Conservation of genetic information: A code for site-specific DNA recognition
(steroid receptors/glucocorticoid response element/genetic code origin)

LESTER F. HARRIS*, MICHAEL R. SULLIVAN, AND DAVID F. HICKOK

Abbot Northwestern Hospital Cancer Research Laboratory, Minneapolis, MN 55407-3799

Communicated by Ralph T. Holman, February 18, 1993 (received for review September 23, 1992)

ABSTRACT We present findings of genetic information conservation between the glucocorticoid response element (GRE) DNA and the cDNA encoding the glucocorticoid receptor (GR) DNA-binding domain (DBD). The regions of nucleotide sub-sequence similarity to the GRE in the GR DBD occur specifically at nucleotide sequences on the ends of exons 3, 4, and 5 at their splice junction sites. These sequences encode the DNA recognition helix on exon 3, a β-strand on exon 4, and a putative α-helix on exon 5, respectively. The nucleotide sequence of exon 5 that encodes the putative α-helix located on the carboxyl terminus of the GR DBD shares sequence similarity with the flanking nucleotide regions of the GRE. We generated a computer model of the GR DBD using atomic coordinates derived from nuclear magnetic resonance spectroscopy to which we attached the exon 5-encoded putative α-helix. We docked this GR DBD structure at the 39-base-pair nucleotide sequence containing the GRE binding site and flanking nucleotides, which contained conserved genetic information. We observed that amino acids of the DNA recognition helix, the β-strand, and the putative α-helix are spatially aligned with trinucleotides identical to their cognate codons within the GRE and its flanking nucleotides.

The glucocorticoid receptor (GR) DNA-binding domain (DBD) has been well characterized and consists of 150 amino acids that have been shown to be biologically active in vitro (1). We reported earlier that a nucleotide sequence within the cDNA encoding the GR DBD shared a high degree of sub-sequence similarity with a well-characterized glucocorticoid response element (GRE) from mouse mammary tumor virus (MMTV) 5' long terminal repeat. This GR DBD cDNA sub-sequence encoded a predicted α-helix structure on the carboxyl flank of the first zinc finger that we proposed as a putative DNA-recognition helix (2). Recently, our prediction of the GR DNA-recognition α-helix amino acid sequence, its location within the GR DBD, and its orientation toward the DNA within the GRE major-groove half-sites was confirmed by NMR (3) and x-ray crystallography (4).

The genomic structure of the human GR gene was recently determined (5); the two zinc fingers of the DBD are separately encoded by 2 of the 10 exons—namely, exons 3 and 4. The DNA-recognition helix encoded in exon 3 is located at the carboxyl terminus of the first zinc finger. Adjacent to the DNA-recognition helix is a structure that has been determined by NMR (3) and x-ray crystallography (4) to be a β-strand. This β-strand encoded in exon 4 is located on the amino terminus of the second zinc finger at the splice junction site of exons 3 and 4. This splice site occurs at a preserved glycine residue within the steroid receptor family that connects the GR DNA-binding recognition helix and β-strand structures as a bridge that joins the two zinc fingers. The carboxyl terminus of the GR DBD contains a structure that we predicted to be an α-helix and proposed to interact at the GRE (2). This putative α-helix is encoded in exon 5 and is located adjacent to the second zinc finger at the exon 4/5 junction splice site.

We report herein that genetic information is conserved between a GRE and its flanking nucleotides within GenBank locus MMTPRGR1 and the cDNA of GR DBD, GenBank locus HUMGCRRA, at nucleotide sequences within exons 3, 4, and 5. These GR DBD cDNA sub-sequences, which are maximally similar to the GRE, encode the GR DNA-recognition helix, a β-strand, and a putative α-helix located in exons 3, 4, and 5, respectively.

MATERIALS AND METHODS

Nucleotide sequence data were taken from cited references and GenBank, a computer data base of DNA and RNA sequences (release no. 68.0; June 1991). LOCAL is a program that searches for maximally similar sub-sequences between any two amino acid or nucleic acid sequences by using a dynamic programming matrix algorithm (6). Gap weighting and mismatch values used were: unity for matches, 0.9 for mismatches and −[0.9 + (1.01 × length)] for gaps. PRSTRC is protein secondary-structure prediction program that uses a modified Chou–Fasman (7) algorithm. LOCAL and PRSTRC are academic software packages distributed by the Harvard Medical School Molecular Biology Computer Research Resource (MBCRR), Dana–Farber Cancer Institute, Harvard School of Public Health, 44 Binney Street JF815, Boston, MA 02115.

All computer models were created by using QUANTA software running on a Silicon Graphics IRIS 4D 320-GTX graphics workstation. QUANTA is a molecular modeling and display tool developed by Molecular Simulations (200 Fifth Avenue, Waltham, MA 02254) that 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 model of the GR DBD was derived from NMR coordinates (3). Atomic coordinates of the exon 5-encoded GR putative DNA-binding α-helix were computed by using the SEQUENCE BUILDER 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 39-bp B-DNA nucleotide sequence from GenBank locus MMTPRGR1 used in Fig. 3.

RESULTS

Earlier we compared nucleotide sequences containing known hormone response elements (HREs) with cDNA sequences

Abbreviations: GRE, glucocorticoid response element; HRE, hormone response element; GR, glucocorticoid receptor; PR, progesterone receptor; ER, estrogen receptor; DBD, DNA-binding domain; MMTV, mouse mammary tumor virus.

*To whom reprint requests should be addressed.
encoding the DBD of steroid receptor proteins (2). Using an algorithm for secondary-structure prediction (7), we located putative DNA-recognition α-helices within the DBDs of members of the steroid receptor protein superfamily. At the time of our initial observations (2), others suggested that the structure for DNA interaction by the steroid hormone receptor proteins was via amino acids found within a “zinc finger” (8). Our predicted structure for the DBD of the steroid hormone receptor protein superfamily suggested that amino acids within a putative α-helix adjacent to the first zinc finger may be important in specific DNA binding. Therefore, we proposed that the zinc finger may serve as a structural backbone for presentation of a putative DNA-recognition α-helix to DNA major-groove half-sites.

Recently the structure for the DBD of the GR was derived from NMR and x-ray crystallography structural determinations (3, 4). In these reports, the location and amino acid sequence of a GR DNA-recognition helix are described. We obtained the NMR-derived atomic coordinates of the GR DBD (R. Kaptein, personal communication) (3). We compared our predicted GR DNA-recognition helix structure and its orientation relative to the GRE with the equivalent NMR-derived GR DNA-recognition helix. Remarkably, the amino acid sequences were 100% identical, and the DNA-directed orientation of the side chains of hydrophilic amino acids Lys-461, Lys-465, Arg-466, Glu-469, Gin-471, and hydrophobic amino acid Val-462 were very similar (2–4) (Fig. 1).

A recent computer modeling and simulation study of the GR, estrogen receptor (ER), and progesterone receptor (PR) DBDs interacting with B-form DNA major-groove half-sites of GRE and the estrogen response element (ERE) via α-helical structures (10) agrees with our original findings (2). Therefore, the NMR determination of the ER structure (11), the NMR and x-ray crystallographic structural determinations of the GR on its corresponding GRE, and computer simulations of ER, GR, and PR interacting with DNA confirm our prediction and support our hypothesis that conservation of genetic information is a site-specific determinant for protein–DNA interaction.

In the present study, using the genomic structure of the GR gene (5) as a guide, we conducted computer-based nucleotide-sequence similarity searches. We compared separately the nucleotide sequence of GR DBD exon 3 (positions 1318–1485), the nucleotide sequence of GR DBD exon 4 (positions 1486–1602), and the nucleotide sequence of GR DBD exon 5 (positions 1603–1626), which encodes a putative DNA-binding helix to MMTPRGR1 nucleotide sequence (−312 to −40). The results are shown in Fig. 2a. The maximally similar sub-sequence for exon 3 and MMTPRGR1 was found within a well-characterized functional GRE of MMTPRGR1 and within exon 3 at a region encoding the DNA-recognition helix of the GR, thus confirming our earlier findings (2). It is significant that this GRE sequence contains the TGTTCCT motif recognition target, and this GRE has been reported by others to be the most critical regulatory element for MMTV gene transcription within the MMTV 5′ long terminal repeat, as determined by nuclease footprinting, methylation studies, and deletion mutation findings (12–15). The maximal nucleotide sub-sequence similarity between MMTPRGR1 and exon 4 occurred at the region encoding the β-strand at the splice junction site that joins exon 4 to exon 3 as described above, but in a different GRE site of MMTPRGR1 than seen with exon 3 (data not shown); however, subsequent comparisons detected sequence similarity for the β-strand-encoded region of exon 4 in the same GRE site as seen with exon 3 (Fig. 2b). The maximally similar sub-sequence between the putative α-helix-encoded region of exon 5 and MMTPRGR1 occurred on the right flanking nucleotide sequence of the GRE (Fig. 2b).

Since there are multiple codons for the majority of amino acids, it is necessary to look at every possible reading frame 5′ to 3′ on both strands of DNA to locate the extent of conserved genetic information. The location of conserved genetic information between the GR and DNA amino acids of the GR DNA-recognition helix encoded by exon 3, the β-strand of exon 4, and the putative α-helix of exon 5 is shown in Fig. 2c. Trinucleotides identical to codons for hydrophilic amino acids lysine, arginine, and glutamic acid as well as hydrophobic amino acids valine and phenylalanine of the DNA-recognition helix are found primarily in the right half-site of the GRE. Genetic information is conserved for β-strand amino acids glutamine, asparagine, tyrosine, leucine, and cysteine within the minor groove between and extending into both major-groove half-sites of the GRE. In addition, trinucleotides identical to codons for amino acids arginine, lysine, and threonine are conserved within the flanking regions of the GRE major-groove half-sites. These amino acids are found within our exon 5-encoded putative DNA-binding α-helix of the GR DBD.

The recent elucidation of the genomic structure of the GR gene (5) and the structural determination of the GR DBD (3) has allowed us to create a model to study separately the nucleotide sequences of the GR gene and exons 3, 4, and 5 as well as the structural modules of the GR DBD that they encode in relation to genetic information conservation at GRE sites. The model of the GR DBD was derived from NMR atomic coordinates of the GR DBD (R. Kaptein, personal communication) (5). However, critical residues following Arg-510 on the carboxyl flank of the GR DBD in the NMR and x-ray structural determinations were disordered, and no coordinates were reported (3, 4). Since the amino acid sequence ranging from Arg-510 to Lys-517 contained our putative α-helix of exon 5, we created an α-helix of the exon 5-encoded amino acids ranging from positions 511 to 517 and attached this structure to Arg-510 in our computer model. We also created a 39-bp B-DNA computer model of the nucleotide sequence of MMTPRGR1 containing the GRE and flanking regions that showed genetic sequence similarity to the GR DBD cDNA exons 3, 4, and 5 (see Fig. 2). A model of the GR DBD docked at this 39-bp B-DNA sequence with areas of conserved genetic information highlighted in the protein and DNA is shown in Fig. 3. Remarkably, the DNA-recognition helix and β-strand structures encoded by exons 3 and 4 are aligned with areas of conserved genetic information within half-site GREs of the GRE (Fig. 3a and b), while the putative α-helix structure encoded by exon 5 aligns with conserved genetic information in the flanking nucleotide regions of the GRE (Fig. 3c). Although not shown separately, we observed that individual codon
sites within the GRE major-groove half-sites and flanking regions are spaced so that they are aligned with their cognate amino acids conserved within the DNA-recognition helix encoded in exon 3 and the β-strand and putative α-helix encoded by exons 4 and 5, respectively, of the GR DBD.

**DISCUSSION**

We have observed that genetic information is conserved within imperfect palindromic nucleotides of the GRE major-groove half-sites and flanking regions for GR DNA-recognition helix amino acids (459–470) encoded by exon 3, for a β-strand structure containing amino acids (471–476) encoded by exon 4, and for a putative α-helix containing amino acids (510–517) encoded by exon 5, respectively. We also observed that flanking the GRE major-groove half-sites are nucleotide sequences with dyad symmetry, 5'-TAAAACGA-3' on the right and 5'-CAAAACT-3' on the left, which appear to extend the original TGTTCCT imperfect palindrome. It is significant that these same flanking nucleotide sequences were determined earlier to be GR binding sites by Scheidereit et al. (13) using nuclease protection patterns (footprinting). The GR amino acids ranging from position 510 to 517 of our putative α-helix encoded by exon 5 share genetic information with the GRE flanking nucleotides. These amino acids have been reported to be partially responsible for nuclear localization of the GR protein and are related in sequence to the nuclear localization signal of simian virus 40 large tumor (T) antigen (16). Therefore, our observations may indicate that in addition to nuclear localization, these GR amino acids, 510–517, may also be important in site-specific DNA recognition and transcription initiation as well.

The origin of GR DNA site-specific recognition has been reported to occur within the GRE major-groove half-site containing the TGTTCCT motif. The GR DBD specifically binds to the TGTTCCT major groove half-site of the GRE as a monomer; this DNA-binding reaction induces cooperative dimerization and a subsequent DNA-binding interaction in the adjacent GRE major-groove half-site (17). Our findings
show that the TGTTCT, or specifically its complementary ACAAGA motif, is rich in codons for the exon 3-encoded GR DNA-recognition helix amino acids (see Fig. 2c). Recently, van der Waals and hydrogen-bonding interactions between GR DBD-recognition helix amino acids and nucleotides in the GRE right major-groove half-site have been reported (4, 10). Many of these observations concur with our earlier predictions (2) and include amino acids interacting at nucleotide sites that we have identified as their cognate codon/anticodon nucleotides. Therefore, conservation of genetic information offers an explanation for the binding preference reported for GR at the GRE major-groove half-site that contains the TGTTCT motif. Furthermore, MMTPRGRI contains numerous GRE sites; however, the GRE site that has been reported to preferentially bind GR and have the highest transcription-enhancing activity (15) shares nucleotide sequence similarity with the DNA nucleotide sub-sequences of exons 3, 4, and 5 of the GR DBD, which encode the DNA-recognition helix, the β-strand, and the putative DNA-binding α-helix, respectively, as described above (Fig. 2a–c). It is interesting to note that the GR has been reported to bind to nucleotide regions within its own gene and down-regulate its expression (18). Our findings show that the nucleotide sequence of the GR gene within exon 3 that encodes the DNA-recognition helix shares a high degree of nucleotide similarity with a well-characterized GRE. Consequently, this GR gene sequence may serve as a regulatory binding site for the GR protein.

The above observations and the findings we reported earlier of conservation of genetic information shared among prokaryotic and eukaryotic DNA regulatory proteins' DBDs and their cognate DNA-binding sites (2, 19) strongly support our hypothesis of a common site-specific DNA recognition code. The basic mechanism of this recognition appears to be stereocchemical complementarity between the proteins' DNA-recognition α-helix amino acids and their cognate codon/anticodon nucleotides within their specific DNA-binding sites. Therefore, our findings support a stereocchemical theory for the origin of the genetic code (20) based on physicochemical complementarity between amino acids and their cognate codon and/or anticodon nucleotides. Three areas of research support this theory and complement our findings: (i) correlations between amino acids' side-chain physicochemical characteristics and the nucleotides of their cognate codons (21–23), (ii) stereocchemical complementarity and structural relationships between amino acids and their cognate codons and/or anticodons nucleotides (24–26), and (iii) direct in vitro binding preference for codon nucleotides by cognate amino acids (27–30).
5-encoded structures of the GR DBD, with trinucleotides identical to their cognate codons within the GRE and its flanks, suggests that these structures may have been template dependent in their evolution (i.e., peptides acting as templates for nucleotide polymerization or vice versa) (32, 33). Therefore, we propose that prebiotic, template-directed autocatalytic synthesis of mutually cognate peptides and polynucleotides resulted in their amplification and evolutionary conservation in contemporary prokaryotic and eukaryotic organisms as a genetic regulatory apparatus.

Our computer analyses were conducted utilizing information derived from published laboratory bench biological research (wet work) of others (1, 3–5, 8, 10–17, 20–33). Our observations are based, therefore, on precedence and offer an explanation for site-specific DNA binding by DNA regulatory proteins. Our findings are consistent with our hypothesis that the origin of the genetic code and a site-specific DNA-recognition code have the same underlying mechanism. Furthermore, our results indicate that our hypothesis, applied to genetic sequence analysis, secondary structural prediction, and molecular model building, can be used as a predictive tool for determining sites on DNA regulatory proteins that recognize cognate DNA-binding sites and vice versa.

We thank Molecular Simulations, Inc., staff for software support with QUANTA; Michael Fenton of Cray Research, Inc., for data reduction programs; Minnesota Supercomputer Institute Scientific Director, Don Truhlar, for support and encouragement; Minnesota Supercomputer Center user services representatives for technical support on the CRAY-2; Temple Smith of the Harvard Medical School Molecular Biology Computing Research Resource for assistance with sequence analysis software; and R. Kaptein for personal communication of GR NMR structural coordinates. Special thanks are due Charlie Larson of Silicon Graphics, Inc., for hardware support with the IRIS 4D 320-GTX workstation. We are sincerely grateful to Prof. Thomas C. Spelsberg of the Department of Biochemistry and Molecular Biology, and Pamela D. Popken-Harris of the Department of Immunology, Mayo Foundation, Rochester, MN, for preliminary review of the manuscript and encouragement. This work was supported in part by a research grant from the Minnesota Supercomputer Institute, Minneapolis.