Computers Math. Applic. Vol. 20, No. 4–6, pp. 25–48, 1990
Printed in Great Britain

0097-4943/90 $3.00 + 0.00
Pergamon Press plc

GENETIC SEQUENCES OF HORMONE RESPONSE ELEMENTS SHARE SIMILARITY WITH PREDICTED ALPHA HELICES WITHIN DNA BINDING DOMAINS OF STEROID RECEPTOR PROTEINS: A BASIS FOR SITE-SPECIFIC RECOGNITION

L. F. HARRIS,† M. R. SULLIVAN and D. F. HICKOK
Cancer Research Laboratory, Abbott-Northwestern Hospital, Minneapolis, MN 55407, U.S.A.

Abstract—The 150 amino acid sequence comprising the DNA binding region of rat glucocorticoid receptor protein, RGRDBR, was compared to amino acid sequences of members of the superfamily of eukaryotic DNA regulatory proteins. Maximal similarity fell within the 86 amino acid sequence of RGRDBR reported to contain both DNA binding and transcription regulating properties and within the reported DNA binding regions of those proteins to which it was compared. Chou–Fasman secondary structure predictions within these DNA binding domains revealed a conserved alpha helix–beta turn–alpha helix motif. The 450 nucleotide sequence comprising the complementary DNA (cDNA) of amino acids making up RGRDBR was compared to a nucleotide sequence (−312 to −38) from mouse mammary tumor virus 5′ long terminal repeat, MMTV5LTR, known to contain glucocorticoid response elements (GREs). The maximally similar subsequence was found within the coding region for predicted alpha helix B of RGRDBR (nucleotides 1376 to 1412) and within a reported GRE of MMTV5LTR (nucleotides −199 to −131). This MMTV5LTR GRE sequence contains an imperfect palindrome of TGTTCT which is the specific recognition motif for DNA binding by both glucocorticoid and progesterone receptors. Since there are multiple coding possibilities for the majority of the 20 known amino acids, the exceptions being methionine and tryptophan which have a single codon, to thoroughly investigate the extent of genetic information conserved between RGRDBR and GRE, we converted this MMTV5LTR GRE nucleotide subsequence (−199 to −131) to amino acids in all three reading frames reading rightward and leftward in both strands. This procedure revealed all coding possibilities within the MMTV5LTR nucleotide subsequence, as well as the location of the codon sites. A comparison of these MMTV5LTR amino acid coding possibilities to RGRDBR predicted helix B amino acids revealed highly conserved genetic information localized within the GRE half-sites, predominantly in the right half-site containing the TGTTCT sequence. In the absence of atomic coordinates for eukaryotic DNA regulatory proteins, a computer model of a eukaryotic/procaryotic hybrid protein was created with RGRDBR predicted helix B replacing helix F of *E. coli* cAMP-dependent regulatory protein (CRP) for which coordinates from X-ray crystallography were available. This hybrid protein was docked onto MMTV5LTR at the region of maximal similarity to helix B. Our computer model shows that the side chains of amino acids within RGRDBR helix B are oriented toward, and appear to be capable of interacting with, nucleotides on both strands of their respective codons within a functional GRE. Calculations of H-bonding in this model indicate that amino acids of helix B are forming H-bonds with nucleotides of their cognate codon/anti-codon sites within the major grooves of the GRE half-sites.

INTRODUCTION

Steroid hormone receptors are members of a superfamily of DNA regulatory proteins [1, 2]. Amino acid sequences are now known for representatives of the three major classes of steroid hormone receptor proteins [3–11]. DNA binding domains among the steroid hormone receptor proteins share amino acid sequence similarity rich in cysteine, lysine and arginine residues [1]. In addition, within the DNA binding domains, a motif of repeating cysteine residues with a pattern reminiscent of the arrangement seen in the *Xenopus laevis* transcription factor, TFIIIA, is also conserved [12]. The rat glucocorticoid receptor protein has been well-characterized; a region of 150 amino acids comprising the DNA binding domain RGRDBR (407 to 556) mediates constitutive enhancement and reportedly contains a sequence of 86 amino acids (440 to 525) with both DNA binding and transcriptional enhancing properties [3, 13].

Steroid receptor proteins are reported to enhance gene transcription by interacting as dimers with short palindromic nucleotide sequences termed hormone response elements (HREs) [1, 2, 14–17]. These observations concur with findings for the procaryotic DNA regulatory proteins' interaction at operator sites [18]. As seen in procaryotic operators, hormone response elements have also been

†To whom all correspondence should be addressed.

shown to constitute a family of related nucleotide sequences [19, 20]. For example, perfect or imperfect palindromes around the hexanucleotide motif TGT T/C CT serve as functional response elements for both glucocorticoid and progesterone receptors in several genes [21, 22]. Androgen receptor also binds to the TGTTCT motif [23]. The TGTTCT motif is also found in operators of the phages 434 [24] and P22 [25]. Recently, a functional estrogen response element (ERE) was shown to be closely related to the glucocorticoid response element (GRE); plasmids in which the hexanucleotide (TGTTCT), the main binding motif of GRE, was altered to TGACCT in both halves of a perfect 15 base pair palindrome, switched gene induction from glucocorticoid to estrogen receptor regulation [20].

Although considerable information is known, the recognition mechanism of site-specific DNA binding leading to transcriptional enhancement by steroid receptor proteins has not been clearly defined. Knowledge of steroid receptor proteins' secondary structure, along with identification of specific amino acids which interact with nucleotides within HREs, is needed to clarify this aspect of gene control. We used computer-derived comparisons of amino acid sequences, predicted secondary structures and hydropathy profiles among representatives of this superfamily of DNA regulatory proteins. We found that an alpha helix–beta turn–alpha helix secondary structural motif, as seen in procaryotic DNA regulatory proteins, is conserved within the DNA binding domains of eukaryotic DNA regulatory proteins. We also compared cDNA which codes for amino acids of the DNA binding domains of steroid receptor proteins with a nucleotide sequence from mouse mammary tumor virus 5′ long terminal repeat (MMTV5LTR) which contains specific functional HREs. We observed that genetic information is conserved between HREs on the DNA and alpha helices within the DNA binding domains of these DNA regulatory proteins. This conservation of genetic information may be the basis for DNA site-specific recognition by DNA regulatory proteins. Finally, we used a computer-based molecular modeling and display tool (QUANTA) to prepare protein and DNA structural models in order to visualize and simulate protein–DNA interaction.

EXPERIMENTAL PROCEDURES

Nucleotide sequence data was taken from cited references and GenBank, a computer database of DNA and RNA sequences distributed by IntelliGenetics Inc. (700 East El Camino Real, Mountain View, CA 94020). Amino acid sequences were taken from cited references and the Protein Identification Resource (PIR), a computer database of amino acid sequences distributed by the National Biomedical Research Foundation (Georgetown University Medical Center, 3900 Reservoir Rd N.W., Washington, DC 20007).

LOCAL is a program which searches for maximally similar subsequences between any two amino acid or nucleic acid sequences using a dynamic programming matrix algorithm [26]. Matrices can be constructed to weight the similarity values of all amino acids based on their known properties. We used a distance matrix based on the Kyte and Doolittle [27] hydropathic index to identify hydropathically similar residues among selected proteins. Gap weighting and mismatch values used were: 1.0 for matches, 0.9 for mismatches and $-(0.9 + 1.01 * \text{length})$ for gaps. The LOCAL program runs on various machines from personal computers to supercomputers. LOCAL was run on a Cray-2 supercomputer at the Minnesota Supercomputer Center in Minneapolis, in a UNICOS environment (a Cray research UNIX operating system; UNIX is a trademark of AT&T). LOCAL is an academic software package distributed by the Harvard Medical School Molecular Biology Computer Research Resource (MBCRR) (Dana-Farber Cancer Institute, Harvard School of Public Health, 44 Binney St JF815, Boston, MA 02115).

PRSTRC identifies potential protein secondary structural domains using the Chou–Fasman pseudo probabilities algorithm [28]. Structures predicted are alpha helix, beta sheet, beta turn and omega loop. PRSTRC was run on the Cray-2 supercomputer and is also distributed by the MBCRR.

HYPHO (a code developed by L. Harris and M. Fenton) combines hydropathic profiles with local sequence alignments of proteins in a single graphic. Hydropathic values for each amino acid were as published by Kyte and Doolittle [27] and are shown as vertical deflections. The HYPHO program was written in C language and runs on a Silicon Graphics Iris 3130 workstation.

QUANTA is a molecular modeling and display tool developed by Polygen Corp. (200 Fifth Ave, Waltham, MA 02254). QUANTA allows the construction of molecular models of DNA sequences, point mutations of existing models and the modeling of small peptides with a selected secondary structure. The RGRDBR alpha helix B atomic coordinates were computed using the BUILD module of the QUANTA program. This module allows the construction of molecular models of small peptides and folds them into a selected secondary structure. This module was also used to generate coordinates for the MMTV5LTR fragment used in Figs 5 and 6. The cAMP-dependent regulatory protein (CRP) coordinates from crystallography data were provided by Dr Irene Weber [29]. CRP helix F excised from this coordinate set had an identical backbone conformation to the same helix F generated by QUANTA. QUANTA was also used to dock RGRDBR helix B in place of helix F of CRP forming the hybrid protein model for structural refinement. QUANTA was also used to interactively manipulate protein/DNA complexes and calculate H-bonds. The following H-bonding parameters were used: for interactions with known H-atom positions, we used a maximum H-bond distance of 2.5 Å, but we also included near H-bonds where the bond length was $\geqslant 2.5$ Å, but < 3.0 Å (indicated by an asterisk) if the amino acid was oriented toward its codon. For interactions without explicit H-positions, we used a maximum H-bond distance of 3.5 Å, but we also included near H-bonds where the bond length was $\geqslant 3.5$ Å, but < 4.0 Å (indicated by an asterisk) if the amino acid's side chain was oriented toward its codon. QUANTA was run on the Silicon Graphics Iris 3130 workstation.

AMBER is a general-purpose molecular mechanics and dynamics program designed for refinement of macromolecular conformations using an analytical potential energy function [30–32]. The AMBER prep modules LINK, EDIT and PARMV were run on the Iris workstation and the minimization module (MIN) was run on the Cray-2 supercomputer. We used the LINK and EDIT modules to define proper connectivity for the hybrid RGRDBR helix B/CRP protein. Atomic parameters were prepared for vector processing using the PARMV module. The hybrid protein was structurally refined, *in vacuo*, using the MIN module. A distance-dependent dielectric function was used. The non-bonded cutoff was set at 12 Å. The r.m.s. gradient criterion was set to a value of 0.1. No periodic boundary conditions were applied. Non-bonded interactions were calculated using a residue-based cutoff. The initial step-length was 0.001 and non-bonded interactions were updated every 10 steps. 550 Steps of a steepest decent method of minimization were run using approx. 1000 s of CPU time on the Cray-2. The final conformation of the minimized hybrid protein was inspected visually and observed to maintain a reasonable geometry. AMBER is an academic software package distributed by the Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, CA 94143.

RESULTS AND DISCUSSION

We compared the amino acid sequences of representative steroid receptor proteins within the superfamily of DNA regulatory proteins to the 150 amino acid DNA binding and transcriptional enhancing domain of RGRDBR; see Fig. 1. As expected, the region of maximal subsequence similarity to RGRDBR was found within the DNA binding domain of all steroid receptor proteins, extending and confirming the observations of others as to the conservation of similar amino acid sequences within this region [1, 2]. Interestingly, the greatest similarity to RGRDBR among the proteins was found within amino acids ranging from 440 to 525 of RGRDBR, the 86 amino acid sequence reported to have both DNA binding and transcriptional enhancing properties [3].

To further characterize the conserved maximally similar subsequences found within the DNA binding domain of these proteins, we compared these sequences for relationship as a function of secondary structural prediction and hydropathy. The results are displayed in Fig. 2 as non-averaged hydropathy values for each amino acid with local amino acid sequence alignment determined without gaps or deletions, along with a secondary structural prediction. A strong conservation of hydropathy among these sequences is apparent. Perhaps more significant are the findings using an algorithm developed by Chou and Fasman for secondary structure prediction [28]. Within the DNA binding domains of these eukaryotic proteins, an alpha helix–beta turn–alpha helix motif is conserved. A similar structural motif is seen conserved in procaryotic DNA regulatory proteins [33–35].

```
RGRDBR AND HUMGCRA  FOR 407-556 & 387-539 THE SIMILARITY VALUE OF 207.27
WAS LOCATED FOR THE MAXIMALLY SIMILAR SUBSEQUENCE
WHICH CONTAINS 144 MATCHES AND 6 MISMATCHES(+GAPS) WITH 3 RELATED AMINO ACIDS.

SVFSNGYSSPGMRPDVSSPPSSSSAATGPPPKLCLVCSDEASGCHYGVLTCGSCKVFFKRAVEGQHNYLCAGRNDCIIDKIRRKNCPACRYRKCLQAGMNLEARKTKKKIKGIQQATAGVQDTSENPNKTIVPAALPQLTPTLVSLLEVI
+*********+************** ******************************************************************************************* ***+************ ***************
TVFSNGYSSPSMRPDVSSPPSSSSATTGPPPKLCLVCSDEASGCHYGVLTCGSCKVFFKRAVEGQHNYLCAGRNDCIIDKIRRKNCPACRYRKCLQAGMNLEARKTKKKIKGIQQATTGVQETSENPNKTIVPATLPQLTPTLVSLLEVI
                        T                                                                                                 S      G

RGRDBR AND HUMMINR  FOR 421-521 & 583-686  THE SIMILARITY VALUE OF 113.49
WAS LOCATED FOR THE MAXIMALLY SIMILAR SUBSEQUENCE
WHICH CONTAINS  80 MATCHES AND 20 MISMATCHES(+GAPS) WITH 16 RELATED AMINO ACIDS

        DVSSPPSSSSAATGPPPKLCLVCSDEASGCHYGVLTCGSCKVFFKRAVEGQHNYLCAGRNDCIIDKIRRKNCPACRYRKCLQAGMNLEARKTKKKIKGIQQ
        +***++* *+  +++*+*+****+**********+***************************************** +********* ***+***+***++
        NVSSSTSVST-GSSRPSKICLVCGDEASGCHYGVVTCGSCKVFFKRAVEGQHNYLCAGRNDCIIDKIRRKNCPACRLQKCLQAGMNLGARKSKKKLKGIHE
              LR                                                                                     LG

RGRDBR AND QRHUP  FCR 436-513 & 563-641 THE SIMILARITY VALUE OF 93.50
WAS LOCATED FOR THE MAXIMALLY SIMILAR SUBSEQUENCE
WHICH CONTAINS 67 MATCHES AND 12 MISMATCHES(+GAPS) WITH 7 RELATED AMINO ACIDS

                PPKLCLVCSDEASGCHYGVLTCGSCKVFFKRAVEGQHNYLCAGRNDCIIDKIRRKNCPACRYRKCLQAGMNLEARKTKK
                *+*+**+*+***********************+***************+************ ***+**** *  ** **
                PQKICLICGDEASGCHYGVLTCGSCKVFFKRAMEGQHNYLCAGRNDCIVDKIRRKNCPACRLRKCCQAGMVLGGRKFKK

RGRDBR AND HARDBR   435-511 & 537-613 THE SIMILARITY VALUE OF  79.50
WAS LOCATED FOR THE MAXIMALLY SIMILAR SUBSEQUENCE
WHICH CONTAINS  52 MATCHES AND 14 MISMATCHES(+GAPS) WITH 10 RELATED AMINO ACIDS

                PPPKLCLVCSDEASGCHYGVLTCGSCKVFFKRAVEGQHNYLCAGRNDCIIDKIRRKNCPACRYRKCLQAGMNLEARK
                **+* **+*+*********+*************+**+++****+**** ***+****** ** *** +*** * ***
                PPQKTCLICGDEASGCHYGALTCGSCKVFFKRAAEGKQKYLCASRNDCTIDKFRRKNCPSCRLRKCYEAGMTLGARK

RGRDBR AND HUMESTR  FOR 440-505 & 185-250 THE SIMILARITY VALUE OF 41.68
WAS LOCATED FOR THE MAXIMALLY SIMILAR SUBSEQUENCE
WHICH CONTAINS  39 MATCHES AND 26 MISMATCHES(+GAPS) WITH 16 RELATED AMINO ACIDS

                  CLVCSDEASGCHYGVLTCGSCKVFFKRAVEGQHNYLCAGRNDCIIDKIRRKNCPACRYRKCLQAGM
                  *+** *+*** **** +* +**+**** ++*+++*+**+ *+* *** *** *+*** *** ++**
                  CAVCNDYASGYHYGVWSCEGCKAFFKRSIQGHNDYMCAT-NQCTIDKNRRKSCQACRLRKCYEVGM
                                                      P

RGRDBR AND HUMRETR  FOR 411-505 & 62-153 THE SIMILARITY VALUE OF  32.37
WAS LOCATED FOR THE MAXIMALLY SIMILAR SUBSEQUENCE
WHICH CONTAINS  41 MATCHES AND 48 MISMATCHES(+GAPS) WITH 32 RELATED AMINO ACIDS

  NGYSSPGMRPDVSSPPSSSSAATGPPPKLCLVCSDEASGCHYGVLTCGSCKVFFKRAVEGQHNYLCAGRNDCIIDKIRRKNCPACRYRKCLQAGM
  +++**++ ++ *+****++++   *++* *+** *+ ** ****  * +** **+*++  *+** *   ++***+*+ *+ *  ** +** ++**
  ETQSSSS-EEIVPSPPSPPPL---PRYKPCFVCQDKSSGYHYGVSACEGCKGFFRRSI--QKNYTCHRDKNCIINKVTRNRCQYCRLQKCFEVGM
                           I                                    MV

RGRDBR AND AEVERBA  FOR 440-513 & 37-115  THE SIMILARITY VALUE OF  28.88
WAS LOCATED FOR THE MAXIMALLY SIMILAR SUBSEQUENCE
WHICH CONTAINS  33 MATCHES AND 42 MISMATCHES(+GAPS) WITH 28 RELATED AMINO ACIDS

                  CLVCSDEASGCHYGVLTCGSCKVFFKRAVEGQHNYLCAGRNDCIIDKIRRKNCPACRYRKCLQAGMNLEARKTKK
                  *+**+*+*+* ** ++** +** **+* ++  * * *     *+**** *++*  **++**+ +**+*    +*+
                  CVVCGDKATGYHYRCITCEGCKSFFRRTIQKNHPYSCTYDGCCVIDKITRNQCQLCRFKKCISVGMDLVLDDSKR
                                                 L T                              MA

RGRDBR AND CVDRDBR  FOR 438-505 & 1-68 THE SIMILARITY VALUE OF  21.08
WAS LOCATED FOR THE MAXIMALLY SIMILAR SUBSEQUENCE
WHICH CONTAINS  26 MATCHES AND 39 MISMATCHES(+GAPS) WITH 28 RELATED AMINO ACIDS

                 KLCLVCSDEASGCHYGVLTCGSCKVFFKRAVEGQHNYLCAGRNDCIIDKIRRKNCPACRYRKCLQAGM
                 ++* **+*+*+* ++** +** **+*     + ++++++++ ** * * +*+++*** +*+++*+**
                 RICGVCGDRATGFHFNAMTCEGCKGFFRR---SMKRKAMFTNGDCKITKDNRRHCQACRLKRCVDIGM
                                                         CPF
```

Fig. 1 (legend opposite).

A summary of Fig. 2 structural predictions is shown in Table 1. The position of the alpha helices, beta turns and omega loops is conserved among the proteins. A high degree of amino acid sequence identity is also conserved within these structures, and amino acid substitutions are predominantly hydropathically similar. HUMESTR, HUMRETR and CVDRDBR had only a loop predicted instead of both a helix and loop in the region designated alpha helix A. There was, however, a helix predicted in close proximity, as shown for HUMRETR and CVDRDBR. Suprisingly, HARDBR showed another alpha helix in the region between helix B and D, which we designated helix C. Although not predicted, a similar sequence is found in the other proteins (Fig. 2), leading us to believe that these proteins may have a similar helical motif to the androgen receptor. Beta turns were also predicted in the helix A region for all proteins except CVDRDBR, which had only a loop. There was also a conserved beta turn 2–helix B–beta turn 4–helix D motif found in all proteins except HUMESTR, AEVERBA and CVDRDBR, which had a beta turn 3 predicted instead of a beta turn 4. AEVERBA did not have a loop or a helix predicted in the helix B region; however, a loop was predicted adjacent to the helix B region which overlapped the helix C region. Beta turns 2 and 4 consist predominantly of four amino acids with the pattern CxxC. Within the helix B region, lysine, arginine, glutamic acid, serine and alanine residues are conserved. These amino acids are also frequently found in alpha helices of procaryotic DNA regulatory proteins and are reportedly involved in DNA binding [33, 36]. Recent observations [37] describing specific amino acid residues comprising alpha helices in eukaryotic proteins for which crystallographic coordinates are available agree with residues found in our putative alpha helices, thus supporting our secondary structural predictions. The frequency of occurrence of certain amino acids at particular locations within known alpha helices was also recently reported [38]. The amino acids serine, lysine and glycine in the amino, middle and carboxyl locations, respectively, are in agreement with the positions of those residues in predicted helix B, and to a lesser extent in predicted helix D.

Flanking helix B on the amino end we observed a predicted beta turn with a CGSC or CEGC sequence. The beta turn 2 CGSC distinguishes glucocorticoid, mineralocorticoid, progesterone and androgen receptors from estrogen, retinoic acid, v-erb A and vitamin D receptors which have CEGC for beta turn 2 (see Table 1). During final preparation of this paper, a report appeared in the literature which described exchanging GS to EG within the CxxC regions above. This substitution of amino acids switched the specificity of the DNA response element activation from glucocorticoid receptor (GR) to estrogen receptor (ER) [39]. Flanking helix B on its carboxyl flank a predicted loop structure, HNYLC, is found. The tyrosine (Y) within the predicted loop HNYLC is flanked by glutamic acid and basic amino acids arginine and lysine residues within helix B spaced 5, 8 and 9, respectively, on the amino side of tyrosine. Tyrosines found in other proteins flanked in a similar manner by these amino acids are reported to be phosphorylated by tyrosine protein kinase [40]. Similar sequences were observed in the same location in all steroid receptor proteins except HUMESTR, which had glutamine substituted for glutamic acid. HUMRETR, AEVERBA and CVDRDBR did not conform to this pattern. Recent point mutation studies exchanging glycine for the tyrosine residue above, show a highly significant reduction in DNA binding by GR [41].

The CGSC and the HNYLC sequences are similar to sequences believed to be involved in forming putative zinc-activated DNA binding "finger" structures [42–47]. A structure in which a

(Figure opposite.)

Fig. 1. The maximally similar amino acid subsequences of several hormone receptor proteins were derived by local sequence comparison with rat GR DNA binding region. Amino acids are numbered from the transcription start site of the proteins. The abbreviations represent the following proteins: RGRDBR, rat glucocorticoid receptor DNA binding region [3]; HUMGCRA, human glucocorticoid receptor A [4]; HUMMINR, human mineralocorticoid receptor [5]; QRHUP, human progesterone receptor [6]; HARDBR, human androgen receptor DNA binding region [7] (Chang, personal communication, 1988); HUMESTR, human estrogen receptor [8]; HUMRETR, human retinoic acid receptor [9]; AEVERBA, v-erb A oncogene product [10]; and CVDRDBR, chicken vitamin D receptor DNA binding region [72]. The Dayhoff [73] one-letter amino acid code is as follows: A = alanine, C = cysteine, D = aspartic acid, E = glutamic acid, F = phenylalanine, G = glycine, H = histidine, I = isoleucine, K = lysine, L = leucine, M = methionine, N = asparagine, P = proline, Q = glutamine, R = arginine, S = serine, T = threonine, V = valine, W = tryptophan and Y = tyrosine. Identical amino acids are starred (*) and hydropathically similar amino acids are indicated by a plus (+) sign. These maximally similar subsequences are shown in decending order of similarity to RGRDBR based on a local sequence comparison [26] with Needleman–Wunsch similarity values [74]. Gaps are indicated by a dash and deletions are subscripted.

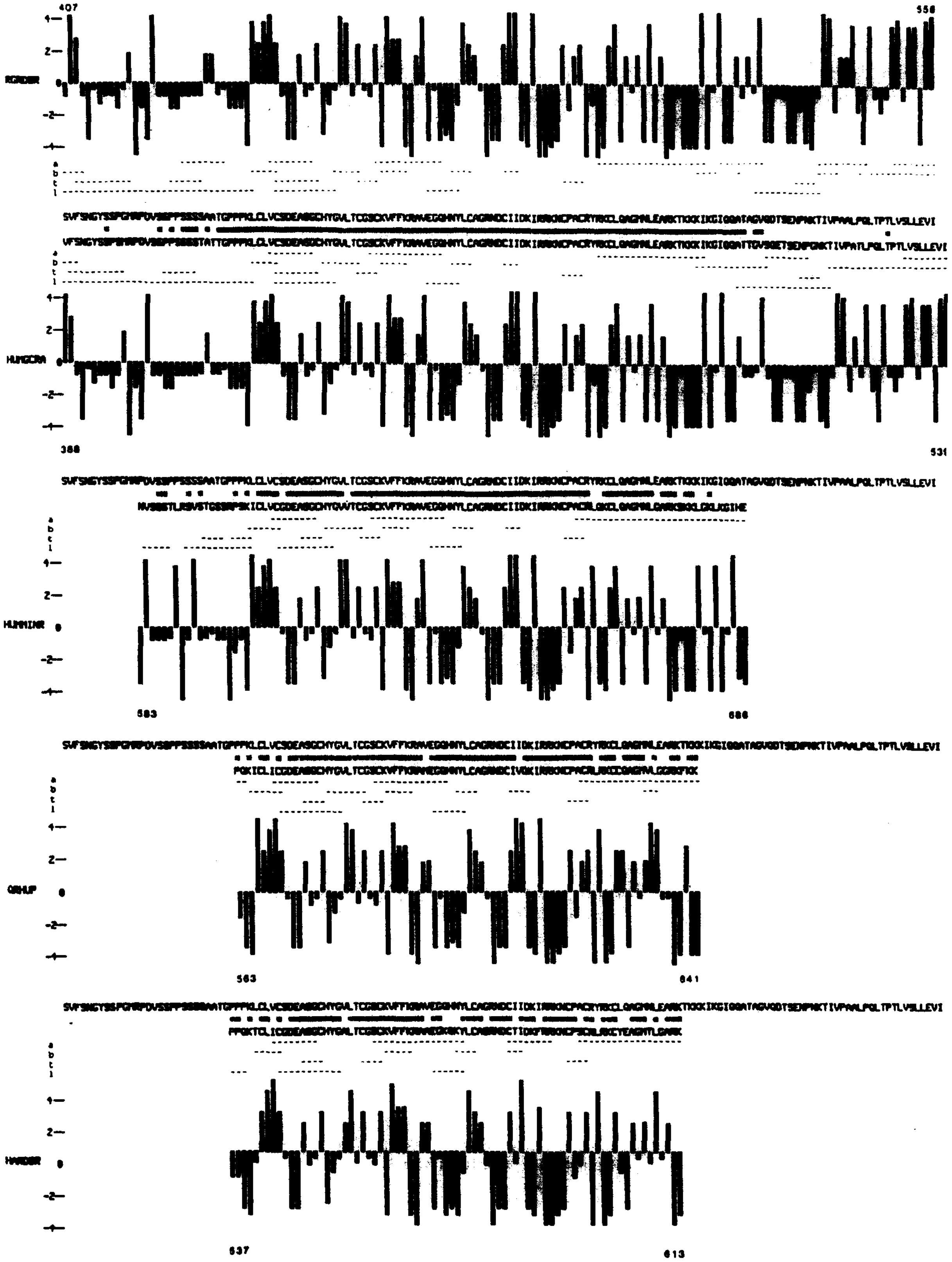


Fig. 2(*continued opposite.*)

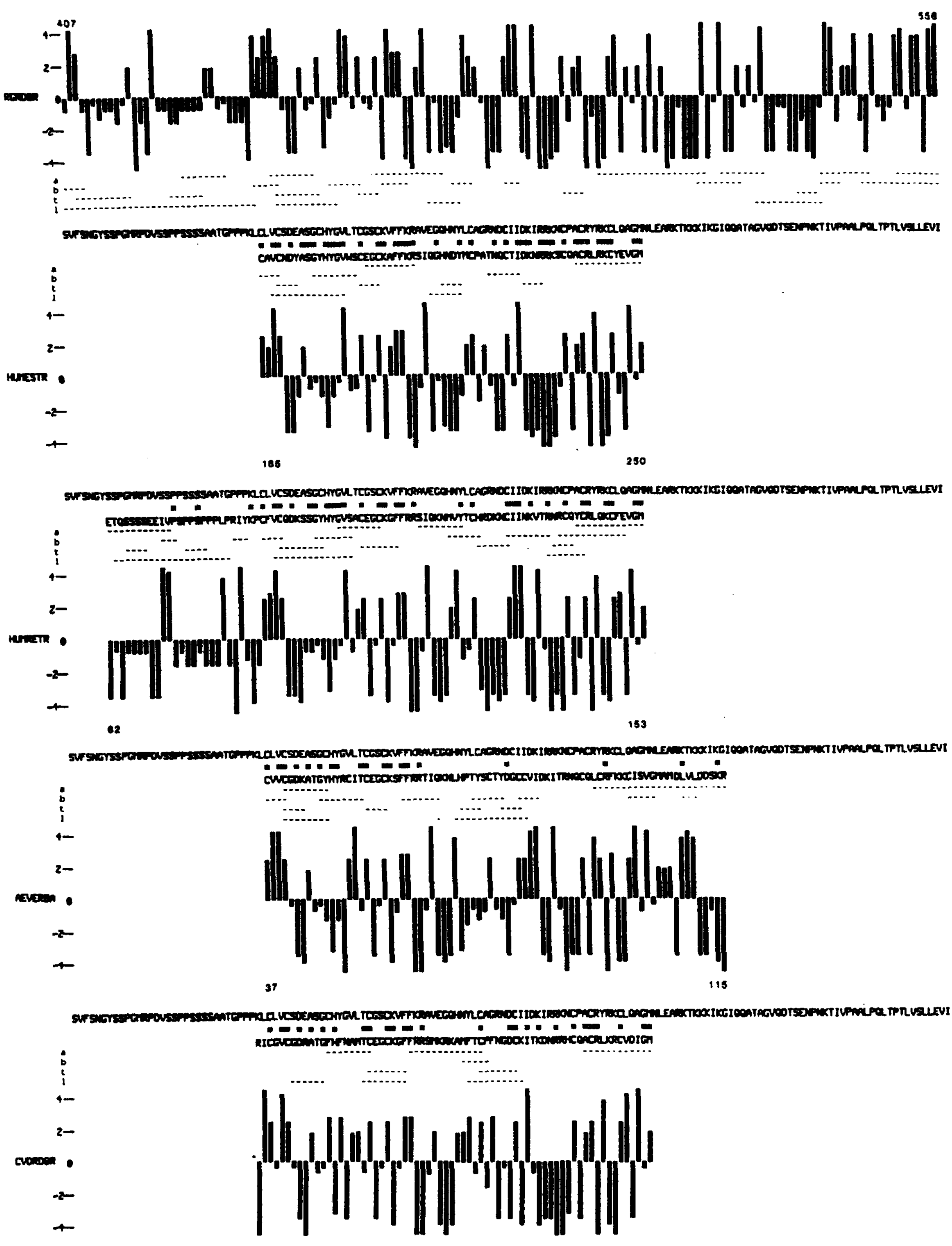


Fig. 2. The hydropathic profiles of the maximally similar subsequences from Fig. 1 were compared to RGRDBR using HYPHO. These subsequences were aligned for best matching without gaps or deletions. The hydropathic values were not averaged so the deflections of each amino acid could be compared (negative values are hydrophilic). The predicted secondary structures from PRSTRC (a = alpha helix, b = beta sheet, t = beta turn, l = omega loop) are also shown as dashed lines below the amino acid sequences. Abbreviations for the proteins are as in Fig. 1.

Table 1. Summary of Chou–Fasman secondary structure predictions of alpha helices, beta turns and *omega* loops from Fig. 2

	Alpha helix A	Alpha helix B	Alpha helix C	Alpha helix D
RGRDBR	VCSDEASG	SCKVFFKRAVEGQ		RKCLQAGMNLEARKTKKKI
HUMGCRA	VCSDEASG	SCKVFFKRAVEGQ		RKCLQAGMNLEARKTKKKI
HUMMINR	VCGDEASG	SCKVFFKRAVEGQ		ACRLQKCLQAGMNLGARKSKKLGKLKGIHE
QRHUP	ICGDEASG	SCKVFFKRAMEGQ		ACRLRKCCQAGMVLGGRKFKK
HARDBR	ICGDEASG	SCKVFFKRAAEGKQKY	CTIDKFRR	SCRLRKCYEAGMTLGARK
HUMESTR		EGCKAFFKR		ACRLRKCYEVGM
HUMRETR	GVSACEGC	SIQKNMVY		YCRLQKCFEVGM
AEVERBA	CGDKATGY			LCRFKKCISVGMAMDLVLDDSKR
CVDRDBR	FHFNAMTC	FRRSMKRKAMFTC		ACRLKRCVDIGM

	Beta turn 1	Beta turn 2	Beta turn 3	Beta turn 4
RGRDBR	CSDEASGC	CGSC		CPAC
HUMGCRA	CSDEASGC	CGSC		CPAC
HUMMINR	ASGC	CGSC		CPAC
QRHUP	ASGC	CGSC		CPAC
HARDBR	ASGC	CGSC		CPSC
HUMESTR	CNDY	CEGC	HNDY DKNR	none
HUMRETR	CQDKSSGY	CEGCKGF	HRDKNC	NRCQ
AEVERBA	CGDK	CEGC	HPTY YDGC	none
CVDRDBR	none	CEGCKGF	CPFNGDC	none

	Omega loop 1	Omega loop 2	Omega loop 3	Omega loop 4
RGRDBR	CSDEASGCHYG		EGQHNY	
HUMGCRA	CSDEASGCHYG		EGQHNY	
HUMMINR	CGDEASGCHYG		EGQHNY	
QRHUP	CGDEASGCHYG		EGQHNY	
HARDBR	CGDEASGCHYG		EGKQKY	
HUMESTR	VCNDYASGYHYGV		QGHNDY	
HUMRETR	VCQDKSSGYHYGVS			RNRCQYC
AEVERBA	CGDKATGY		LHPTYSCTYDGCC	
CVDRDBR	GDRATG	TCEGCKGF	FTCPFNGDCK	

"zinc finger" domain folds into a conformation with an antiparallel beta ribbon and an alpha helix has also been suggested [46]. In addition, in the LAC9 regulatory protein of *Kluyveromyces lactis*, a structure containing a "zinc finger" with an adjacent alpha helix has been proposed [48]. These latter findings are consistent with our results, especially with the predicted beta turn 1–omega loop 1–beta turn 2–helix B structures shown in Fig. 2 and summarized in Table 1. Recently, investigators have suggested that target gene HRE binding specificity exists in a "zinc finger" structure within the amino end of the steroid receptor DNA binding domain (cI) [49]. A "zinc finger" exchange between GR and ER indicates that DNA binding specificity does reside in the cI domain [50]. However, detailed examination of point mutation findings indicate that specificity does not reside in the putative "zinc finger" region of the cI domain [49], but resides in adjacent amino acids on its carboxyl flank [39, 41 51, 52]. These point mutations [41] which affect either specific DNA binding, transcription stimulation or both, involve amino acids located within our predicted beta turn 2–helix B structure of RGRDBR/HUMGCRA and HUMESTR beta turn 2–helix B–loop structure. A recent 3-D solution structure of a single "zinc finger" of Xenopus protein Xfin DNA binding domain indicated that the polypeptide backbone folds into a well-defined helix which contains basic and polar amino acids believed to be involved in DNA binding [53].

To investigate the genetic relationship of RGRDBR with known GRE DNA binding sites, we compared the 450 nucleotide cDNA segment corresponding to the 150 amino acids comprising the DNA binding domain of rat GR protein, RGRDBR, with a DNA segment of MMTV5LTR [54], ranging from nucleotides −312 to −38, located upstream of the transcription start site, a region known to contain functional GREs. We observed a maximally similar subsequence located within our predicted alpha helix B of RGRDBR and a nucleotide sequence of MMTV5LTR, encompassing a known GRE, see Fig. 3a. More specifically, the nucleotide sequence −181 to −171 of MMTV5LTR, which is maximally similar to RGRDBR helix B, contains the TGTTCT motif, the most critical regulatory element for MuMTV gene transcription as determined by nuclease footprinting, methylation studies and deletion mutation findings [18, 55–58].

Since there are multiple coding possibilities for the majority of the 20 known amino acids, the exceptions being methionine and tryptophan which have a single codon, to thoroughly investigate the extent of genetic information conserved between RGRDBR and GRE, we converted this MMTV5LTR GRE nucleotide subsequence (−199 to −133) to amino acids in all three reading frames reading rightward and leftward in both strands (Fig. 3b). This procedure revealed all coding possibilities within the nucleotide subsequence, as well as the location of the codon sites. A comparison of these MMTV5LTR amino acid coding possibilities to RGRDBR predicted helix B amino acids revealed highly conserved genetic information localized within the GRE half-sites. Specifically, within the GRE palindrome, codons for amino acids in predicted beta turn 2 (C, G and S)–helix B (K, V, F, R, E, G and Q)–omega loop 3 (H, N, Y, L and C) of RGRDBR are conserved (Fig. 3b). Interestingly, these nucleotide regions correspond to exact GR binding sites as described by Scheidereit *et al.* [18, 58]. Subsequent to our computer-derived predictions above, results from point mutation studies [39, 41, 52] were published which reveal that amino acids, specifically C, E/G, G/S, C, K, V, F, F, K, R, and E of HUMGCRA/RGRDBR and HUMESTR are important to specific DNA binding and transcription activation. Remarkably, these amino acids are located within our predicted beta turn 2–helix B region (Fig. 2, Table 1) and have codons within the GRE region of MMTV5LTR, which contain the maximal subsequence similarity to RGRDBR beta turn 2–helix B shown in Figs 3a, b above. The genetic information conserved within the GRE was predominantly in the right half-site of the palindrome encompassing the hexanucleotide TGTTCT. This finding may offer an explanation for the TGTTCT half-site binding preference of the GR protein in GREs of different genes [16, 18].

Recent findings show that eukaryotic DNA regulatory proteins interact with adjacent DNA major grooves as dimers [16, 17]. Our findings indicate that eukaryotic DNA regulatory proteins have a helix–turn–helix structural motif within their DNA binding domain (Fig. 2), as seen in procaryotic DNA regulatory proteins. These findings suggest a similar structural backbone may be conserved between procaryotes and eukaryotes for presentation of specific alpha helical amino acids to the DNA major groove sites. Therefore, in the absence of atomic coordinates for eukaryotic DNA regulatory proteins, to further study the potential nucleotide interactions of alpha helix B amino acids, we generated a computer model of a eukaryotic/procaryotic hybrid protein.

We used X-ray crystallographic coordinates from CRP, a procaryotic protein which regulates DNA transcription in *Escherichia coli* by specific nucleotide interactions as a dimer with its helix F amino acids [59]. In preparation for creating a eukaryotic/procaryotic hybrid protein, we compared amino acids of RGRDBR alpha helix B with amino acids of CRP alpha helix F for sequence similarity and hydropathy. The results are shown in Fig. 4a. As can be seen, the two helical sequences have six hydropathically similar amino acid residues with two identical residues, glutamic acid (E) and glutamine (Q), in perfect alignment. Helices B and F were constructed as 3-D molecular models using QUANTA (see Experimental Methods). In Fig. 4b–d the structures are compared. The end and lateral views of the 3-D structures are strikingly similar with amino acid side chains of like polarity in nearly identical orientations.

In Fig. 5, RGRDBR helix B is shown attached to the CRP backbone in place of helix F. The hybrid protein is docked onto MMTV5LTR at the site of maximum sequence similarity for helix B, as shown in Figs 3a, b. This docking site encompasses the imperfect palindrome TGTTCT on the sense strand in the right half-site and CAATGT on the antisense strand in the upstream half-site. It can be seen in the CRP/helix B dimer that helix B amino acids are in a position to potentially interact with two adjacent DNA major grooves. When the codon matches and helix B amino acids from Figs 3a, b are highlighted on this model, several of the potential DNA binding side chains of helix B amino acids (hydrophilic amino acids KKRE and hydrophobic amino acids VF) are oriented toward, and appear to be capable of interacting with, nucleotide subsequences identical to their respective codon base pairs. It is noteworthy that these codons are found in the DNAase I protected areas reported by Payvar *et al.* [56, 57] and are virtually identical in location to the nuclease footprint findings reported by Scheidereit *et al.* [58]. In addition, these sequences are palindromic and similar to reported procaryotic DNA operator half-sites [24, 25, 59–61] and include the 5′ TGTTCT 3′ motif conserved in GREs as reported by Scheidereit *et al.* [58].

H-bond formation, van der Waals, polar and hydrophobic interactions have been suggested to be mechanisms of protein/DNA site-specific binding [36]. Both van der Waals contacts and/or hydrophobic interactions with proteins are generally believed to occur within the major grooves of B-DNA at the location of a 5-methyl group of thymine. A TGT motif is found in GRE and progesterone response element half-sites. Likewise, an invariant TGT is found in 12 of the phage 434 operator half-sites [24], and in the phage P22 OR1 half-site [25]. A similar motif, TGTGA, is the consensus sequence for CRP binding and gene transcription [59].

To further characterize the potential nucleotide interaction of helix B amino acids, we highlighted van der Waals fields separately for each potential DNA binding amino acid of RGRDBR helix B and its available codons within the palindromic half-sites. The protein is shown docked at a distance approx. 10 Å from the DNA for visual clarity, see Figs 6a–f.

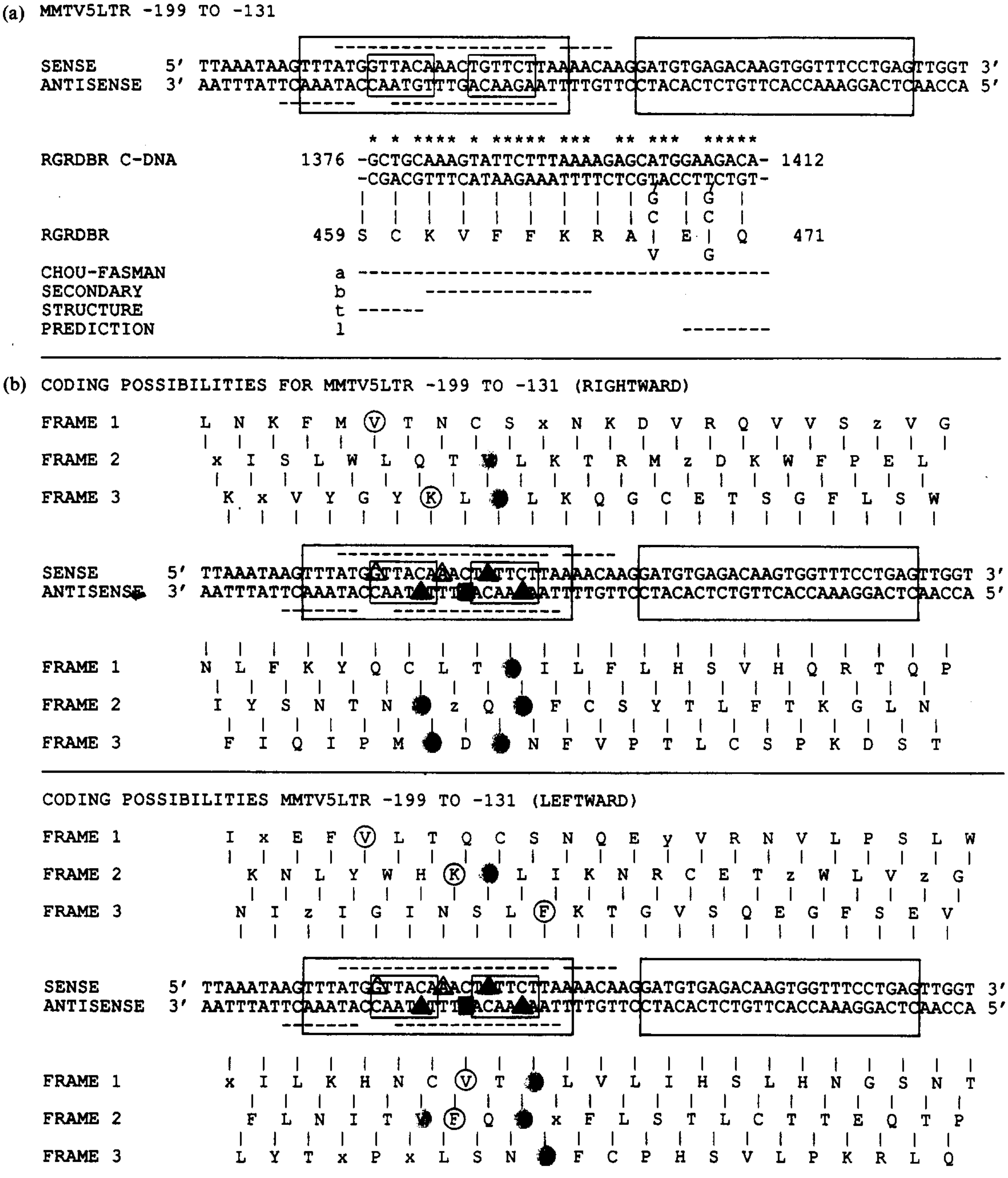
(Figures opposite.)

Fig. 3a. The top of the figure shows the nucleotide sequence of MMTV5LTR ranging from −199 to −131 upstream from the transcription start site. GR receptor binding sites have been detected with nuclease footprinting studies by others and are shown as large boxes [57, 58] or dashed underlines/overlines [59]. Small boxes contain the two GR binding half-sites GTTACA and TGTTCT, respectively. Below MMTV5LTR is the maximally similar subsequence obtained using LOCAL comparing MMTV5LTR nucleotides ranging from −312 to −38 to the 450 nucleotide cDNA sequence which codes for the 150 amino acid sequence of the rat GR DNA binding region (RGRDBR), shown in Figs 1 and 2. Matches are starred. Below the RGRDBR cDNA sequence is shown its corresponding amino acid sequence in the Dayhoff one-letter code as in Fig. 1. Below that, the Chou–Fasman secondary structural prediction for this subsequence of amino acids is shown. Abreviations used for structural predictions are as in Fig. 2.

Fig. 3b. The MMTV5LTR DNA sequence has been converted to amino acid code in all three frames, in both strands and in both directions resulting in all coding possibilities. Amino acids are in the Dayhoff one-letter code as in Fig. 1. Stop codons which do not code for amino acids, i.e. TAA, TAG and TGA, are represented as lower case x, y and z, respectively. Boxes and underlines/overlines are as in Fig. 3a. Methylation inhibition studies [16] are summarized with the following symbols: ▲ show nucleotides where methylation inhibits receptor binding; △ show residues which, when methylated, do not inhibit receptor binding but cannot be methylated after the receptor is bound; ■ shows a residue that is hypermethylated in the presence of bound receptor; amino acids in RGRDBR helix B with codon matches in MMTV5LTR GRE half-sites, specifically K461, V462, F463, K465, R466 and E469, are circled (residue numbers correspond to those from Fig. 2); ● are amino acids which appear to be sterically available to interact with their cognate codons as suggested by our model building and are confined to the two half-sites described above; and ○ are matches with unfavorable steric availability.

Figure 6a is a composite of all potential DNA binding amino acids of helix B in both monomers along with their codons in the half-sites of the two major grooves. Methylation sites [18] within the codon regions described in Figs 6b, c, f are shown in Fig. 3b.

In Fig. 6b, the lysines of RGRDBR helix B are highlighted in both monomers of the CRP/helix B dimer. In addition, lysine codons within the right half-site of the palindrome, AAGAA on the antisense strand, are highlighted. The lysine residues within the right helix (K461 and K465) appear to be capable of interacting with its AAG codons reading in both directions from adenine to guanine. Interestingly, methylation of guanine within AAGAA blocks GR binding [18]. Within the side chain of lysine, ϵ-nitrogen is a H-bond donor and could form a H-bond to the thymine O4-atom of the A/T base pair within its codon, which is a H-bond acceptor. In addition, the λ


(a) MMTV5LTR -199 TO -131
SENSE 5' TTAAATAAGTTTATGGTTACAAACTGTTCTTAAAACAAGGATGTGAGACAAGTGGTTTCCTGAGTTGGT 3'
ANTISENSE 3' AATTTATTCAAATACCAATGTTTGACAAGAATTTTGTTCCTACACTCTGTTCACCAAAGGACTCAACCA 5'
* * **** * ***** *** ** *** *****
RGRDBR C-DNA 1376 -GCTGCAAAGTATTCTTTAAAAGAGCATGGAAGACA- 1412
-CGACGTTTCATAAGAAATTTTCTCGTACCTTCTGT-
G G
C C
RGRDBR 459 S C K V F F K R A E Q 471
V G
CHOU-FASMAN a
SECONDARY b
STRUCTURE t
PREDICTION l
(b) CODING POSSIBILITIES FOR MMTV5LTR -199 TO -131 (RIGHTWARD)
FRAME 1 L N K F M V T N C S x N K D V R Q V V S z V G
FRAME 2 x I S L W L Q T L K T R M z D K W F P E L
FRAME 3 K x V Y G Y K L L K Q G C E T S G F L S W
SENSE 5' TTAAATAAGTTTATGGTTACAAACTGTTCTTAAAACAAGGATGTGAGACAAGTGGTTTCCTGAGTTGGT 3'
ANTISENSE 3' AATTTATTCAAATACCAATGTTTGACAAGAATTTTGTTCCTACACTCTGTTCACCAAAGGACTCAACCA 5'
FRAME 1 N L F K Y Q C L T I L F L H S V H Q R T Q P
FRAME 2 I Y S N T N z Q F C S Y T L F T K G L N
FRAME 3 F I Q I P M D N F V P T L C S P K D S T
CODING POSSIBILITIES MMTV5LTR -199 TO -131 (LEFTWARD)
FRAME 1 I x E F V L T Q C S N Q E y V R N V L P S L W
FRAME 2 K N L Y W H K L I K N R C E T z W L V z G
FRAME 3 N I z I G I N S L F K T G V S Q E G F S E V
SENSE 5' TTAAATAAGTTTATGGTTACAAACTGTTCTTAAAACAAGGATGTGAGACAAGTGGTTTCCTGAGTTGGT 3'
ANTISENSE 3' AATTTATTCAAATACCAATGTTTGACAAGAATTTTGTTCCTACACTCTGTTCACCAAAGGACTCAACCA 5'
FRAME 1 x I L K H N C V T L V L I H S L H N G S N T
FRAME 2 F L N I T F Q x F L S T L C T T E Q T P
FRAME 3 L Y T x P x L S N F C P H S V L P K R L Q

Figs 3a,b *(legends opposite).*

and ϵ lysine-methylene groups could make van der Waals contacts with the thymine 5-methyl groups. Although not highlighted, the middle adenine within the AAA between the two half-sites on the sense strand has also been shown to be methylated without interference of receptor binding [18]. The lysines appear to be sterically unable to interact with the AAA codon. Nonetheless, binding of GR prevents methylation of AAA [18]. Point mutation studies converting lysines to glycine [41] within HUMGCRA in the region of our predicted helix B reveal that a lysine corresponding to RGRDBR K461 is essential for stimulation of transcription and that the lysine corresponding to RGRDBR K465 is important in DNA binding specificity and transcription activity.

Figure 6c shows the valines of helix B highlighted in both monomers. The codons within both half-sites of the palindrome are highlighted, CTGTT on the sense strand in the left site and ATGTT of the antisense strand in the right site. Methylation of the guanines within CTGTT and ATGTT has also been shown to block subsequent binding of GR [18]. The valines appear to be capable of interacting with their codons in the left and right half-sites reading from the guanine outward in both directions. The thymine 5-methyl group of TGT could form van der Waals contacts or hydrophobic interactions with the methyl groups of the valine residues. Although not highlighted, within the minor groove on the antisense strand, methylation of guanine within TTGA is enhanced after receptor binding [18]. Valines appear to be sterically unable to interact at the TTGA site. If bending of the DNA occurs with GR/DNA binding, as proposed for procaryotic protein/DNA complexes [33, 34, 36, 62], then the guanine of TGA within the minor groove may be more accessible for methylation after receptor binding. Recent observations [52] from selective mutation studies reveal that the valine of our HUMGCRA predicted helix B corresponding to RGRDBR helix B V462 is involved in target gene specificity.

The phenylalanines of helix B are highlighted in Fig. 6d, along with their codons in the right half-site, TTCTT on the sense strand, and in the left half-site, TTT on the antisense strand. The phenylalanines appear to be capable of interacting with both codon sites. Interaction at the TTCTT site is suggested by the TTC codons reading inward toward cytosine from both directions. Phenylalanines could form van der Waals interactions with thymine 5-methyl groups within their codon regions. A hydrophobic pocket between the TGT half-sites and within the minor groove could be formed by phenylalanine and valine residues of helix B when the hybrid dimer is docked as shown (see Fig. 6a). Point mutation studies exchanging these phenylalanine residues for glycine residues completely eliminate DNA binding and transcription activation by HUMGCRA [41].

In Fig. 6e, arginines (R466) in both monomers are highlighted along with the AGA codon on the antisense strand in the right half-site. Arginine does not appear to be sterically capable of interacting with AGA nucleotides. However, mutation of this arginine residue in HUMGCRA to glycine reduces DNA binding and transcription activation [41], suggesting that the RGRDBR R466 arginine side chain may be flexible enough to interact with thymine, as described for RGRDBR K465 above (Fig. 6b).

Figure 6f shows glutamic acids highlighted along with their codons AAGAA on the antisense strand in the right half-site. Methylation of guanine within this sequence has been shown to inhibit GR binding [18]. Likewise, the glutamic acid side chain appears to be capable of interacting with the GAA codons reading outward from guanine in both directions. In wild-type CRP, glutamic acid 181 of helix F is proposed to interact with a G/C base pair at positions 7 and 16 of the lac operator [53]. We observed that these positions contain the codon site GAG for glutamic acid on

(Figure opposite.)

Fig. 4. (a) Hydropathic profile comparison of RGRDBR predicted helix B (KVFFKRAVEGQ) to the putative DNA binding alpha helix of CRP from *E. coli* helix F (VGRILKMLEDQ) using the HYPHO program (hydropathic values and amino acid one-letter code as in Fig. 2). (b) Computer graphic image of an end view of the RGRDBR helix B (top) and the CRP helix F (bottom) oriented so their side chain positions can be compared. The image was created using the QUANTA program. Color coding of amino acids is based on polarity: blue = positively charged side chains; red = negatively charged side chains; yellow = uncharged polar side chains; and light blue = non-polar side chains. (c) A 90° rotation on the vertical axis of Fig. 4b showing a lateral view of both helices. Image created using QUANTA with color coding as in Fig. 4b. RGRDBR helix B is at the top. (d) Space-filled model of Fig. 4c. Image created using QUANTA with color coding as in Fig. 4b.

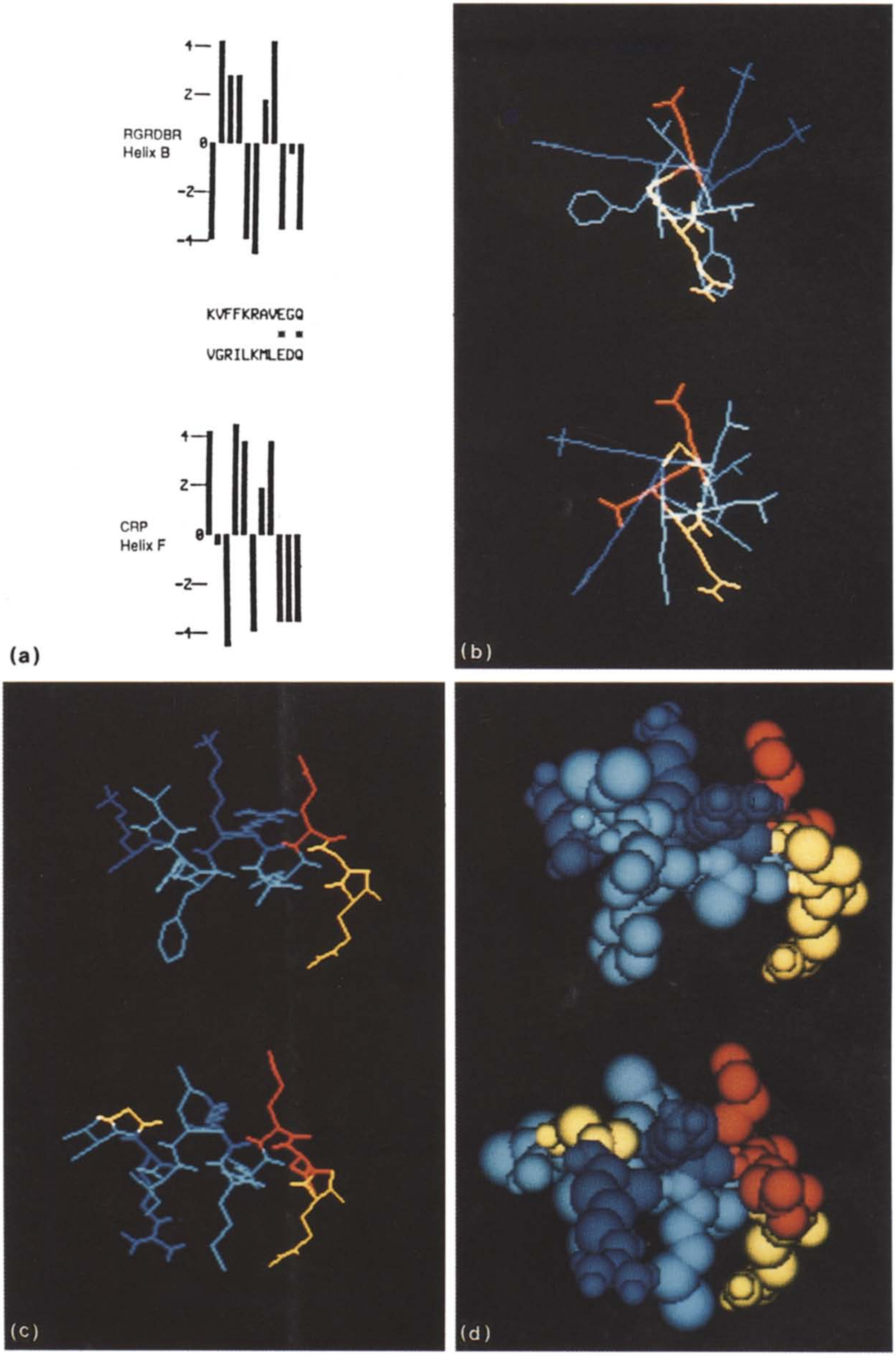


Fig. 4*(legend opposite.)*

```
SENSE     -199 5' T T A A A T A A G|T T T A T G G T T A C A A A|C T G T T C T T|A A|A A C A A G G  3' -160
ANTISENSE      3' A A A T T A T T C|A A A T A C C A|A T G T T T|G A C|A A G A A|T T|T T G T T C C  5'
```

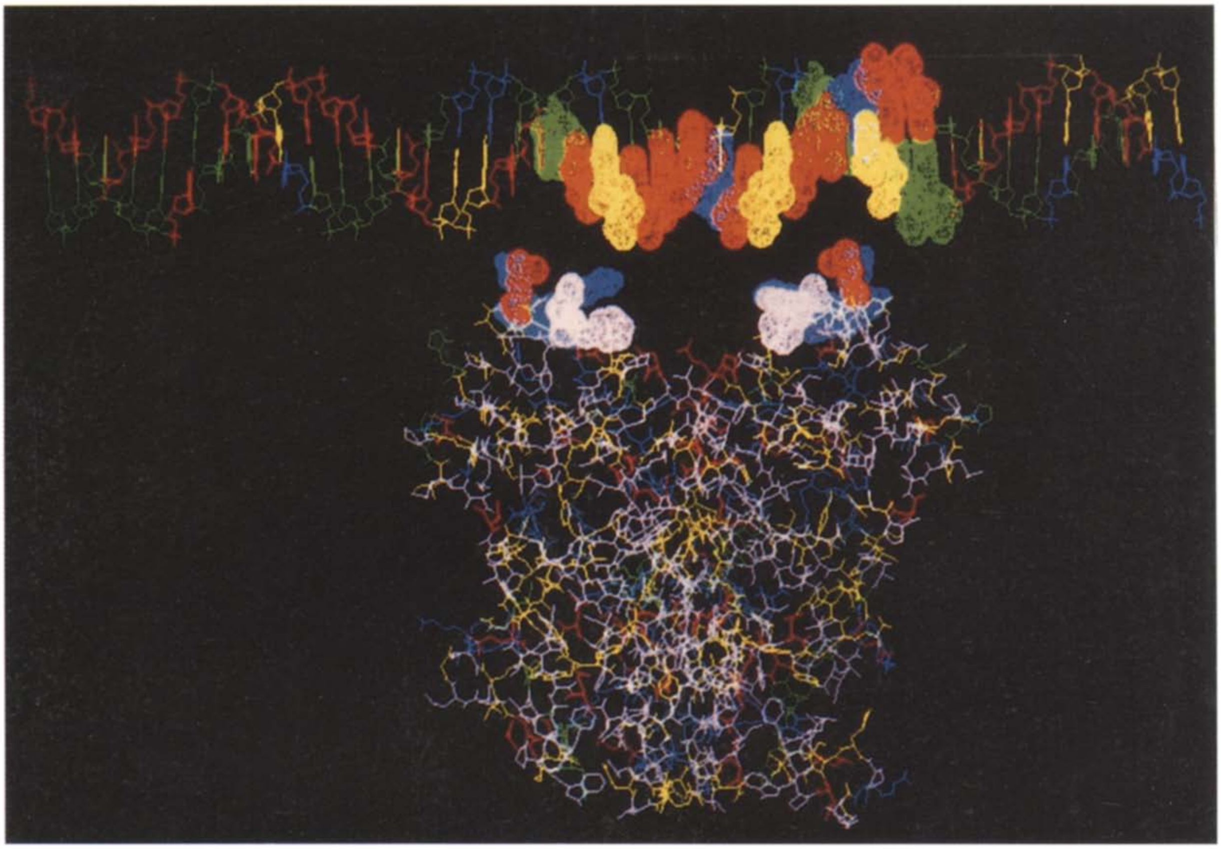
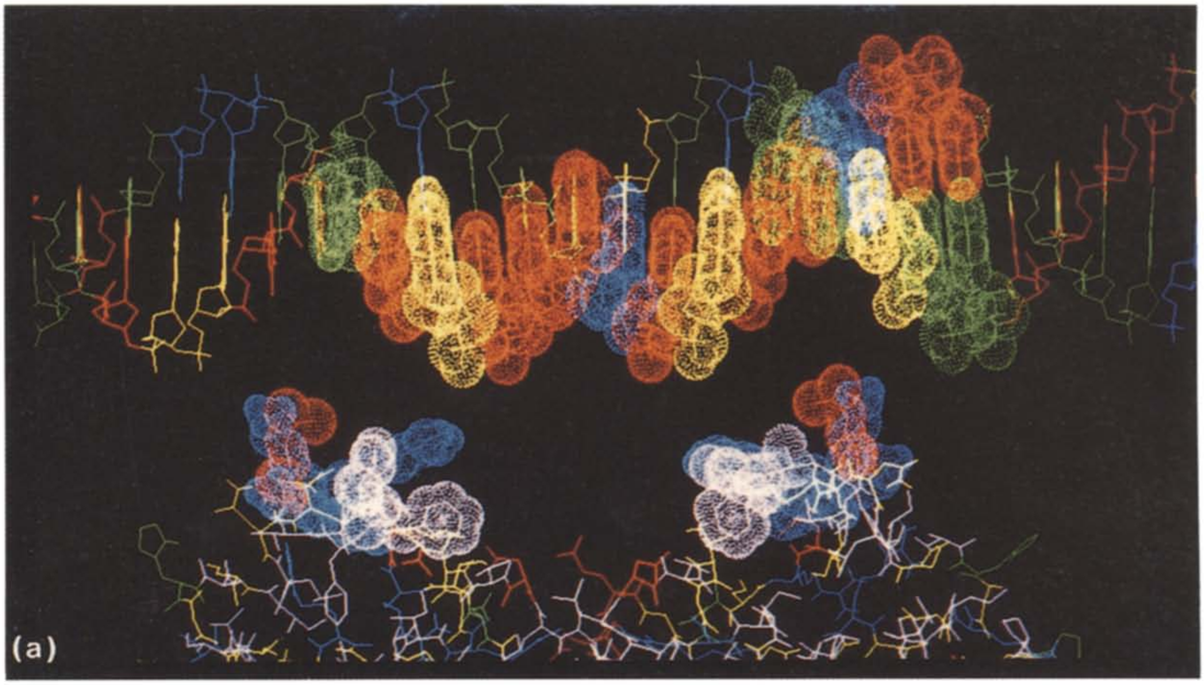


Fig. 5. A computer model of a hybrid protein consisting of CRP from *E. coli* with its DNA binding helix F replaced by the predicted helix B from RGRDBR was constructed. This hybrid protein model was docked onto the region of maximal similarity between RGRDBR cDNA and MMTV5LTR (as shown in Fig. 3a). This nucleotide sequence also coincides with a reported GRE [16] within MMTV5LTR. At the top of the figure, the nucleotide sequence ranging from −199 to −160 for both strands is shown. The large box and underlines/overlines represent protected areas, as described in Figs 3a, b. Codon and amino acid matches from Figs 3a, b are highlighted on the DNA and protein, respectively, and above the model the nucleotide sequence is shown with highlighted nucleotides boxed. All highlighted residues have a dot surface indicating the van der Waals fields of each atom in that residue. Color coding for the protein is as in Fig. 4b, except for the non-polar residues which are colored purple in this and subsequent figures for contrast. Color coding for the DNA is: green = ADE; red = THY; yellow = GUA; and blue = CYT.



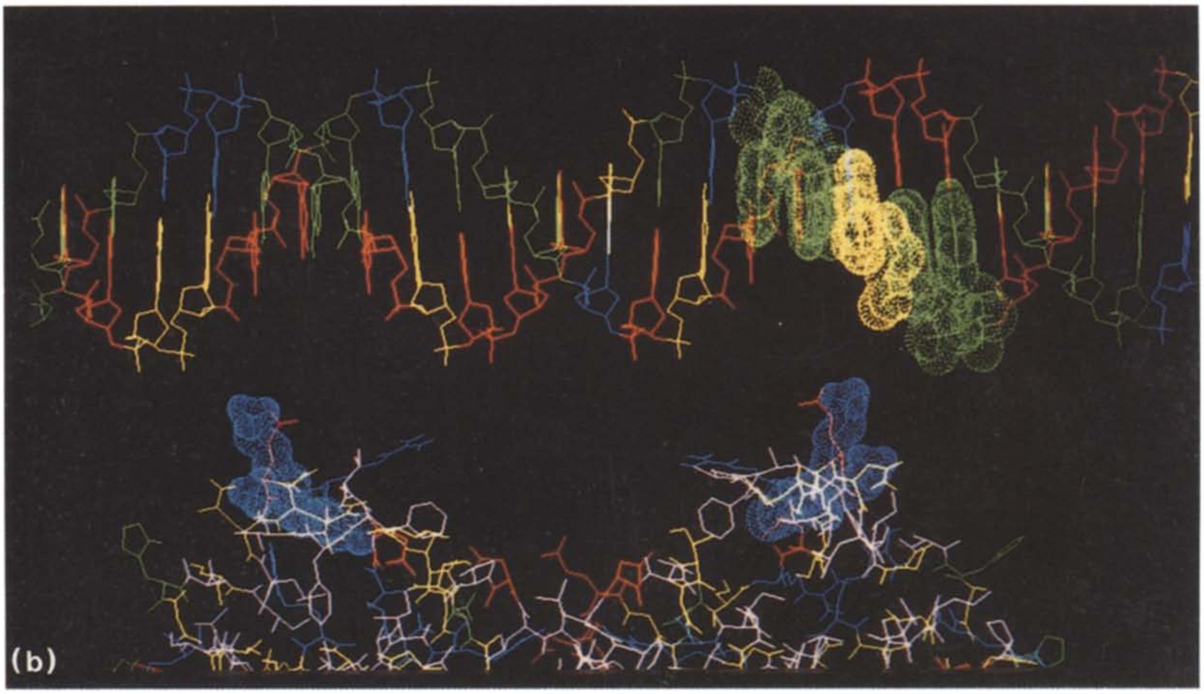
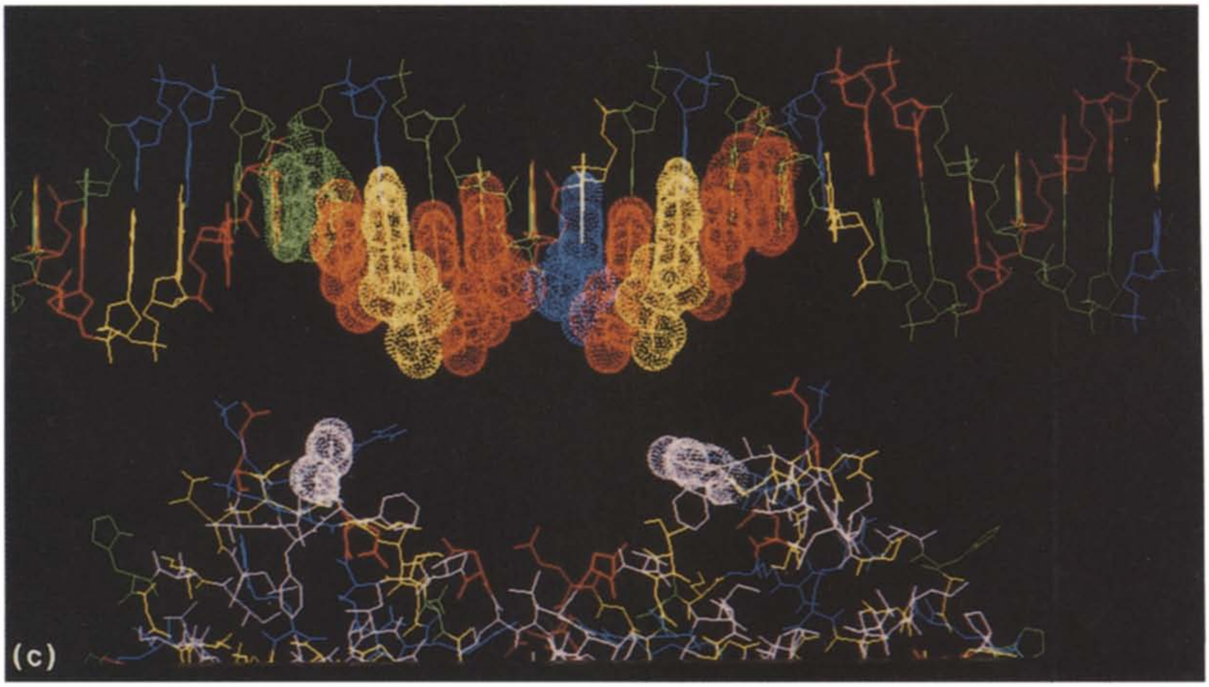


Fig. 6a. A composite of Figs 6b–6f.

Fig. 6b. Lysine residues (K461, K465) and their available codons are highlighted. (Residues corresponding to K465 in both monomers are in a vertical orientation.)



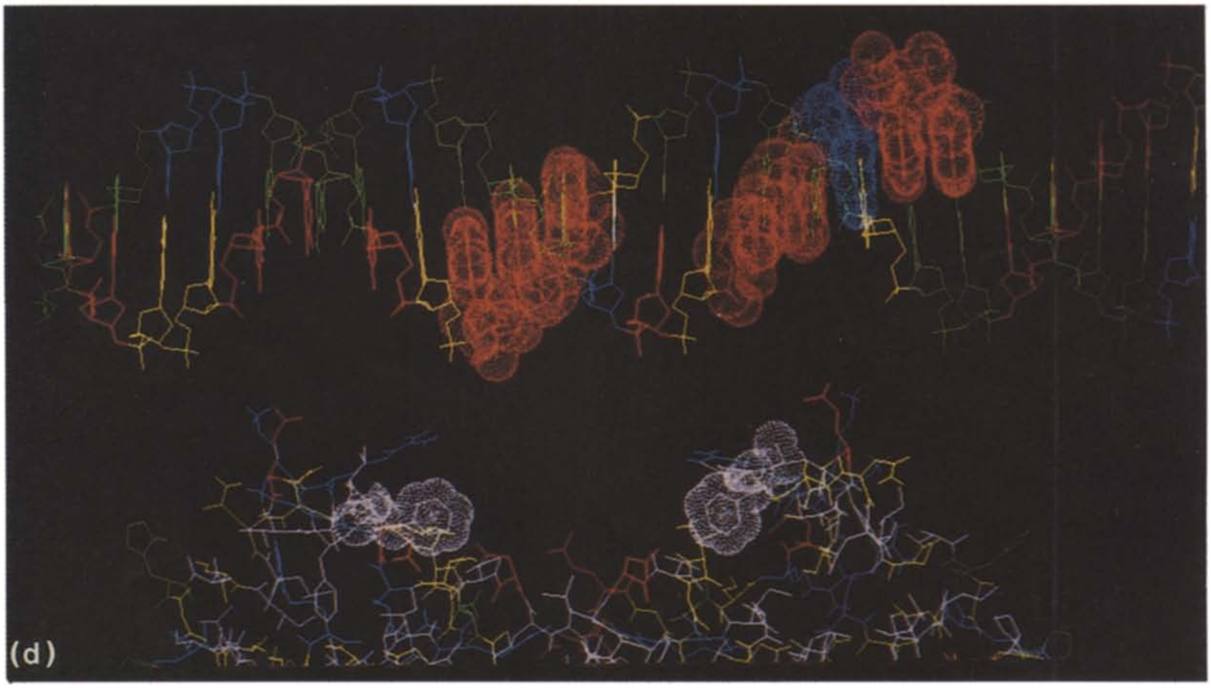
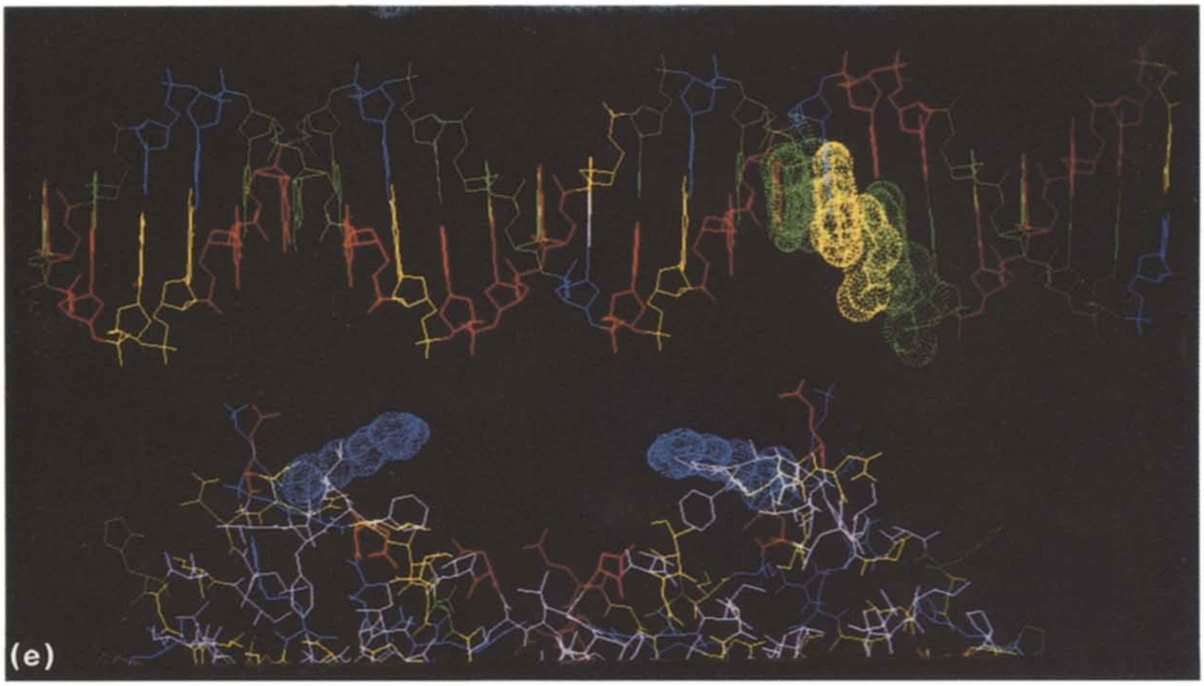


Fig. 6c. Valine residues (V462) and their available codons are highlighted.

Fig. 6d. Phenylalanine residues (F463) and their available codons are highlighted.



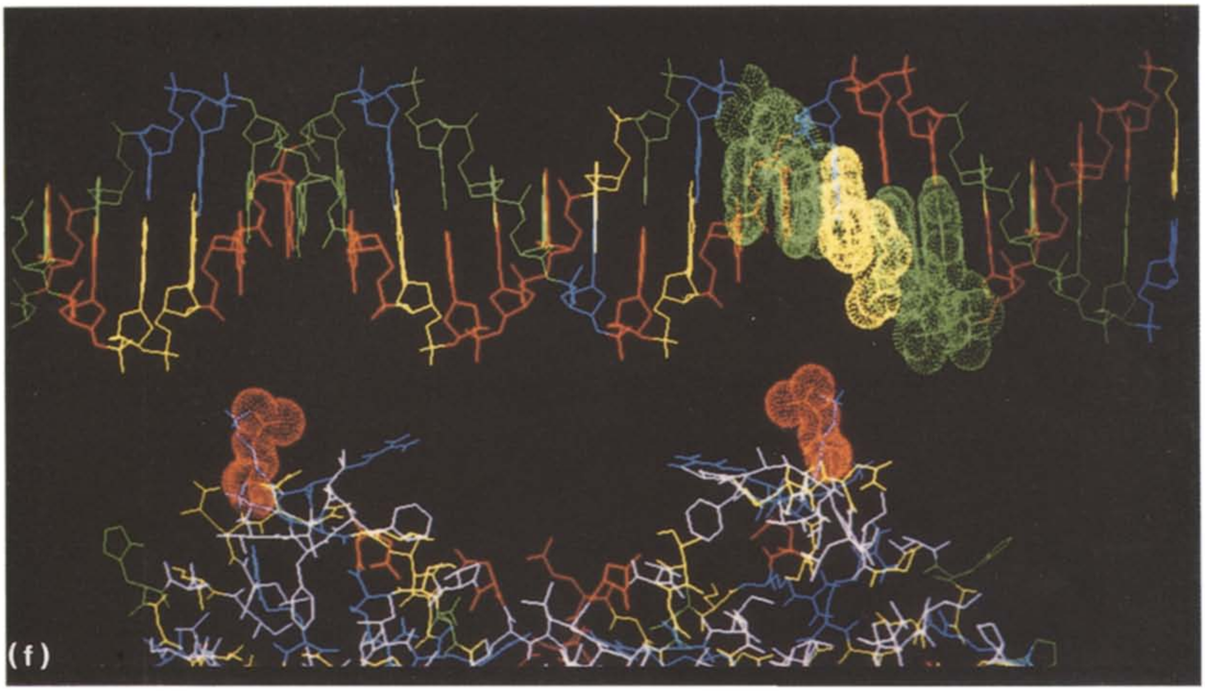


Fig. 6e. Arginine residues (R466) and their available codon are highlighted.

Fig. 6f. Glutamic acid residues (E469) and their available codons are highlighted.

Figs 6a–f. Close-up views of the structure in Fig. 5 with the nucleotide sequence of MMTV5LTR ranging from −188 to −164. RGRDBR helix B (see Fig. 4a) residues that appear to be able to interact with the nucleotides within major grooves of the DNA molecule are italicized as follows: *KVF* F *KR* AV *E* GQ. In all figures, the potential DNA binding amino acids of helix B and their sterically available cognate codons have been highlighted as in Fig. 5.

the sense and antisense strands in the left and right half-sites, respectively, of the lac operator [63]. Methylation of guanine at these positions is inhibited by bound CRP [64]. The glutamic acid carboxylate side chain is a H-bond acceptor and could form a H-bond in the major groove with one glutamate O-atom and cytosine at N4 [65]. In our CRP helix B hybrid model, a similar reaction could occur with helix B glutamic acid E469 forming a H-bond with cytosine N4 within the G/C base pair of its GAA codon region. Point mutations exchanging glycine for glutamic acid of HUMBCRA helix B corresponding to RGRDBR E469 reduces transcription stimulation and DNA binding [41].

Although not shown, we docked the CRP helix B hybrid around the TGTTCT half-site and the adjacent downstream half-site TTGTTC (see Figs 3a, b). This procedure resulted in a reversal of the helix B orientation within the TGTTCT half-site. We observed codon conservation in both

Table 2a. Our H-bond calculations summary for the CRP helix B hybrid/MMTV5LTR GRE complex left half-site

5′ 1 ATGGTTACAAACTGTTCTTAAAAC 24 3′
3′ 48 TACCAATGTTTGACAAGAATTTTG 25 5′

DNA Binding domain residue	No. of H-bonds to DNA	H-bond with DNA base pair	Codon or anti-codon interaction	Major or Minor groove interaction	Watson–Crick d/a interaction	Ribose–backbone interaction
K461	5	*A43, A43, *A43,	— — —	NZ/*A43–N7 Major		NZ/*A43–O5′
		*T42, T42	— —			NZ/A43–O1P
						N/*T42–O5′
						N/T42–O1P
V462	2	*T42, *A43	C, C/AC	N/*A43–H61 Major		N/*T42–O1P
F463	0					
F464	0					
K465	1	T6	—	N/T6–O4 Major	T6–O4	
R466	5	A7, T42, T42	— — —	NE/A7–N1 Minor	A7–N1	NE/T42–N1
		T42, G41	— — .			NE/T42–N3
						NH1/T42–O4′
						NH1/G41–N9
A467	0					
V468	1	T5	C			N/T5–N1
E469	2	A44, *T6	— —	OE2/A44–H62 Major	A44–H62	N/T6–O1P
G470	1	*T5	C			N/T5–O5′
Q471	2	G4, T5	AC, AC			N/G4–O3′
						N/T5–O5′

An asterisk indicates "near" H-bonds as defined in the text. For each amino acid the interacting atom is followed by a slash, the nucleotide type, the nucleotide number and the nucleotide atom interacting. The interacting atom is indicated with IUPAC nomenclature.

Table 2b. Summary of H-bond calculations for the CRP helix B hybrid/MMTV5LTR GRE complex right half-site

5′ 1 ATGGTTACAAACTGTTCTTAAAAC 24 3′
3′ 48 TACCAATGTTTGACAAGAATTTTG 25 5′

DNA Binding domain residue	No. of H-bonds to DNA	H-bond with DNA base pair	Codon or anti-codon interaction	Major or Minor groove interaction	Watson–Crick d/a interaction	Ribose–backbone interaction
K461	7	A33, C17, *T18	C, AC, AC	O/A33–H61 Major		
		T18, *G32, *T15	AC, C, AC	NZ/C17–N3 Mid	C17–N3	
		T16	AC	NZ′/*T18–N3 Minor		
				NZ/T18–O4 Major	T18–O4	
				NZ/*G32–N1 Mid		
				N/*T15–O4 Major	*T15–O4	
				N/T16–O4 Major	T16–O4	
V462	0					
F463	0					
F464	3	G32, *G32, *A33	AC, AC, AC	N/*G32–N7 Major		N/G32–N9
				N/*A33–N7 Major		
K465	3	T16, T16, A33	AC, AC, C	NZ/T16–N3 Minor		
				NZ/T16–O4 Major	T16–O4	
				NZ/A33–N1 Minor	A33–N1	
R466	5	G14, *T15, T15	— — —	NH1/G14–N1 Mid		
		A34, C35	— —	NH1/*T15–N3 Minor		
				NH2/T15–O4 Major	T15–O4	
				NH1/A34–N1 Minor	A34–N1	
				NH1/C35–N3 Mid	C35–N3	
A467	1	*A33	—			N/*A33–O1P
V468	1	A33	AC			N/A33–O1P
E469	0					
G470	0					
Q471	0					

An asterisk indicates "near" H-bonds as defined in the text. For each amino acid the interacting atom is followed by a slash, the nucleotide type, the nucleotide number and the nucleotide atom interacting. The interacting atom is indicated with IUPAC nomenclature.

half-sites. In the TGTTCT half-site, helix B amino acids in this orientation had poorer access to their codons than in the reverse orientation shown in Fig. 5. Specifically, the potential interaction of lysine at its A/T codon site, Fig. 6b, and glutamic acid at its G/C codon site is lost; conversely, the arginine of helix B becomes available to react at its AGA codon site, Fig. 6e. It is possible that glutamic acid may interact in the G/C base pair region in the AACAA codon region of glutamine (Q), see Fig. 3b. However, methylation and mutation studies [18, 55] support our docking model as shown in Figs 5 and 6a–f. Furthermore, the results highlighted in Figs 6a–f support codon conservation as a basis for site-specific recognition and also offer an explanation for the right half-site TGTTCT binding preference reported for GR and progesterone receptors in the GRE/PRE of tyrosine amino transferase gene [16]. The interactions discussed above for RGRDBR helix B amino acids K461, V462, F463, K465 and E469 are supported by the van der Waals surface complimentarity and amino acid orientation apparent in Figs 6a–f. In addition, point mutation studies conducted in the HUMGCRA DNA binding domain indicate that amino acids within our predicted helix B are involved in specific DNA binding and transcription activation [41]. To further test our predicted amino acid nucleotide interactions, described in Figs 6a–f, we docked the CRP helix B hybrid protein within H-bonding distance with its DNA GRE site. The guanine methylation sites [18] within the GRE palindrome were used to position the CRP helix B hybrid for DNA docking.

Our H-bond calculations of the CRP helix B hybrid nucleotide interaction at the GRE site (Tables 1a, b) support the interactions discussed in Figs 6b–f. Specific H-bond interactions between amino acids of helix B and codon/anti-codon nucleotides of MMTV5LTR GRE are shown in Fig. 7. The donor/acceptor (d/a) pattern shown in Fig. 7 is as described in Fig. 3 of the companion article in this issue [63]. Our model indicates that both strands of the DNA making up the codon sites are required for chemical interaction between amino acid and nucleotides. A d/a pocket is formed on the strand opposite the codon sites as we observed in procaryotic models [63].

Interestingly, a TGT dyad symmetry is conserved among both procaryotic operator sites and eukaryotic HREs. The T/A-rich regions create an acceptor-rich pocket within the half-sites of the HREs as we reported for the procaryotic promoter half-sites [63]. Procaryotic DNA binding proteins with a helix–turn–helix motif interact in a cooperative manner as dimers within major grooves of B-DNA [66]. Recent findings that steroid receptor proteins bind in a cooperative manner as dimers to half-sites within their response elements [16, 17] are similar to the mechanism described for procaryotic DNA regulatory protein interaction at promoter sites [33, 66]. Our findings suggest that the DNA binding domain of eukaryotic DNA regulatory proteins is like the procaryotic helix–turn–helix motif. An alpha helix swap among procaryotic DNA regulatory proteins suggests that DNA site-recognition resides within alpha helices [67]. Therefore, we propose that the amino acids of helix B lie across the major groove of the HRE making base contacts which confer both transcription stimulation and DNA binding specificity; point mutation studies by others indicate that amino acids of our predicted helix B are necessary for both specific DNA binding and transcription activation [41]. Amino acids of our predicted helix D may also make specific and non-specific DNA contacts. In this regard, point mutations exchanging glycine for amino acids R and M in our predicted HUMGCRA helix D indicate these amino acids are involved in receptor/DNA binding and transcription activation [41]. These amino acids correspond to identical amino acids located within our RGRDBR predicted helix D. Recent extended X-ray absorption fine structure spectroscopy (EXFACS) findings suggest that two zinc ions are tetrahedrally coordinated in RGRDBR [51]. Since the EXFACS data is inconclusive as to which 4 of the 5 cysteine residues of each "zinc finger" are actually coordinating the metal ions, other folding schemes are possible [51], including cysteine/histidine metal bridges [47]. These metal bridges may also serve to fold the structure into a helix–turn–helix motif presenting our predicted helix B amino acids to the major groove. During final preparation of this paper, an article appeared in the literature describing the 3-D solution structure of a single "zinc finger" of the Xenopus protein Xfin DNA binding domain. These findings indicated that the polypeptide backbone folds into a well-defined helix which contains basic and polar amino acids believed to be involved in DNA binding [53].

The mechanism of site-specific DNA binding recognition by DNA regulatory proteins leading to enhanced gene transcription is unknown [60, 68, 69]. On the basis of our findings we propose that specific amino acids within alpha helices of the proteins' DNA binding domains may specifically interact with their codon/anti-codon nucleotides within the specific DNA regions (HREs) to which they bind. Comparison of the cDNA of the amino acid sequences comprising the DNA binding domains of regulatory proteins to the nucleotide sequences of their respective binding sites on the DNA revealed that genetic information is conserved between nucleotide

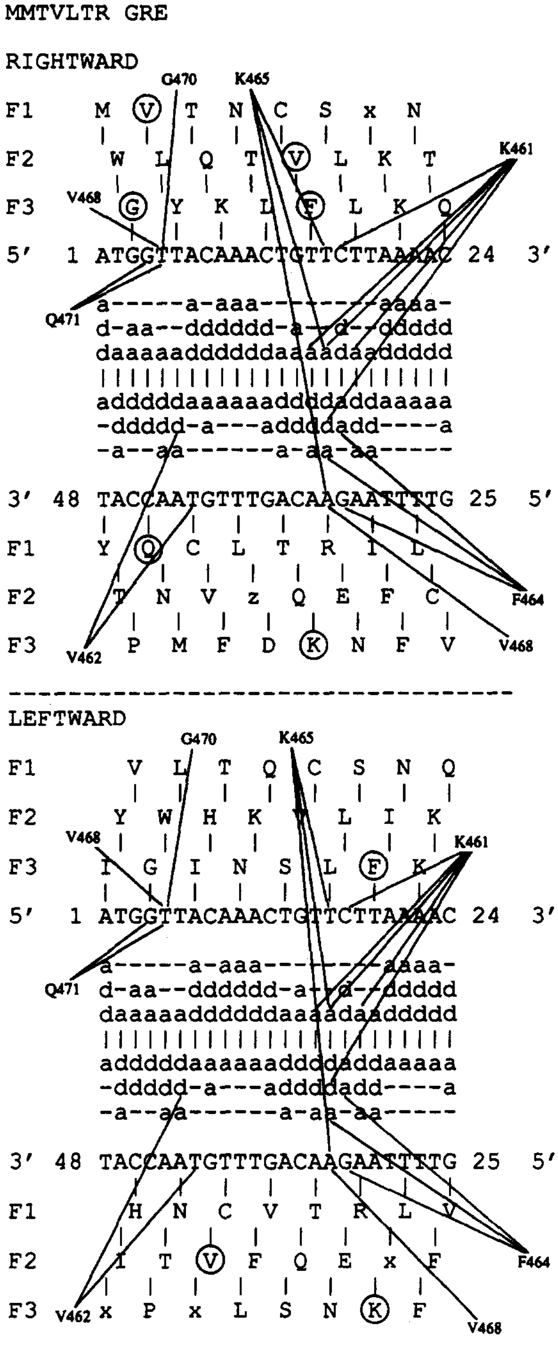


Fig. 7. CRP helix B hybrid protein amino acid/MMTV5LTR GRE codon/anti-codon interactions from our H-bond calculations using molecular modeling. Those amino acids interacting with GRE nucleotides which we identified as codon/anti-codon base pairs are indicated by one-letter nomenclature and residue number with arrows drawn from them to the nucleotide d/a sites with which they are predicted to H-bond. The position of these codons are indicated in the coding possibilities section of the figure by circled amino acids aligned with the first nucleotide of their codon reading rightward and leftward on both the sense and antisense strands. The d/a pattern is as described in Fig 3 of the companion paper in this issue [63].

sequences coding for alpha helices and nucleotides sequences of their specific HRE DNA binding sites. These coding regions are spaced within the major grooves of the GRE half-sites and are compatibly positioned so as to allow direct alignment of amino acid side chains, within helix B, with their cognate codon/anti-codon nucleotides located on the GRE binding half-sites. In addition, a pattern of d/a sites within the GRE is formed by these codon/anti-codon nucleotides. These d/a sites in the GRE are complimentary with d/a sites on side chains of the cognate amino acids in helix B (Fig. 7). For reference see Fig. 2b of the companion paper in this issue [63].

In order to further characterize the amino acid/nucleotide interactions involved in sequence-specific recognition, we are currently working on electrostatic and molecular dynamics simulations of wild-type and mutant protein/operator complexes using 434 cro and cI proteins, *E. coli* CRP protein and the CRP helix B hybrid protein. Atomic coordinates of eukaryotic DNA regulatory proteins are required for visualization of their tertiary structure. However, hybrid protein computer models with procaryotic structural backbones and eukaryotic helices, as described herein, can be created which will allow molecular dynamic studies to proceed. Recently, specific nucleotide/amino acid interactions based on the 434 cI protein/434 operator complex, the TRP repressor/operator complex and the 434 cro/operator complex co-crystal findings were reported [24, 70, 71]. This information agrees with our observations of codon conservation within the respective operator half-sites for amino acids within the DNA binding helices of the respective proteins (see the companion paper in this issue [63]).

Acknowledgements—We thank C. J. R. Ernst and C. E. Crutchfield, III of the Abbott-Northwestern Hospital Cancer Research Laboratory for technical assistance, T. Lybrand of the University of Minnesota for assistance with the AMBER software, T. Smith of the Harvard Medical School Molecular Biology Computer Research Resource (MBCRR) for assistance with sequence analysis software and discussions, I. Weber of the U.S. Department of Commerce National Bureau of Standards for CRP coordinates, M. Fenton of Cray Research for software development and the Minnesota Supercomputer Center user service representatives for technical support on the Cray-2. We also thank Polygen Corp. for software support, especially M. Dygert. Special thanks are due to J. Sood of Silicon Graphics Inc. for hardware support with the Iris 3130, T. Walsh of the Minnesota Supercomputer Institute for encouragement and support and C. Chang of the University of Chicago for providing unpublished nucleotide sequences of human and rat androgen receptors. This work was supported, in part, by a grant of Cray-2 computer time from the Minnesota Supercomputer Institute.

REFERENCES

1. R. Evans, The steroid and thyroid hormone receptor superfamily. *Science* **240,** 889–895 (1988).
2. R. Schleif, DNA binding by proteins. *Science* **241,** 1182–1187 (1988).
3. R. Miesfeld, P. Godowski, B. Maler and K. Yamamoto, Glucocorticoid receptor mutants that define a small region sufficient for enhancer activation. *Science* **236,** 423–427 (1987).
4. S. Hollenberg, C. Weinberger, E. Ong, G. Cerelli, A. Oro, R. Lebo, E. Thompson, M. Rosenfeld and R. Evans, Primary structure and expression of a functional human glucocorticoid receptor cDNA. *Nature* **318,** 635–641 (1985).
5. J. Arriza, C. Weinberger, G. Cerelli, T. Glaser, B. Handelin, D. Housmann and R. Evans, Cloning of human mineralocorticoid receptor complimentary DNA: structural and functional kinship with the glucocorticoid receptor. *Science* **237,** 268–275 (1987).
6. M. Misrahi, M. Atger, L. d'Auriol, H. Loosfelt, C. Meriel, F. Fridlansky, A. Guiochon-Mantel, F. Galibert and E. Milgrom, Complete amino acid sequence of the human progesterone receptor deduced from cloned cDNA. *Biochem. biophys. Res. Commun.* **143,** 740–748 (1987).
7. C. Chang, J. Kokontis and S. Liao, Molecular cloning of human and rat complementary DNA encoding androgen receptors. *Science* **240,** 324–326 (1988).
8. S. Green, P. Walter, V. Kumar, A. Krust, J. Bornert, P. Argos and P. Chambon, Human oestrogen receptor cDNA: sequence, expression and homology to v-*erb* A. *Nature* **320,** 134–139 (1986).
9. V. Giguere, E. Ong, P. Segui and R. Evans, Identification of a receptor for the morphogen retinoic acid. *Nature* **330,** 624–629 (1987).
10. B. DeBuire, C. Henry, M. Benaissa, G. Biserte, J. Claverie, S. Saule, P. Martin and D. Stehelin, Sequencing the *erb*A gene of avian erythroblastosis virus reveals a new type of oncogene. *Science* **224,** 1456–1459 (1984).
11. D. Lubahn, D. Joseph, P. Sullivan, H. Willard, F. French and E. Wilson, Cloning of human androgen receptor complimentary DNA and localization to the X chromosome. *Science* **240,** 327–330 (1988).
12. R. Evans and S. Hollenberg, Zinc fingers: guilt by association. *Cell* **52,** 1–3 (1988).
13. S. Rusconi and K. Yamamoto, Functional dissection of the hormone and DNA binding activities of the glucocorticoid receptor. *EMBO Jl* **6,** 1309–1315 (1987).
14. G. Ringold, Steroid hormone regulation of gene expression. *A. Rev. pharmac. Toxic.* **25,** 529–566 (1985).
15. K. Yamamoto, Steroid receptor regulated transcription of specific genes and gene networks. *A. Rev. Genet.* **19,** 209–252 (1985).
16. S. Tsai, J. Carlstedt-Duke, N. Weigel, K. Dahlman, J-A. Gustafsson, M-J. Tsai, B. O'Malley, Molecular interactions of steroid hormone receptor with its enhancer element: evidence for receptor dimer formation. *Cell* **55,** 361–369 (1988).
17. L. Klein-Hitpass, S. Tsai, G. Greene, J. Clark, M-J. Tsai and B. O'Malley, Specific binding of estrogen receptor to the estrogen response element. *Molec. cell. Biol.* **9,** 43–49 (1989).

18. C. Scheidereit and M. Beato, Contacts between receptor and DNA double helix within a glucocorticoid regulatory element of mouse mammary tumor virus. *Proc. natn. Acad. Sci. U.S.A.* **81**, 3029–3034 (1984).
19. D. von der Ahe, S. Janich, C. Scheidereit, R. Renkawitz, G. Schutz and M. Beato, Glucocorticoid and progesterone receptors bind to the same sites in two hormonally regulated promotors. *Nature* **313**, 706–709 (1985).
20. G. Klock, U. Strahle and G. Schutz, Oestrogen and glucocorticoid responsive elements are closely related but distinct. *Nature* **329**, 734–736 (1987).
21. U. Strahle, G. Klock and G. Schultz, A DNA sequence of 15 base pairs is sufficient to mediate both glucocorticoid and progesterone induction of gene expression. *Proc. natn. Acad. Sci. U.S.A.* **84**, 7871–7875 (1987).
22. A. Cato, R. Miksicek, G. Schutz, J. Arnemann and M. Beato, The hormone regulatory element of mouse mammary tumor virus mediates progesterone induction. *EMBO Jl* **5**, 2237–2240 (1986).
23. A. Cato, D. Henderson and H. Ponta, The hormone response element of the mouse mammary tumor virus DNA mediates the progestin and androgen induction of transcription in the proviral long terminal repeat region. *EMBO Jl* **6**, 363–368 (1987).
24. A. Aggarwal, D. Rodgers, M. Drottar, M. Ptashne and S. Harrison, Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science* **242**, 899–907 (1988).
25. A. Johnson, A. Poteete, G. Lauer, R. Sauer, G. Ackers and M. Ptashne, Lambda repressor and cro-components of an efficient molecular switch. *Nature* **294**, 217–223 (1981).
26. T. Smith and M. Waterman, Identification of common molecular subsequences. *J. molec. Biol.* **147**, 195–197 (1981).
27. J. Kyte and R. Doolittle, A simple method for displaying the hydropathic character of a protein. *J. molec. Biol.* **157**, 105–132 (1982).
28. P. Chou and G. Fasman, Empirical predictions of protein conformation. *A. Rev. Biochem.* **47**, 251–276 (1978).
29. D. McKay, I. Weber and T. Steitz, Structure of catabolite gene activator protein at 2.9 Å resolution. Incorporation of amino acid sequence and interactions with cyclic AMP. *J. biol. Chem.* **257**, 9518–9524 (1982).
30. P. Weiner and P. Kollman, AMBER: assisted model building with energy refinement. A general program for modelling molecules and their interactions. *J. comput. Chem.* **2**, 287 (1981).
31. U. Singh and P. Kollman, An approach to computing electrostatic charges for molecules. *J. comput. Chem.* **5**, 129 (1984).
32. S. Weiner, P. Kollman, D. Case, U. Singh, C. Ghio, G. Alagona, S. Profeta Jr and P. Weiner, A new force field for molecular mechanics simulation of nucleic acids and proteins. *J. Am. chem. Soc.* **106**, 765 (1984).
33. M. Ptashne, *A Genetic Switch. Gene Control and Phage Lambda.* Blackwell/Cell Press, London (1987).
34. T. Steitz, D. Ohlendorf, D. McKay, W. Anderson and B. Matthews, Structural similarity in the DNA-binding domains of catabolite gene activator and *cro* repressor proteins. *Proc. natn. Acad. Sci. U.S.A.* **79**, 3097–3100 (1982).
35. Y. Takeda, D. Ohlendorf, W. Anderson and B. Matthews, DNA-binding proteins. *Science* **221**, 1020–1026 (1983).
36. I. Weber and T. Steitz, Model of a specific complex between catabolite gene activator protein and B-DNA suggested by electrostatic complimentarity. *Proc. natn. Acad. Sci. U.S.A.* **81**, 3973–3977 (1984).
37. L. Presta and G. Rose, Helix signals in proteins. *Science* **240**, 1632–1641 (1988).
38. J. Richardson and D. Richardson, Amino acid preferences for specific locations at the ends of alpha helices. *Science* **240**, 1648–1652 (1988).
39. M. Danielson, L. Hinck and G. M. Ringold, Two amino acids within the knuckle of the first zinc finger specify DNA response element activation by the glucocorticoid receptor. *Cell* **57**, 1131–1138 (1989).
40. T. Patschinsky, T. Hunter, F. Esch, J. Cooper and B. Sefton, Analysis of the sequence of amino acids surrounding sites of tyrosine phosphorylation. *Proc. natn. Acad. Sci. U.S.A.* **79**, 973–977 (1982).
41. S. Hollenberg and R. Evans, Multiple and cooperative *trans*-activation domains of the human glucocorticoid receptor. *Cell* **55**, 899–906 (1988).
42. J. Miller, A. McLachlan and A. Klug, Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *EMBO Jl* **4**, 1609–1614 (1985).
43. L. Fairall, D. Rhodes and A. Klug. Mapping the sites of protection on a 5 S RNA gene by the *Xenopus* transcription factor IIIA. *J. molec. Biol.* **192**, 577–591 (1986).
44. J. Berg, More metal binding fingers. *Nature* **319**, 264–265 (1986).
45. J. Berg, Potential metal binding domains in nucleic acid binding proteins. *Science* **232**, 485–487 (1986).
46. J. Berg, Proposed structure for the zinc-binding domains from transcription factor IIIA and related proteins. *Proc. natn. Acad. Sci. U.S.A.* **85**, 99–102 (1988).
47. M. Danielson, J. Northrop and G. Ringold, The mouse glucocorticoid receptor: mapping of functional domains by cloning, sequencing and expression of wild-type and mutant receptor proteins. *EMBO Jl* **5**, 2513–2522 (1986).
48. M. Witte and R. Dickson, Cysteine residues in the zinc finger and amino acids adjacent to the finger are necessary for DNA binding by the LAC9 regulatory protein of *Kluyveromyces lactis*. *Molec. cell. Biol.* **8**, 3726–3733 (1988).
49. G. Chalepakis, J. Postma and M. Beato, A model for hormone receptor binding to the mouse mammary tumour virus regulatory element based on hydroxyl radical footprinting. *Nucleic Acids Res.* **16**, 10237–10247 (1988).
50. S. Green, V. Kumar, I. Theulaz, W. Wahli and P. Chambon, The N-terminal DNA-binding "zinc finger" of the oestrogen and glucocorticoid receptors determines target gene specificity. *EMBO Jl* **7**, 3037–3044 (1988).
51. L. Freedman, B. Luisi, Z. Korszun, R. Basavappa, P. Sigler and K. Yamamoto, The function and structure of the metal coordination sites within the glucocorticoid receptor DNA binding domain. *Nature* **334**, 543–546 (1988).
52. M. Beato, Gene regulation by steroid hormones. *Cell* **56**, 335–344 (1989).
53. S. M. Lee, G. P. Gippert, K. V. Soman, D. A. Case and P. E. Wright, Three-dimensional solution structure of a single zinc finger DNA-binding domain. *Science* **245**, 635–637 (1989).
54. N. Fasel, K. Pearson, E. Buetti and H. Diggelmann, The region of mouse mammary tumor virus DNA containing the long terminal repeat includes a long coding sequence and signals for hormonally regulated transcription. *EMBO Jl* **1**, 3–7 (1982).
55. E. Buetti and B. Kuhnel, Distinct sequence elements involved in the glucocorticoid regulation of the mouse mammary tumor virus promoter identified by linker scanning mutagenesis. *J. molec. Biol.* **190**, 379–389 (1986).
56. F. Payvar, O. Wrange, J. Carlstedt-Duke, S. Okret, J. Gustafsson and K. Yamamoto, Purified glucocorticoid receptors bind selectively *in vitro* to a cloned DNA fragment whose transcription is regulated by glucocorticoids *in vivo*. *Proc. natn. Acad. Sci. U.S.A.* **78**, 6628–6632 (1981).

57. F. Payvar, D. DeFranco, G. Firestone, B. Edgar, O. Wrange, S. Okret, J. Gustafsson and K. Yamamoto, Sequence-specific binding of glucocorticoid receptor to MTV DNA at sites within and upstream of the transcribed region. *Cell* **35**, 381–392 (1983).
58. C. Scheidereit, S. Geisse, H. Westphal and M. Beato, The glucocorticoid receptor binds to defined nucleotide sequences near the promotor of mouse mammary tumor virus. *Nature* **304**, 749–752 (1983).
59. B. de Crombrugghe, S. Busby and H. Buc, Cyclic AMP receptor protein: role in transcription activation. *Science* **224**, 831–838 (1984).
60. M. Beato, Modulation of gene expression through DNA binding proteins: is there a regulatory code? *Haemat. Blood Transf.* **29**, 217–223 (1985).
61. B. Gicquel-Sanzey and P. Cossart, Homologies between different procaryotic DNA-binding regulatory proteins and between their sites of action. *EMBO Jl* **1**, 591–595 (1982).
62. J. Warwicker, B. Engelman and T. Steitz, Electrostatic calculations and model-building suggest that DNA bound to CAP is sharply bent. *Proteins* **2**, 283–289 (1987).
63. L. F. Harris, M. R. Sullivan and D. F. Hickok, Conservation of genetic information between regulatory protein DNA binding alpha helices and their cognate operator sites: a simple code for site-specific recognition. *Computers Math. Applic.* **20**(4–6), 1–23 (1990).
64. J. Majors, Specific binding of CAP factor to lac promotor DNA. *Nature* **256**, 672–674 (1975).
65. R. Ebright, P. Cossart, B. Gicquel-Sanzey and J. Beckwith, Molecular basis of DNA sequence recognition by the catabolite gene activator protein: detailed inferences from three mutations that alter DNA sequence specificity. *Proc. natn. Acad. Sci. U.S.A.* **81**, 7274–7278 (1984).
66. A. Kolb, A. Spassky, C. Chapon, B. Blazy and H. Buc, On the different binding affinities of CRP at the *lac*, *gal* and *mal* T promoter regions. *Nucleic Acids Res.* **11**, 7833–7852 (1983).
67. R. Wharton, E. Brown and M. Ptashne, Substituting an alpha helix switches the sequence specific DNA interactions of a repressor. *Cell* **38**, 361–369 (1984).
68. C. Pabo and R. Sauer, Protein–DNA recognition. *A. Rev. Biochem.* **53**, 293–321 (1984).
69. B. Matthews, Protein–DNA interaction: no code for recognition. *Nature* **335**, 294–295 (1988).
70. Z. Otwinowski, R. Schevitz, R. Zhang, C. Lawson, A. Joachimiak, R. Marmorstein, B. Luisi and P. Sigler, Crystal structure of *trp* repressor/operator complex at atomic resolution. *Nature* **335**, 321–329 (1988).
71. C. Wolberger, Y. Dong, M. Ptashne and S. Harrison, Structure of a phage 434 cro/DNA complex. *Nature* **335**, 789–795 (1988).
72. D. McDonnell, D. Mangelsdorf, J. Pike, M. Haussler and B. O'Malley, Molecular cloning of complimentary DNA encoding the avian receptor for vitamin D. *Science* **235**, 1214–1217 (1987).
73. M. Dayhoff, *Atlas of Protein Sequence and Structure*. National Biomedical Research Fdn, Silver Spring, Md (1978).
74. S. Needleman and C. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. molec. Biol.* **48**, 443–453 (1970).