGKITAQILFDN		Mgu39698_1	*Mge*	uvrC, exonuclease ABC subunit C
9	73	KELIKTNGVGPKLALAILSG	1	Ruva_Ecoli	*Eco*	ruvA, Holliday junction DNA helicase subunit
	108	GALVKLPGIGKKTAERLIVE	2	Ruva_Ecoli	*Eco*	ruvA, Holliday junction DNA helicase subunit
	72	LALLSVSGVGPRLAMATLAV	1	Ruva_Mycle	*Mle*	ruvA, Holliday junction DNA helicase subunit
	107	ASLTRVPGIGKRGAERIVLE	2	Ruva_Mycle	*Mle*	ruvA, Holliday junction DNA helicase subunit
10	12	EALRCLPGVGPKSAQRMAFT		Recr_Ecoli	*Eco*	recR, DNA repair protein
	12	DSFMKLPGIGPKTAVRLAFF		Recr_Bacsu	*Bsu*	recR, DNA repair protein
	5	QPLHVLPGVGPKSAEKYAKL		Spnmsagen_2	*Spn*	recR, DNA repair protein
11	73	EELSSIPGIGMVKAIQILAA		Radc_Bacsu	*Bsu*	radC, DNA repair protein
	66	EQFSGVHGIGVAKFAQLKGI		Radc_Ecoli	*Eco*	radC, DNA repair protein
	76	KAFCSVKGLGITQFIQLQAI		Radc_Haein	*Hin*	radC, DNA repair protein
12	269	EDLALCPGLGPQKARRLFDV		Ercl_Human	*Hsa*	ERCC-1, DNA excision repair protein
	211	DELEQLEGWGPTKVNRFLEA		Swi0_Schpo	*Spo*	swi10, DNA excision repair protein
13	16	IALTSIYGVGKTRSKAILAA		Rs13_Ecoli	*Eco*	S13, ribosomal protein
	10	IALTYIFGIGLSSAKTILKK		U01733_1	*Mge*	S13, ribosomal protein
	16	ISLTYIFGIGRTTAQQVLKE		Rs13_Bacsu	*Bsu*	S13, ribosomal protein
	63	YSLQYIHGIGRTRARQILVD		Rs13_Arath	*Ath*	S13, ribosomal protein
	17	ISLTYIYGIGPALSKEIIAR		Chtrpsl_2	*Ctr*	S13, ribosomal protein
	25	YALAMIKGIGYNTAMIIIRK		Rs13_Sulac	*Sac*	S13, ribosomal protein
	39	RALTELNGIGHRAARIIAQK		Rs13_Halma	*Hma*	S13, ribosomal protein
	17	IASTKIDGIGPKKAIQVRYR		Rt13_Tobac	*Nta*	S13, ribosomal protein
	17	QACAQIYGLGHHHCLQICDV		Pwu02970_16	*Pwi*	S13, ribosomal protein
	28	FAITAIKGVGRRYAHVVLRK		Rs18_Human	*Hsa*	S18, ribosomal protein
14	195	SLLATIPGIGKKTLPHLLVV		Piv_Morla	*Mla*	Pilin gene inverting protein
	281	EILLSFPGLGPLLGARVLAE		Ym3_Strco	*Sco*	Mini-circle hypothetical protein
	270	EIIESMPGMGPVLGAEFVAI		Yil1_Strcl	*Scl*	Insertion element (IS116)
	192	EVLMALPGVGPQVAAAVLAL		TTHISOR_5	*Taq*	DNA-binding protein 1A
	274	PVLTSMPGVGVRTAAVLLVT		Yis1_Strco	*Sco*	Insertion element (IS110)
	210	QRLLTIPGIGTITASLLATK		Yenis1328_1	*Yen*	Insertion element (IS1328)
	213	QRVQSIPGVGYLTALSVYAS		Coxtranspo_1	*Cbu*	Transposase (IS1111a)
	194	CILQSMKGIGKIASASIISN		A32816	*Pat*	Insertion element (IS492)
misc.	14	TELTDISGVGPSKAESLREA	1	Rl32_Halma	*Hma*	HL5, ribosomal protein
	47	SALADVSGIGNALAARIKAD	2	Rl32_Halma	*Hma*	HL5, ribosomal protein
	272	GITLVNIGVGPSNAKTICDH		Amn_Ecoli	*Eco*	AMP nucleosidase
	242	KLLSGVPNIGKKLAAEILKD		Asu18466_54	*Asv*	ORF (EP364R)
	135	DDLKRIKGIGPKISDWLNAQ		Ynq2_Parde	*Pde*	ORF
	145	IPLQPVPGVGPKTLDKLKKA		Yqxk_Bacsu	*Bsu*	ORF (yqxk)
	230	EHLMSYNGVGPKVADCVCLM		S49801	*Sce*	ORF (YM9958.02)

Fourteen homologous families of HhH sequences. Every pair of motifs shares <60% sequence identity. The HhH motif, from the 5' nuclease domain of *Taq* pol I, has been added to this set as a consequence of its known tertiary structure (15). Locations of the initial residues of the HhH sequences and (Swissprot or GenBank) accession codes are shown preceding and following the alignment respectively. Abbreviations: Asv, African swine fever virus; Ath, *Arabidopsis thaliana*; Bsu, *B.subtilis*; BT5, bacteriophage T5; Cbu, *Coxiella burnetii*; Cel, *C.elegans*; Cfa, *Crithidia fasciculata*; Ctr, *Chlamydia trachomatis*; Dme, *D.melanogaster*; Dno, *Dichelobacter nodosus*; Eco, *E.coli*; Hin, *Haemophilus influenzae*; Hma, *H.morismortui*; Hsa, *Homo sapiens*; Mge, *Mycoplasma genitalium*; Mla, *Moraxella lacunata*; Mle, *Mycobacterium leprae*; Mth, *Methanobacterium thermoformicicum*; Mpu, *Mycoplasma pulmonis*; Mtu, *Mycobacterium tuberculosis*; Nta, *Nicotiana tabacum*; Pat, *Pseudomonas atlantica*; Pde, *P.denitrificans*; Pwi, *Prototheca wickerhamii*; Rhm, *Rhodothermus marinus*; Rno, *Rattus norvegicus*; Rsp, *Rhodobacter sphaeroides*; Sac, *Sulfolobus acidocaldarius*; Sce, *S.cerevisiae*; Scl, *Streptomyces clavuligerus*; Spn, *Streptococcus pneumoniae*; Spo, *S.pombe*; Ssp, *Synechocystis* sp.; Taq; *Thermus aquaticus*; Tth, *Tetrahymena thermophila*; Yen, *Yersinia enterocolitica*; Zym, *Zymomonas mobilis*; other abbreviations in text.

was warranted by the observation that the five known structures that contain the motif (pol β, AlkA, endonuclease III and *Taq* polymerase I; see below) do not adopt a common fold and may be results of localised sequence and structural convergence.

Secondary structure predictions were provided by the PHD server (11) using multiple alignments of all homologues in Table 1 whose tertiary structures remain unknown, as query information. At an expected accuracy of prediction of 72% (11), eight out of 13 alignments produced a prediction of two α-helices, 4 (radC, uvrC, NAD^+-dependent ligase motif 1 and ruvA motif 2) yielded an uncertain prediction for the first half of the motif followed by prediction of an α-helix, and 1 (tranposase homologues) yielded a β-strand–β-strand prediction. These predictions are consistent with the proposal that the majority of these sequences represents a bi-helical structure; the possibility that a minority of these do not contain a N-terminal α-helix can not be discounted.

Structure-based searches

These observations of sequence similarities were able to be correlated with structural similarities given that the crystal structures of five of these motifs have been determined. Rat polymerase β (pol β; containing two motifs) (14), *T.aquaticus* polymerase I (*Taq* pol I) (15), *E.coli* AlkA (T. Ellenberger, personal communication) and *E.coli* endonuclease III (endo III) (16) do not all adopt a common fold yet each contains a bi-helical structure with a short inter-helical loop that coincides with their sequence-similar motifs. Since tandem helices are a common occurrence in protein structures (17) it was important to assess the significance of the perceived similarities between the bi-helical motifs of pol β, pol I, endo III and AlkA. Henceforth these motifs shall be termed 'helix–hairpin–helix' (HhH) motifs in accordance with Thayer *et al.* (18) (see below).

The STAMP algorithm (12) was used to compare, in a pair-wise manner, each of four HhH motifs of known structure (AlkA, endo III and two from pol β) with all structures contained in the Brookhaven database, and with other structures obtained locally; the resolution of the pol I HhH structure was inappropriate to allow its comparison with databases. STAMP-derived values of scores (S_c) quantified the pairwise similarities between the query structure and all other structures in a length-independent manner (Table 2). The most striking finding of the STAMP search was that each of the HhH probe structures was identified as being similar to each of the other predicted HhH motifs with significantly high S_c values (Table 2). S_c values ≥ 8.33 demonstrated the signficance of their structural similarities; these values are similar to the top scores obtained when searching using other query structures, for example the helix–turn–helix motif (results not shown). The reliability of these scores was supported by $P_{ij'}$ values that exceeded 10 throughout the alignment. Russell and Barton (12) have previously shown that $P_{ij'}$ values >6.0 represent regions where the structural alignment are of high reliability; this was confirmed by direct visualisation of superpositions. Superposition of Pol β 1 and 2, Endo III and AlkA HhH structures demonstrated the remarkable similarity between the topology of these structures (Fig. 1), highlighting that the HhH motif is composed of two helices, of conserved length, linked by a Type II β-bend held by a single hydrogen bond, and crossing at similar angles (Table 3). The *Taq* pol I HhH, although poorly ordered in the crystal structure, also demonstrates many of these characteristics. It is evident that these bi-helical structures are significantly alike, both in sequence and in structure. Furthermore, they are significantly dissimilar to other structural motifs, including other bi-helical motifs such as helix–turn–helix motifs, in the arrangement and crossing angle of the helices and also in the type of turn that bridges the helices.

Figure 1. Structural comparison of four HhH motifs. Superposition of the α-carbon atoms of HhH motifs (positions 1–20) of the pol β HhH 1 (yellow) and HhH 2 (blue), endonuclease III (green), AlkA (red) using STAMP (12). The root-mean-square deviations of backbone atoms between HhH positions 3 and 18 were calculated to be 0.85 Å (pol β HhH 1 and 2), 0.88 Å (pol β HhH 1 and endo III HhH), 0.65 Å (pol β HhH 2 and endo III HhH), 0.98 Å (pol β HhH 1 and AlkA HhH), 0.65 Å (pol β HhH 2 and AlkA HhH) and 0.75 Å (endo III HhH and AlkA HhH). The figure was prepared using RASMOL (13).

The 'helix–hairpin–helix' motif

Comparison was made at this stage with literature sources. The motif derived here corresponded exactly with the 'helix–hairpin–helix' (HhH) DNA-binding motif suggested by Thayer *et al.* (18) to occur in 13 sequences; the number of putative HhH-containing sequences was extended by five by Seeberg *et al.*(19). However, three homologous sequences, human, mouse and yeast replication factors C, suggested by Thayer *et al.* (18) and Seeberg *et al.* (19), to contain HhH motifs were not contained among the 107 sequences identified in this study. On closer inspection it was apparent that these three sequences (consensus sequence GPPG[V/I]GKT) conformed better to the 'P-loop' consensus sequence ([A/G]xxxxGK[S/T]) (20) than to the HhH motif, described here. Although similar in sequence, P-loops contain β-strand–loop–α-helix secondary structures (20,21) in contrast with the α-helix–loop–α-helix structures of HhH motifs.

The five HhH-containing proteins, whose structures have been determined, do not share a common fold, indicating that the HhH motifs, in these examples, are likely to have arisen either by local sequence convergence or by gene duplication and insertion. The crystal structures of the five HhH motifs, and the DNA-bound form of one of these (the second HhH of rat pol β), have allowed

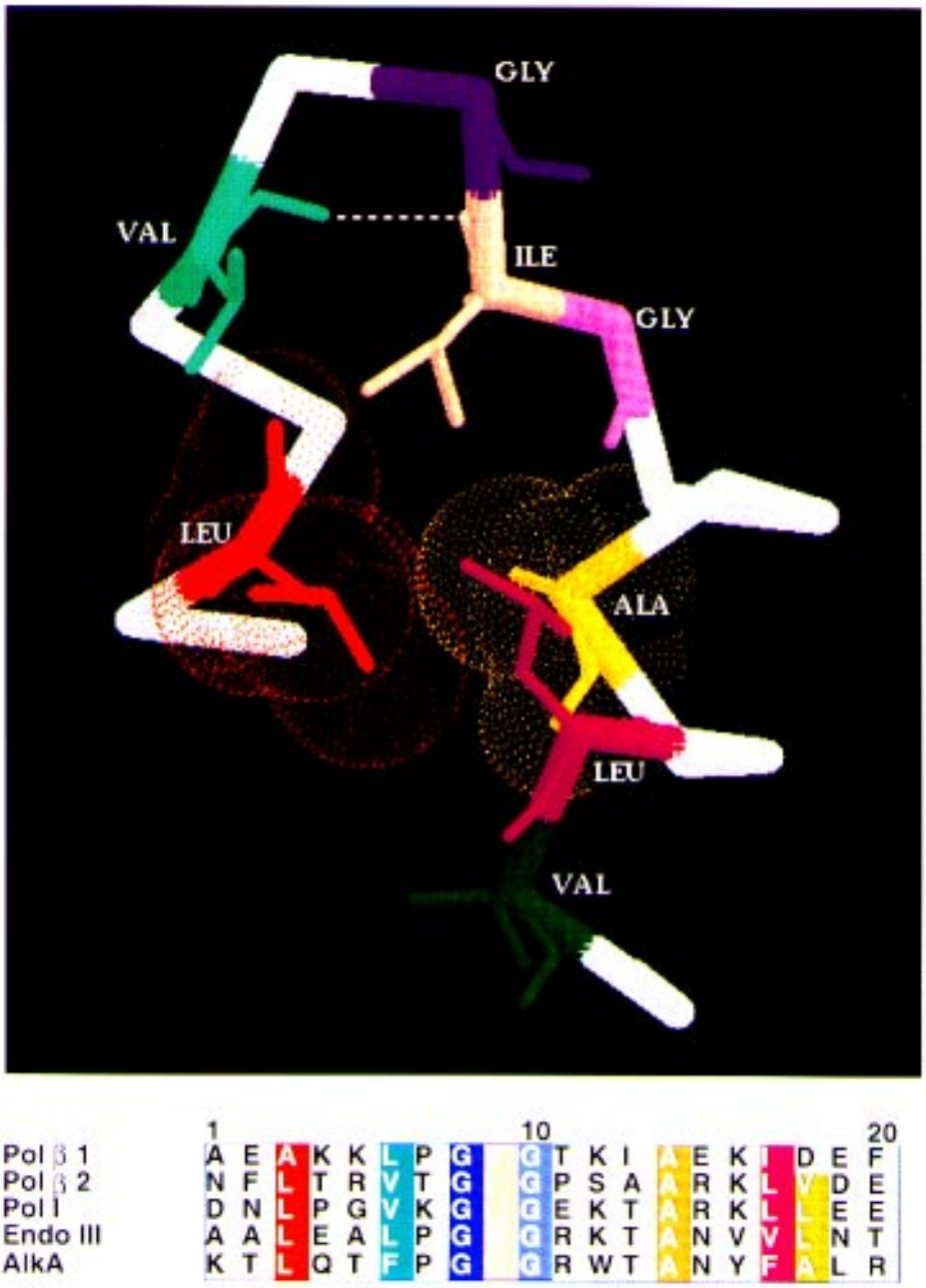

Figure 2. HhH motif. The α-carbon backbone of the second HhH motif from rat pol β. Positions with high propensities for particular types of amino acids have been coloured consistent with the colour scheme of the sequence alignment (below). The side chains of these residues have also been included to demonstrate their contribution to the HhH motif structure. High propensity for hydrophobic residues at HhH positions 3 and 14 (van der Waals surfaces shown in red and orange respectively) suggest their importance in the maintenance of the relative orientations of the two helices. An alanine or small hydrophobic residue is most commonly found at HhH 14 preventing steric clashes between the helices at their crossing point. The multiple sequence alignment of HhH motifs of known structure, as displayed using Alscript (45), shows hydrophobic residues (VILMFYWA) at positions 3, 6, 9, 17 and 18, glycines at positions 8 and 10, and small residues (AGCS) at position 14, with coloured backgrounds.

the first detailed examination of the structural basis for the significant sequence similarities among HhH motifs (Fig. 2), and the amino acids that are important for DNA recognition. It is particularly striking that several positions with high propensities for hydrophobic residues (positions 3, 6, 9, 14, 17 and 18; numbering as Table 1) are all buried at the interface between the two helices. It would appear that residues at positions 3, 17 and 18 stabilise helical packing whereas residues at positions 6 and 9 stabilise the β-turn. An anti-parallel type hydrogen bond between these amino acids also stabilises the hairpin. Glycines at positions 8 and 10 form important elements of the hairpin loop: glycine HhH8 appears to be important for the formation of the type II β-turn, whereas glycine HhH10 contributes to a pronounced extended surface that mediates DNA-binding (discussed below). An alanine or small hydrophobic residue is most commonly found at HhH 14. The absence of a bulky side chain at this

position on the second helix in the majority of putative HhH sequences appears to prevent steric clashes with helix 1 at the helices' crossing point (Fig. 2).

Table 2. The top scoring hits and 'false positives' in a search of structural databases

Search query PDB code (residues)	Name/PDP (residues)	Score (Sc)
Pol β 1	Pol β 1 (59–74)	9.17
2BPF 1 (59–74)	Pol β 2 (100–115)	8.67
	Endo III (111–126)	8.59
	AlkA (209–224)	8.33
	1pdn (C29–C42)	7.42
	3wrp (60–74)	7.32
	1trh (362–377)	7.30
	1edd (191–206)	6.97
Pol β 2	Pol β 2 (100–115)	9.17
2BPF 2 (100–115)	Endo III (111–126)	8.91
	AlkA (209–224)	8.75
	Pol β 1 (59–74)	8.67
	1ecl (377–391)	7.95
	3wrp (60–74)	7.79
	1trh (362–377)	7.74
	1mys (539–553)	7.68
Endo III	Endo III (111–126)	9.17
2ABK (111–126)	Pol β 2 (100–115)	8.88
	AlkA (209–224)	8.75
	Pol β 1 (59–74)	8.59
	1gky (132–146)	7.65
	2dnj (235–249)	7.54
	1ecl (377–391)	7.42
	1bip (79–93)	7.19
AlkA (209–224)	AlkA (209–224)	9.33
	Pol β 2 (100–115)	8.91
	Endo III (111–126)	8.75
	pol β 1 (59–74)	8.33
	1ctf (72–86)	8.03
	1apl (165–179)	7.84
	1mss (98–112)	7.68
	1lib (20–33)	7.53

HhH motifs in Endo III, AlkA and pol I

A DNA-binding function has been proposed (18) for the putative HhH motif in endonuclease III, based on its identification as the binding site of thymine glycol (15), a known inhibitor of the *N*-glycosylase activity of the enzyme. In this structure the electron density for the inhibitor was weak and the authors could not identify unambiguously the nature of its interactions with the HhH motif (16). Resolution of this question awaits the determination of the tertiary structure of the DNA-bound form of endonuclease III. The crystal structure of another DNA glycosylase, AlkA, has recently been determined (T. Ellenberger, personal communication) and, as predicted, it also contains a HhH motif. *Escherichia coli* AlkA is involved in base excision repair; this work indicates that the AlkA HhH motif is likely to mediate its affinity for DNA during repair processes.

The crystal structure of DNA polymerase I from *T.aquaticus* (*Taq* pol I), containing the first description of the structure of a 5′-nuclease domain, has been reported recently (15). A HhH motif, predicted to reside in the 5′ nuclease domain (residues 191–211), does indeed adopt a helix–hairpin–helix-like structure although the structure of this region could not be determined unambiguously due to high crystallographic B-factors and poor electron density, with no density present for residues 200 and 201 (15). However it was possible to superimpose accurately the helices of the pol I HhH with other HhH structures; in addition it has a crossing angle similar to other HhH structures (Table 3). The *Taq* pol I structure does not contain bound DNA, but Kim *et al.* (15) have proposed three metal ion binding sites formed by conserved carboxylates situated at the base of the major cleft in the 5′ nuclease domain as constituting an active site. Interestingly, the HhH protrudes into this cleft adjacent to the metal binding sites and it seems plausible that the motif presents DNA to the nuclease active site.

The 5′-nuclease domain of polymerases I from a diverse range of organisms is highly conserved (22). Analysis of mutations within the *E.coli* pol I 5′ nuclease domain (23) reveals that two mutations (Gly_{184}→Asp and Gly_{192}→Asp), which result in defective 5′–3′-nuclease activity, occur in or near the predicted HhH motif (Fig. 3a). The Gly_{184}→Asp variant (*polA480ex*) has a markedly reduced 5′–3′-nuclease activity with little effect on polymerase activity, whereas the Gly_{192}→Asp variant (*polA214*) has a more pronounced effect on both polymerase and nuclease activities. *In vitro*, both these activities are thermolabile in this mutant, and *in vivo*, the mutation is lethal at high temperatures suggesting an essential role for this residue. It has recently been reported that substitution of the corresponding glycine residues by aspartates, in *B.caldotenax* pol I, results in the abolition of the 5′–3′ exonuclease activity (24). These data indicate that the pol I HhH motif is essential for 5′–3′ nuclease activity.

Table 3. HhH crossing angles

HhH structure	Helix 1 (residues)	Helix 2 (residues)	Crossing angle (degrees)
Pol β 1	56–61	67–75	154
Pol β 2	97–102	108–116	133
Endo III	108–113	119–127	130
AlkA	201–211	217–228	131
Taq Pol I	189–197	204–212	129

HhH motifs in pol β: determinants of HhH DNA-binding function

The crystallographic structure of pol β bound to a DNA template–primer (Fig. 3) (14) and the identification of putative HhH motifs has enabled us to predict the mode of DNA-binding for each putative HhH sequence. Seeberg *et al.* (19) suggested that the central residues G[I/V]G of the HhH hairpin bind to DNA through hydrophobic interactions with the bases in the grooves. However this proposal is not supported by the pol β structure. As discussed by Pelletier *et al.* (14), the pol β–DNA template complex structure reveals that pol β backbone nitrogens form non-specific hydrogen bonds with DNA phosphate oxygens. Of

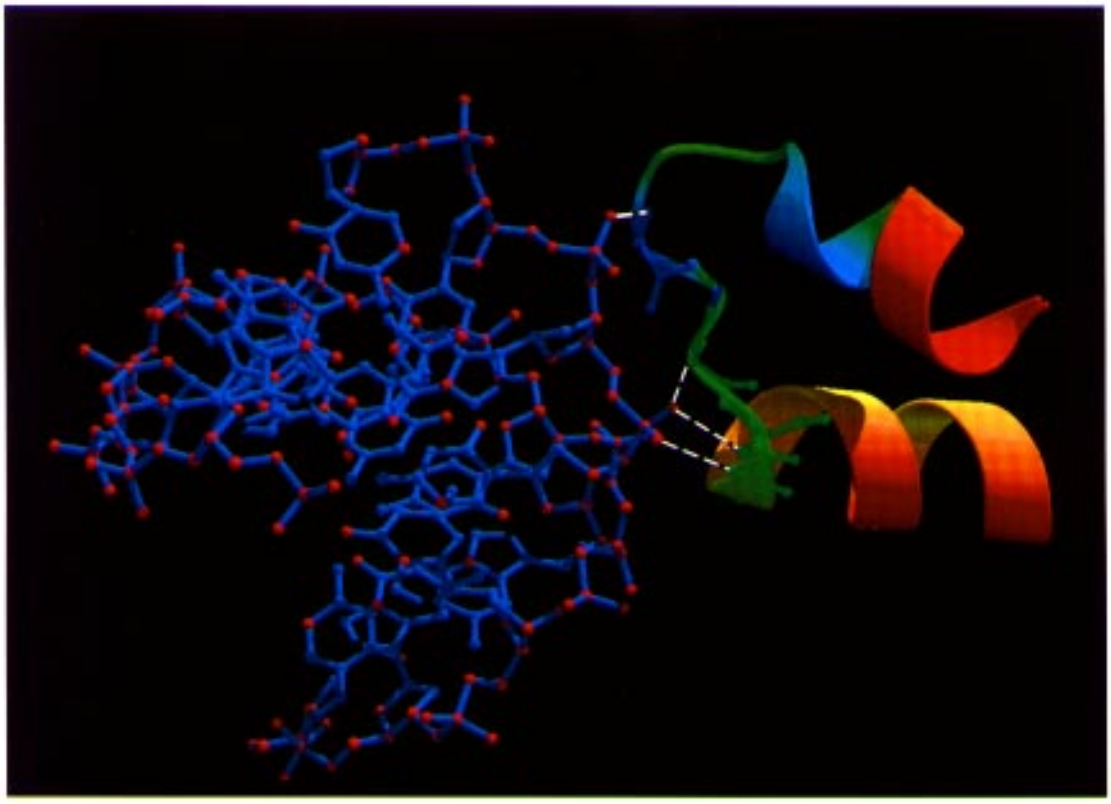

Figure 3. DNA recognition by the second pol β HhH motif. Non-specific hydrogen bond interactions between the backbone nitrogens of the second HhH motif of rat pol β and phosphate oxygens of DNA (primer strand) based on the co-crystal structure of pol β bound to a DNA template–primer (14). The HhH motif has been coloured according to the temperature factor (B factor) of the residues [red (high) to blue (low)]. This indicates that the DNA-bound hairpin loop is the most rigid region of the HhH motif while the helices are more flexible. The figure was prepared using RASMOL (13).

two regions in pol β involved in this interaction, four backbone nitrogens in the second HhH motif, between Gly105–Ala110 (HhH 8–13) form hydrogen bonds with phosphates of the primer strand (Fig. 3). It is suggested that all HhH motifs bind DNA in an analogous manner to that of the second pol β HhH motif. It is notable that there is a high propensity for glycine residues at HhH positions 8 and 10 (Table 1), which in pol β are critical for DNA-recognition and provide an extended surface for DNA–protein recognition. A high propensity for lysine at HhH 12 and threonine or serine at HhH 13 within a subset of proposed HhH sequences suggests that these interact with DNA phosphate groups in a similar manner to the same residues in P-loop structures (20). DNA recognition by HhH motifs, in the manner proposed above, would provide non-sequence-specific interactions of proteins with DNA. This type of interaction would contrast with the sequence-specific interactions of other motifs such as helix–turn–helix motifs (25).

A second HhH, predicted by our search, but not others (18,19), is located in the 8 kDa domain of pol β which appears to be responsible for the short-gap filling activity of the enzyme (26,27). The crystal structures of pol β (14,28) do not demonstrate DNA-binding to the 8 kDa domain, which is assumed to be a result of it adopting one of many non-productive conformations in the crystal. However, other experimental evidence strongly implicates the 8 kDa domain, and its HhH motif, in binding DNA. Kumar *et al*. (29) have demonstrated single-stranded (ss) DNA binding to the 8 kDa domain; this interaction is mediated by the two helices of the HhH motif as shown by nuclear magnetic resonance chemical shift data (30). Lys 72, located in the HhH motif, has been implicated in binding dNTP from the results of pyridoxal phosphate modification studies (31). The first putative HhH motif of the pol β homologue terminal deoxynucleotide transferase also appears to possess affinity for ssDNA (32) and nucleotides (33).

Recently it was reported that pol β can be specifically inhibited by its N-terminal 14 kDa domain (residues 1–140) (34) that contains both its HhH motifs. This domain, like the intact enzyme, binds both ss and double-stranded (ds) DNA but is deficient in polymerase activity. A smaller 8 kDa fragment (residues 3–75), encompassing its first HhH motif binds ss DNA but not ds DNA while another fragment containing the second motif and the catalytic domain (residues 87–334) binds only ds DNA. This evidence supports the prediction that the first HhH is important in ss DNA recognition and we propose that a minimal region containing both HhH motifs (residues 50–120) would also inhibit pol β activity by competing for the DNA substrate.

Homologues of the pol I 5′-nuclease domain

Structural and functional homology between the 5′-nuclease domain of pol I enzymes and a range of other nucleases has been reported (22,35). During the clustering of sequences into homologous classes it became apparent that these nuclease sequences show significant similarities to a family of endonucleases (36) that includes mammalian FEN1 (or DNase IV) and ERCC-5 (or XPG) and yeast RAD2 (Fig. 4). This observation corroborates the

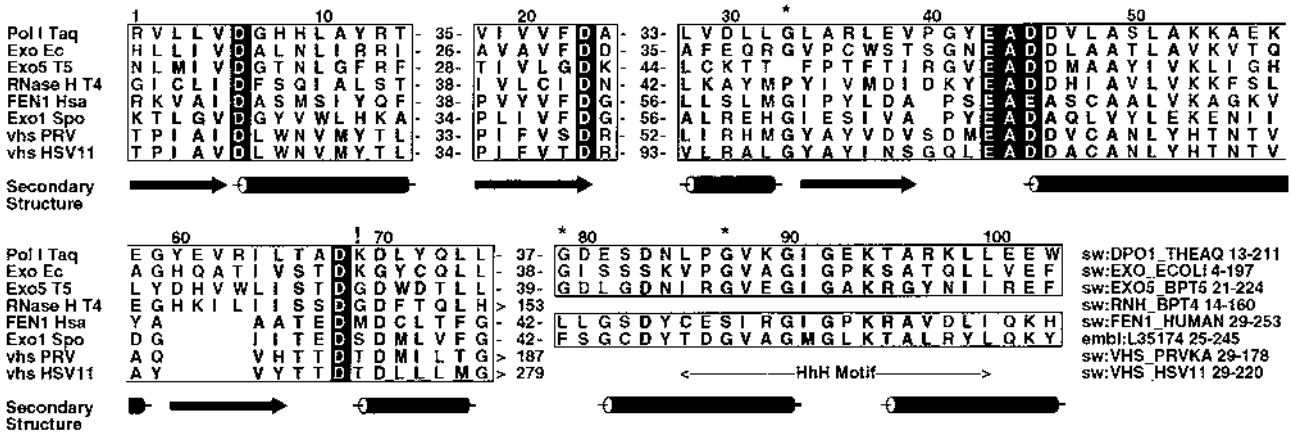

Figure 4. Alscript (45) representation of an alignment of sequences representing the *Taq* pol I 5′-nuclease, *E.coli* and bacteriophage T5 exonucleases, T4 RNase H, human flap endonuclease I (FEN1 or DNase IV), *S.pombe* exonuclease and two virion host shutoff (vhs) proteins from pseudorabies virus (PRV) and Herpes simplex virus (HSV) type 1, strain 17. Numbers represent the number of amino acids between sequence blocks. Accession codes and domain limits are given following the alignment. The alignment was constructed using MACAW (9). No putative HhH-like sequences could be found in vhs proteins; although a sequence homologous to pol I-like HhH motifs is present in T4 RNase H, it differs substantially from the HhH consensus sequence (22) and is not shown here. Positions marked with an asterisk are mutated in *E.coli* polymerase I variants with defective exonuclease activities (reviewed in 22). The position marked with an exclamation mark is that mutated in an HSV vhs variant with reduced function (46). The known secondary structure of *Taq* pol I (15) is represented below the alignment: cylinders represent α-helices, and arrows represent β-strands. Positions where the chemical character of residues are conserved in >66% of sequences are shaded; absolutely conserved aspartic acid and/or glutamic acid putative active site residues, and alanine residues at position 44, are shown as white-on-black.

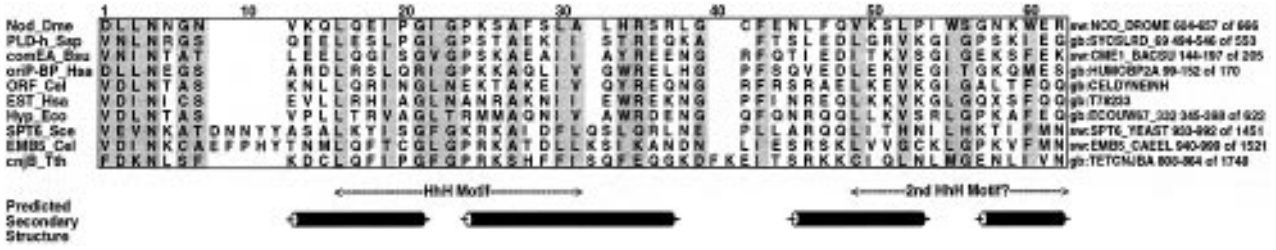

Figure 5. Multiple alignment of nod-like DNA-binding domain (NDD) sequences, displayed using Alscript (45). Positions where the chemical character of residues are conserved in >66% of sequences are shaded. Putative HhH motifs are apparent between positions 16–31, and possibly 49–62. Sequence similarities outside these regions, particularly positions 1–7, suggest that these sequences are homologous. A subset of these sequences exhibit significant sequence similarity (p-value = 1.6×10^{-10}) over a range of ~200 amino acids that encompasses the NDD; these are a hypothetical *E.coli* sequence (Hyp_Ec) and several eukaryotic sequences, including *S.cerevisiae* SPT6, the *C.elegans emb-5* gene product, *Tetrahymena thermophila* cnjB, a human expressed sequence tag (EST) and a *C.elegans* ORF. These observations are novel, excepting previously observed similarities between emb-5 and SPT6 (47), and oriP-BP and *comE* ORF A (4). Accession codes, domain limits and total number of amino acids in these sequences are given following the alignment. Secondary structure, predicted using PHD (11), is shown beneath the alignment, as Figure 4.

original finding of Robins *et al.* (37) that these two families are homologous. Scanning current sequence databases against a multiple alignment of FEN1- and pol I-like 5′-nuclease domain sequences using scanps (6), identified the family of alpha-herpesvirus virion host shutoff (vhs) proteins as candidate homologues. Sequence conservation, particularly in three alignment 'blocks', indicates that these three protein families are homologous (Fig. 4). The alignment shows conservation of several Asp and/or Glu residues that have been suggested to coordinate metal ions within the *Taq* polymerase I 5′-nuclease structure (14). The alphaherpesvirus vhs proteins are known to degrade host and viral mRNAs during infection and therefore have been proposed, although not shown, to function as nucleases (38). The observation of their sequence similarities to exonucleases is consistent with this proposal. Additional support is provided by a vhs mutant with defective activity which contains a substitution of a threonine with an isoleucine in a conserved tripeptide (Asp–Thr–Asp) containing two of the four proposed active site carboxylate groups. Interestingly, both vhs and phage T4 RNase H (39) appear to lack the HhH DNA-binding motif present in their homologous counterparts (Fig. 4).

A homologous family of DNA-binding domains

Sequence clustering also indicated that regions of *B.subtilis comE* ORF A, the *Synechocystis* sp. phospholipase D homologue, human oriP-BP, *D.melanogaster* nod and other proteins are homologous (Fig. 5) and are likely to possess DNA-binding functions. A subset of the putative nod-like DNA-binding domain (NDD) homologues possibly possess a second HhH motif (positions 49–62 in Fig. 5). The observation of one or two HhH motifs at the C-terminal end of nod is particularly intriguing. Nod is a kinesin-like molecule required for proper segregation of non-exchange chromosomes in female meiosis (40). Although the

NDD sequence has been shown not to be essential for the binding of nod to chromosomes (41,42), deletion of the C-terminal 12 residues of the NDD sequence at the nod C-terminus renders it non-functional (43).

A structural basis for non-sequence-specific DNA recognition

Prior to recent advances in structural biology it was evident that the α-helices could be accommodated in the major groove of B-DNA and therefore could mediate sequence-specific contacts with DNA bases (44). Elucidation of the structures of sequence-specific DNA-binding proteins has confirmed that α-helices of helix–turn–helix, zinc-finger, helix-loop–helix and leucine zipper motifs play an important role in DNA-recognition. In addition, α-helices are used commonly to orientate recognition helices enabling interaction with DNA.

In this paper we have presented evidence to support and extend the proposition of Thayer *et al.* (18) that the helix–hairpin–helix motif is a distinct and novel class of DNA binding motif. This highly conserved motif contains features characteristic of sequence-specific motifs such as helix–turn–helix (HtH) structures, namely the use of α-helices as a structural element for the correct orientation of a DNA-recognition element. However, although both are bi-helical structures the HhH motif differs from the HtH motif in its structure and its mode of recognition of DNA. By analogy with the DNA-bound structure of pol β (14) the HhH motif represents a novel structural motif involved in the non-sequence-specific recognition of both ss and ds DNA via hydrogen bond-mediated interactions with the DNA–phosphate backbone. Such interactions appear to be essential for the functions of many non-sequence-specific proteins, particularly those involved in base excision repair processes. Interestingly, on the occasions that the HhH motif is found in multiple copies, these are invariably separated by between 12 and 21 residues, suggesting that a particular spatial arrangement of HhH motifs may be required for multiple-sites of interaction with DNA. Future determination of further HhH-containing structures shall examine the proposition that this α-helical motif is prevalent among many base excision repair enzymes, in which it adopts a common structure and fulfills a common role.

ACKNOWLEDGEMENTS

We would like to thank Asim Siddiqui, Robert Russell and Geoff Barton for assistance in the use of STAMP, Hyeon Son for assistance with ACTIVE, and Steve Ashford for assistance with figures. We are indebted to Dr Tom Ellenberger for sending the AlkA coordinates prior to publication. We are grateful to Drs Soo Hyun Eom, Joe Jager and Tom Steitz for allowing access to the coordinates of *Taq* pol I. C.P.P. is an MRC Training Fellow, and a member of the Oxford Centre for Molecular Sciences, which is supported by EPSRC, BBSRC and the MRC. C.P.P. wishes to thank Dr C. M. Dobson for support and encouragement.

REFERENCES

1 Harrison, S.C. (1991) *Nature* **353**, 715–719.
2 Doherty, A.J., Worrall, A.F. and Connolly, B.A. (1995) *J. Mol. Biol.* **251**, 366–377.
3 Ponting, C.P. and Kerr, I.D. (1996) *Protein Sci.* **5**, 914–922.
4 Inamine, G.S. and Dubnau, D. (1995) *J. Bacteriol.* **177**, 3045–3051.
5 Doolittle, R.F. (1994) *Trends Biochem. Sci.* **19**, 15–18.
6 Barton, G.J. (1993) *Comput. Appl. Biosci.* **8**, 729–734.
7 Gibson, T.J., Hyvönen, M., Musacchio, A., Saraste, M. and Birney, E. (1994) *Trends Biochem. Sci.* **19**, 349–353.
8 Altschul, S.F., Boguski, M.S., Gish, W. and Wootton, J.C. (1994) *Nature Genet.* **6**, 119–129.
9 Schuler, G.D., Altschul, S.F. and Lipman, D.J. (1991) *Proteins Struct. Funct. Genet.* **9**, 180–190.
10 Henikoff, S. and Henikoff, J.G. (1992) *Proc. Natl Acad. Sci. USA* **89**, 10915–10919.
11 Rost, B. and Sander, C. (1993) *J. Mol. Biol.* **232**, 584–599.
12 Russell, R.B. and Barton, G.J. (1992) *Proteins Struct. Fuct. Genet.* **14**, 309–323.
13 Sayle, R.A. and Milnerwhite, E.J. (1995) *Trends Biochem. Sci.* **20**, 374–376.
14 Pelletier, H., Sawaya M.R., Kumar, A., Wilson, S.H. and Kraut, J. (1994) *Science* **264**, 1891–1903.
15 Kim, Y., Eom, S.H., Wang, J., Lee, D-S., Suh, S.W. and Steitz, T.A. (1995) *Nature* **376**, 612–616.
16 Kuo, C.-F., McRee, D.E.., Fisher, C.L., Cunningham, R.P. and Tainer, J.A. (1992) *Science* **258**, 434–440.
17 Efimov, A.V. (1991) *Protein Eng.*, **4**, 245–250.
18 Thayer, M.M., Ahern, H., Xing, D., Cunningham, R.P. and Tainer J.A. (1995) *EMBO J.* **14**, 4108–4120.
19 Seeberg, E., Eide, L. and Bjoras, M. (1995) *Trends Biol. Sci* **20**, 391–397.
20 Saraste, M., Sibbald, P.R. and Wittinghofer, A. (1990) *Trends Biochem. Sci* **15**, 430–434.
21 Bossemeyer, D. (1994) *Trends Biochem. Sci.* **19**, 201–205.
22 Gutman, P.D. and Minton, K.W. (1993) *Nucleic Acids Res.* **21**, 4406–4407.
23 Joyce, C.M., Fujii, D.M., Laks, H.S., Hughes, C.M. and Grindley, N.D.F. (1985) *J. Mol. Biol.* **186**, 283–293.
24 Ishino, Y., Takahashi-fugi, A., Uemori, T., Imamura, M., Kato, I. and Doi, H. (1995) *Protein Eng.*, **8**, 1171–1175.
25 Harrison, S.C. and Aggarwal, A.K. (1990) *Annu. Rev. Biochem.* **59**, 933–969.
26 Singhal, R. and Wilson, S. (1993) *J. Biol. Chem.* **268**, 15906–15911.
27 Prasad, R., Beard, W.A. and Wilson, S.H. (1994) *J. Biol. Chem.* **269**, 18096–18101.
28 Sawaya, M.R., Pelletier, H., Kumar, A., Wilson, S.H. and Kraut, J. (1994) *Science* **264**, 1930–1935.
29 Kumar, A., Widen, S.G., Williams, K.R., Kedar, P., Karpel, R.L. and Wilson, S.H. (1990) *J. Biol. Chem.* **265**, 2124–2131.
30 Liu, D., DeRose, E.F., Prasad, R, Wilson, S.H. and Mullen, G.P. (1994) *Biochemistry* **33**, 9537–9545.
31 Basu, A., Kedar, P, Wilson, S.H. and Modak, M.J. (1989) *Biochemistry* **28**, 6305–6309.
32 Farrar, Y.J.K., Evans, R.K., Beach, C.M. and Coleman, M. (1991) *Biochemistry* **30**, 3075–3082.
33 Pandey,V.N. and Modak, M.J. (1988) *J. Biol. Chem.* **263**, 3744–3751.
34 Husain, I., Morton, B.S., Beard, W.A., Singhal, R.K., Prasad, R., Wilson, S.H. and Besterman, J.M. (1995) *Nucleic Acids Res.* **23**, 1597–1603.
35 Sayers, J.S. and Eckstein, F. (1991) *Nucleic Acids Res.* **19**, 4127–4132.
36 Harrington, J.J. and Lieber, M.R. (1994) *Genes Dev.* **8**, 1344–1355.
37 Robins, P., Pappin, D.J.C., Wood, R.D. and Lindahl, T. (1994) *J. Biol. Chem.* **269**, 28535–28538.
38 Pak, A.S., Everly, D.N., Knight, K. and Read, G.S. (1995) *Virology* **211**, 491–506.
39 Hollingsworth, H.C. and Nossal, N.G. (1991) *J. Biol. Chem.* **266**, 1888–1897.
40 Carpenter, A.T.C. (1973) *Genetics* **73**, 393–428.
41 Afshar, K., Barton, N.R., Hawley, R.S. and Goldstein, L.S.B. (1995) *Cell* **81**, 129–138.
42 Afshar, K., Scholey, J. and Hawley, R.S. (1995) *J. Cell Biol.* **131**, 833–843.
43 Rasooly, R., Zhang, P., Tibolla, A.K. and Hawley, R.S. (1994) *Mol. Gen. Genet.* **242**, 145–151.
44 Zubay, G. and Doly, P.J. (1959) *J. Mol. Biol.* **7**, 1–10.
45 Barton, G.J. (1993) *Protein Eng.*, **6**, 37–40.
46 Berthomme, H., Jacquemont, B. and Epstein, A. (1993) *Virology* **193**, 1028–1032.
47 Nishiwaki, K., Sano, T. and Miwa, J. (1993) *Mol. Gen. Genet.* **239**, 313–322.