DIRECTED MOLECULAR EVOLUTION

LESTER F. HARRIS[1], MICHAEL R. SULLIVAN[1] and DOLPH L. HATFIELD[2]
[1] *David F. Hickok Memorial Cancer Research Laboratory, Abbott Northwestern Hospital, 800 E. 28th St., Minneapolis, MN 55407, U.S.A.;* [2] *Section on the Molecular Biology of Selenium, Laboratory of Basic Sciences, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, U.S.A.*
(Correspondence should be sent to: Lester F. Harris or Dolph Hatfield; e-mail: harrislf@sp.msl.umn.edu; hatfield@dc37a.nci.nih.gov)

(Received 10 June 1998)

Abstract. We propose the existence of a relationship of stereochemical complementarity between gene sequences that code for interacting components: nucleic acid-nucleic acid, protein-protein and protein-nucleic acid. Such a relationship would impose evolutionary constraints on the DNA sequences themselves, thus retaining these sequences and governing the direction of the evolutionary process. Therefore, we propose that prebiotic, template-directed autocatalytic synthesis of mutally cognate peptides and polynucleotides resulted in their amplification and evolutionary conservation in contemporary prokaryotic and eukaryotic organisms as a genetic regulatory apparatus. If this proposal is correct, then the relationships between the sequences in DNA coding for these interactions constitute a life code of which the genetic code is only one aspect of the many related interactions encoded in DNA.

1. Directed Molecular Evolution

Genetic information which is stored in DNA may be retained in DNA as regulatory DNA, or may be transferred to RNA and retained in RNA as tRNA, rRNA, snRNA, etc. or may be reversed from mRNA back to DNA by reverse transcriptase or may be further transferred from RNA (through mRNA) to protein (see Figure 1). This flow of genetic information constitutes the central Dogma of Biology. A relationship between sequences in DNA that are coded to interact would clearly facilitate the formation of genetic information, which in turn would facilitate the emergence of multi-component systems and influence evolutionary direction. For example, sequences in DNA that code for nucleic acid interactions are related through stereochemical complementarity of the encoded base sequences (Watson, J.D. and Crick, 1953) (i.e., those DNA sequences that code for codon:anticodon interactions (Crick, 1966; Osawa *et al*., 1991; Osawa, 1995; Hatfield *et al*., 1992) and other nucleic acid interactions (Saenger, 1988), including those sequences that direct structural conformation in DNA and RNA (Saenger, 1988). In the absence of this relationship, the various components would evolve indepently of each other and evolution of the system would thus be left to chance alone. We propose the existence of a similar relationship between gene sequences that code for interacting

Origins of Life and Evolution of the Biosphere **29:** 425–435, 1999.

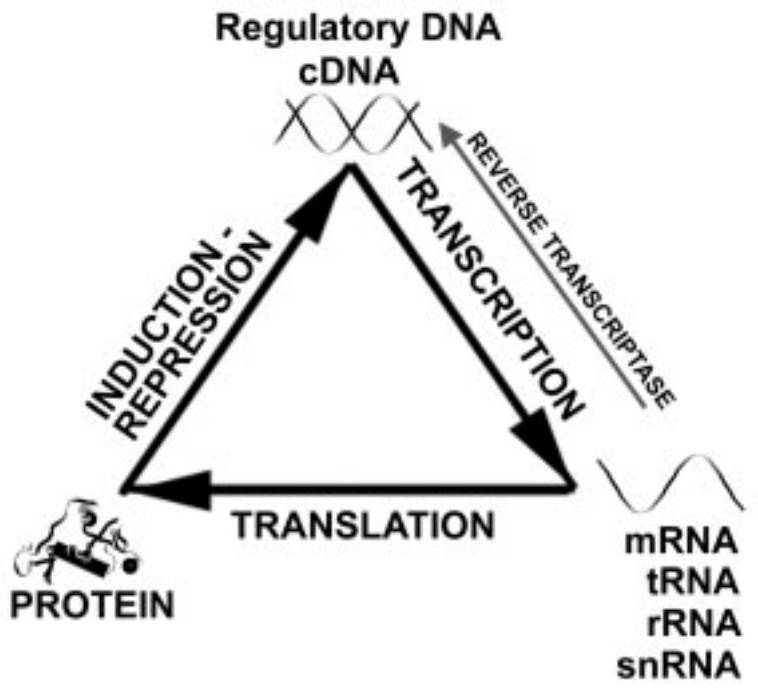

Figure 1. The central dogma of biology: the flow of genetic information.

components (protein:protein and protein:nucleic acid). Such a relationship would impose evolutionary constraints on the DNA sequences themselves, thus retaining these sequences and governing the direction of the evolutionary process.

There is mounting evidence that in the case of protein:nucleic acid interactions the sequences in DNA responsible for directing the specific recognition between these two components, like those involving nucleic acid: nucleic acid interactions, are related. A code for recognition between regulatory proteins and the corresponding specific sequences in DNA to which they bind has been deciphered whereby the mechanism of recognition is determined by the stereochemical complementarity between sites on anino acid sidechains and sites on their cognate codon:anticodon nucleotide bases (Harris *et al*., 1990a, 1993). This notion of a recognition code between proteins and their specific DNA binding sites was initially based on the findings that the c-DNA sequences which encode amino acids in eukaryotic and prokaryotic regulatory proteins' DNA recognition helices have a high degree of nucleotide subsequence similarity with the sequences in DNA to which they specifically bind and regulate transcription, operators or hormone response elements (Harris *et al*., 1990a,b, 1993).

Application of this recognition code led to the discovery of DNA recognition helices for members of the steroid/thyroid hormone receptor superfamily of DNA regulatory proteins (Harris *et al*., 1990, 1993). For example, the c-DNA that en-

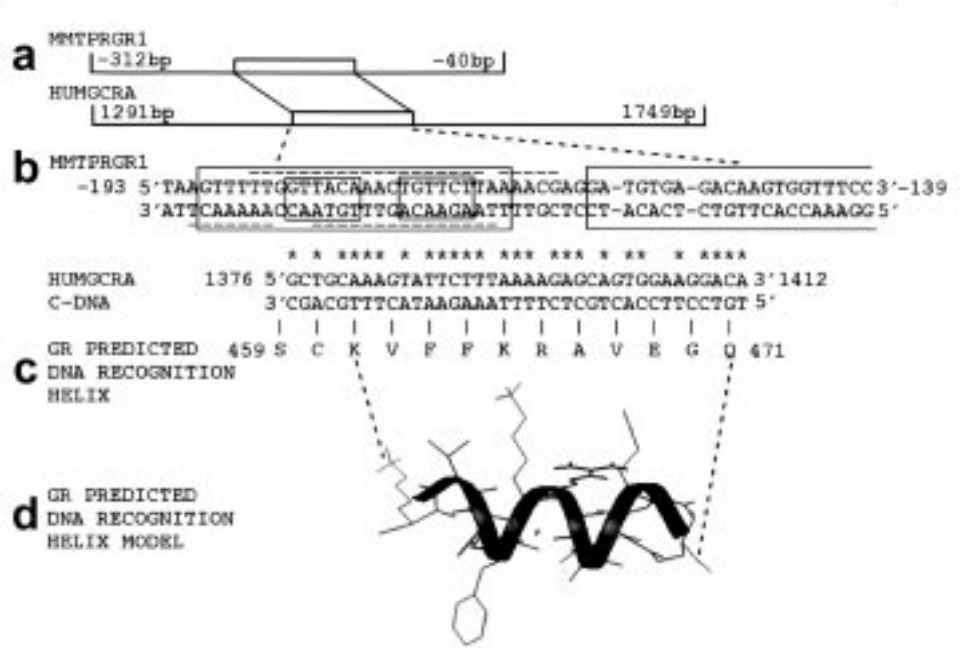

Figure 2. Comparison of the c-DNA sequence encoding the GR DBD with a sequence within the LTR upstream of the MMTV transcription start site. (a), a schematic of a maximally similar nucleotide sequence alignment for the mouse mammary tumor virus (MMTV) 5' long terminal repeat (LTR) (Genbank locus MMTPRGR1) consisting of 272 nucleotides ranging from –312 to –40 upstream from the MMTV transcription start site versus 458 nucleotides of c-DNA encoding the human glucocorticoid receptor DNA binding domain (GR DBD) (GENBANK locus HUMGCRA) ranging from nucleotides 1291 to 1749 are shown. In (b), the MMTPRGR1 nucleotide sequence with maximal nucleotide subsequence alignment with the human c-DNA GR DBD HUMGCRA sequence is shown. Within the MMTPRGR1 sequence are GR binding sites, glucocorticoid response elements (GREs), as detected by nuclease footprinting studies by others and are shown as large boxes (Payvar *et al*., 1983) and dashed underlines and overlines (Scheidereit *et al*., 1983; Scheidereit and Beato, 1984). Within one of the GR DNA binding sites are small boxes containing GRE DNA major groove half-sites, GTTACA and THTTCT, respectively. Nucleotide base pair matches between the MMTPRGR1 and HUMGCRA sequence alignment are shown as stars. Below the HUMGCRA c-DNA sequence is the corresponding amino acid sequence shown in Dayhoff one letter code (Dayhoff, 1978) with the amino acids numbered as in the rat GR (Miesfeld *et al*., 1987). In (c), this amino acid sequence was predicted to be an alpha helix using the Chow-Fasman secondary structure prediction algorithm (Chou and Fasman, 1978). In (d), a computer model of the predicted GR DNA recognition helix is shown.

codes the DNA binding domain (DBD) of the glucocorticoid receptor (GR) was compared with a DNA sequence within the long terminal repeat (LTR) upstream of the mouse mammary tumor virus (MMTV) gene transcription start site, a region known to contain several functional glucocorticoid response elements (GREs). A maximal similarity was detected between a well-characterized GRE within the MMTV LTR sequence and a nucleotide subsequence of the GR DBD c-DNA (see

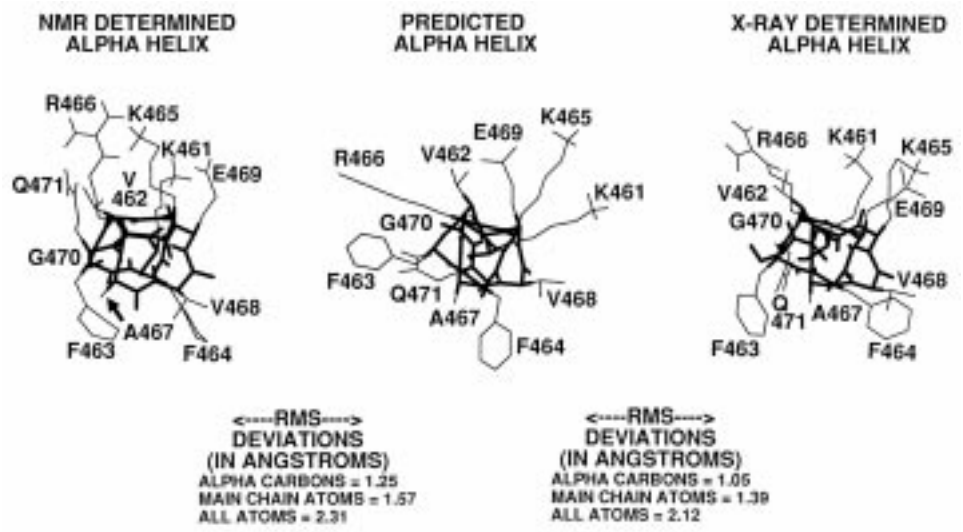

Figure 3. Comparison of the putative GR DNA recognition helix to GR DNA recognition helix structures subsequently determined by NMR and X-ray methodology. A computer model of the predicted GR DNA recognition helix (center) is compared to a computer model of the GR DNA recognition helix from NMR coordinates (Hard *et al*., 1990). (left) and a computer model of the GR DNA recognition helix from x-ray crystallography coordinates (Luisi *et al*., 1991) (right). End views of the alpha helices are shown with amino acids designated using the Dayhoff one letter code (Dayhoff, 1978) and numbered as in the rat GR (Misfeld *et al*., 1987). Main chains are shown with heavy lines, side chains are shown as light lines, The Root Mean Square (RMS) deviations for the predicted and NMR structures are 1.25 Å for the alpha carbon atoms (11 atoms), 1.57 Å for the main chain atoms (44 atoms) and 2.31 Å overall (76 atoms). The RMS deviations for the predicted and x-ray crystallography structures are 1.05 Å for the alpha carbon atoms (11 atoms), 1.39 Å for the main chain atoms (44 atoms) and 2.12 Å overall (76 atoms).

Figures 2a–b). The region of maximal similarity within the GR DBD c-DNA was predicted to encode an alpha helix (see Figures 2c–d). This finding resulted in the discovery of a putative GR DNA recognition helix (see Figure 2d) which was subsequently confirmed by NMR (Hard *et al*., 1990) and X-ray crystallography (Luisi *et al*., 1991) (see Figure 3).

Subsequently, genetic sequence similarity was reported between a GRE DNA sequence and nucleotide sequences at the exon splice junction sites of exons 3, 4 and 5 within the DNA binding domain of the GR protein (see Figures 4a–c) (Harris *et al*., 1993). The GR protein structures encoded at the exon 3, 4 and 5 splice junction sites included the GR DNA recognition helix, a beta strand and a predicted alpha helix, respectively. The primary nucleotide sequences at the exon splice site junctions encoding these GR structures and the nucleotide sequence of the GRE were not identical, but were very similar (see Figure 4b). Since there are multiple codons for the majority of the 20 known amino acids, the GRE sequence was examined in all three reading frames on both strands to determine the extent

of genetic information within the GRE sequence. This procedure revealed trinucleotides identical to codon:anticodon nucleotides in overlapping reading frames for amino acids of the structures encoded at the exon splice junction sites (See Figure 4c).

By model building and conducting molecular dynamics simulations of the GR DBD protein in complex with its cognate GRE DNA, it was observed that amino acids within GR DBD structures encoded at the splice junctions of exons 3, 4 and 5 specifically interact with their cognate codon-anticodon nucleotides within the GRE and its flanking regions (Harris *et al*., 1994, 1995, 1996a,b). The interactions observed were manifest by strong electrostatic and H-bonding between amino acids of the GR DNA recognition structures and nucleotide base sites of their cognate codon-anticodons. As an example, during molecular dynamics arginine 466 of the GR DNA recognition helix shows movement and orientation of its sidechain toward its cognate codon site, AGA, within the GRE sequence (see Figures 5a-d). This movement and orientation of the arginine 466 sidechain is manifest by strong electrostatic energy of attraction and the formation of H-bonds between the arginine sidechain and its codon nucleotides (see Figures 5b and 5d). Strong electrostatic attraction for cognate codon-anticodon nucleotides was also observed for other hydrophilic amino acids of the GR DNA recognition helix (see Figures 6a–f). Similar findings of amino acid- codon nucleotide recognition have been reported in *Tetrahymena group* I self-splicing intronic RNA by arginine (Yarus, 1991). The arginine side-chain shows stereoselective binding for its codons, AGA, CGA and AGG, which are conserved at the catalytic sites of 66 group I self splicing intronic RNA sequences (Yarus, 1989).

Trinicleotides identical to codon-anticodons for amino of regulatory proteins' DNA recognition helices were also observed for other eukaryotic and prokaryotic regulatory proteins within their cognate response elements or operators (Harris *et al*., 1990a,b, 1993, 1994, 1995, 1996a,b). This relationship allows concentration of genetic information within both strands of the specific DNA recognition motifs for both hydrophilic and hydrophobic amino acids of the regulatory proteins' DNA recognition helices. Recently, a gal/S operator that specifically binds the gal/S repressor protein was found within the c-DNA that encodes the gal/S repressor protein; this gal/S operator sequence is identical to the c-DNA sequence which encodes amino acid residues 1–6 within the DNA recognition helix of the gal/S repressor protein (Muller-Hill and Kolkhof, 1994) (see Figure 7). Therefore, genetic subsequence similarity between DNA regulatory proteins[1]; c-DNA encoding DNA recognition helices and their cognate operators or response elements is, in most cases, very similar and in one case identical.

All of these observations support the existence of a code for DNA site specific recognition by DNA regulatory proteins based on stereochemical complementarity and interaction between amino acid sidechains and their cognate codon-anticodon nucleotides. In all of the above studies the amino acids of the DNA recognition helices are consistantly found lined up with their cognate codon nucleotides in their

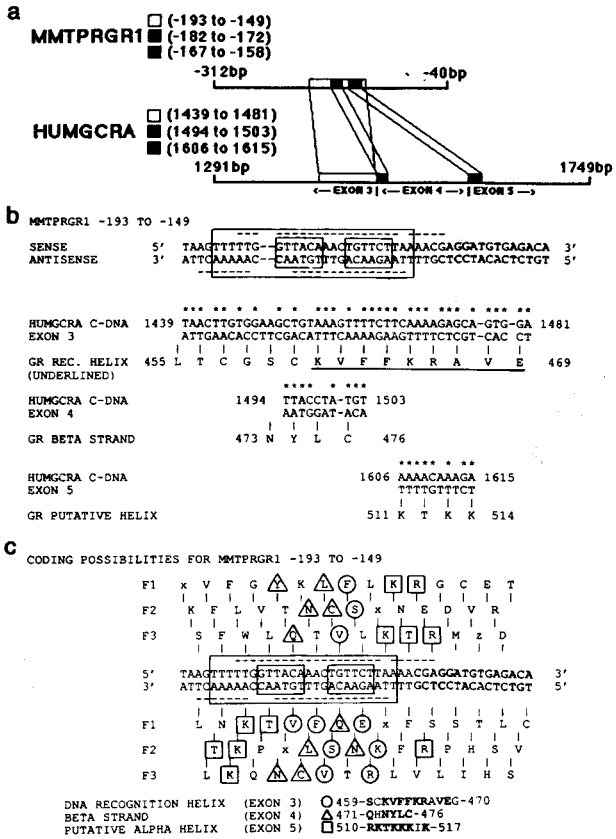

Figure 4. Genetic sequence similarity between a glucocorticoid response element (GRE) and the c-DNA encoding the DNA binding domain (DBD) of the glucocorticoid receptor protein (GR), exons 3, 4 and 5. (a) A schematic of local nucleotide sequence alignments for exon 3 (nucleotide positions 1318 to 1485), exon 4 (positions 1486 to 1602) and exon 5 (positions 1603 to 1626) of the DNA encoding the human GR DBD (GENBANK locus HUMGCRA) vs. mouse mammary tumor virus (MMTV) 5' long terminal repeat (GENBANK locus MMTPRG1) nucleotides ranging from –312 to –40 upstream from the MMTV transcription start site. (b) Nucleotide sequence alignments from (4a). At the top is shown the MMTPRGR1 nucleotide sequence within which GR binding sites (GREs) have been detected with nuclease footprinting studies by others and are shown as large boxes (Payvar *et al*., 1983) and dashed underlines and overlines (Scheidereit *et al*., 1983; Scheidereit and Beato, 1984). Small boxes contain the two GR binding (GRE) half-sites GTTACA and TGTTCT, respectively. Nucleotide base pair matches are starred. Below the HUMGCRA cDNA sequences are shown their corresponding amino acid sequences in Dayhoff (Dayhoff, 1978) one-letter code with the amino acids numbered as in the Rat GR (Miesfeld *et al*., 1987). The DNA recognition helix in the exon 3 alignment is underlined. (c) The nucleotide sequence of the GRE within MMTPRGR1 showing maximum subsequence similarity (see 4b) is translated to amino acids (in Dayhoff one-letter code) in all reading frames (F1, F2, and F3), on both strands: top (rightward: sense 5'–3') and bottom (leftward: antisense 5'–3'). Circles, triangles and squares indicate codons in the DNA sequence with which cognate amino acids from the GR DBD are aligned. Circles = codons for exon 3 encoded DNA recognition helix amino acids, triangles = codons for exon 4 encoded beta strand amino acids and squares = codons for exon 5 encoded putative alpha helix amino acids. The amino acid sequences of these structures are shown at the bottom of the figure. Amino acids aligned with cognate codons are in boldface type.

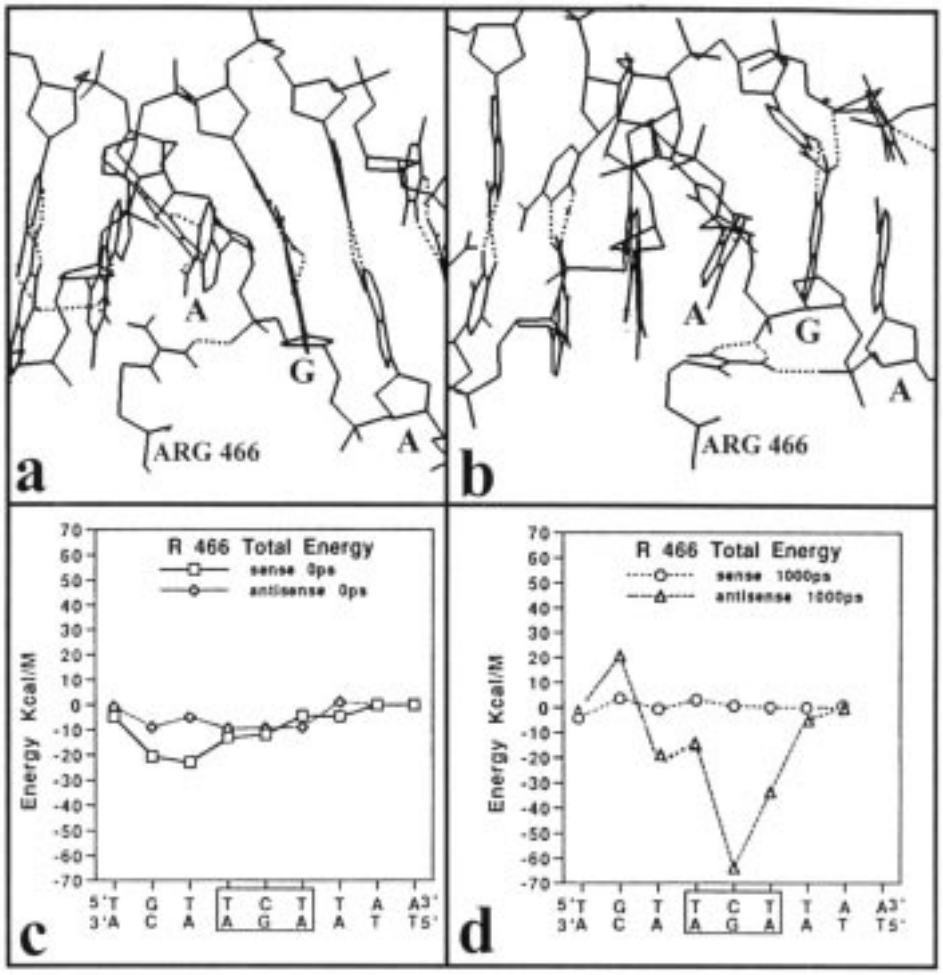

Figure 5. Computer models and energy plots of Arginine 466 of the GR DBD DNA recognition helix interacting with GRE nucleotides. Dashed lines represent H-bonds. (a) Arginine 466 and GRE nucleotides at the beginning of molecular dynamics. (b) Arginine 466 and GRE nucleotides at the end of 1000 picoseconds of molecular dynamics. (c) A plot of interaction energy between Arginine 466 and GRE nucleotides at the beginning of molecular dynamics. (d) A plot of interaction energy between Arginine 466 and GRE nucleotides at the end of 1000 picoseconds of molecular dynamics.

specific DNA binding sites. These observations suggest that these structures may have been template dependant in their evolution (i.e., peptides acting as templates for nucleotide polymerization or vice versa) (Nelsestuen, 1978, 1979). Therefore, we propose that prebiotic, template-directed autocatalytic synthesis of mutally cognate peptides and polynucleotides resulted in their amplification and evolutionary conservation in contemporary prokaryotic and eukaryotic organisms as a genetic regulatory apparatus.

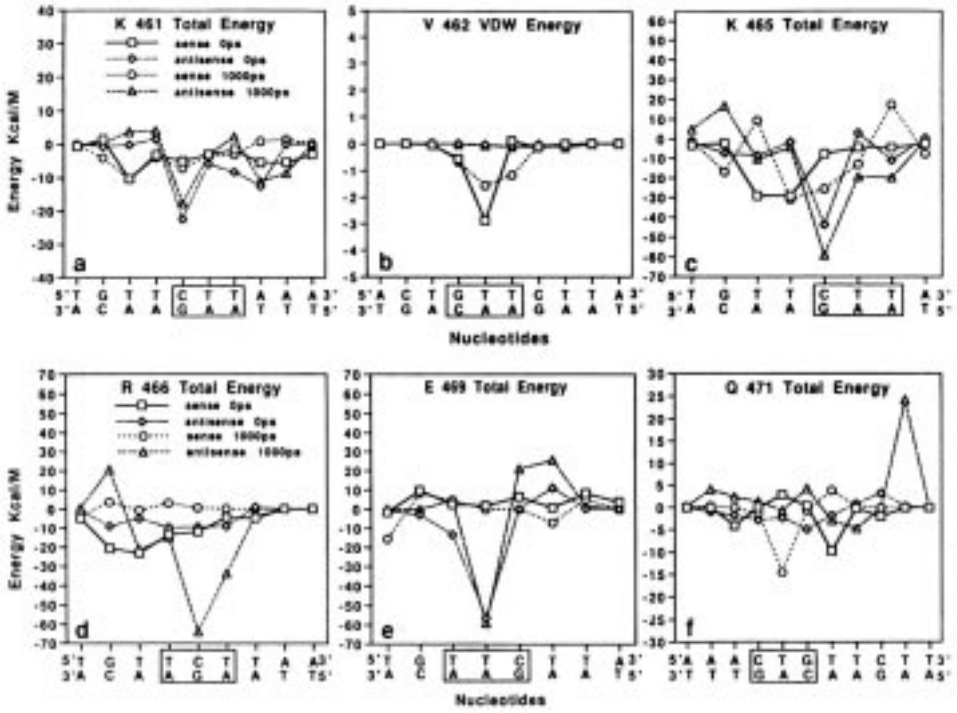

Figure 6. Plots of interaction energy between GRE nucleotides and GR DBD DNA recognition helix amino acids from a GR DBD protein/29 bp GRE DNA model after energy minimization, heating and equilibration (0 picoseconds) and after 1000 picoseconds of molecular dynamics. Codon/anticodon nucleotides reading 5' to 3' for the respective amino acids are boxed. The sense strand is on top. (a) Lysine 461 total interaction energies with GRE nucleotides 17–26 on the sense strand and 3–42 on the antisense strand. (b) Valine 462 VDW interaction energies with GRE nucleotides 15–24 on the sense strand and 35–44 on the antisense strand. (c) Lysine 465 total interaction energies withGRE nucleotides 17–24 on the sense strand and 35–42 on the antisense strand. (d) Arginine 466 total interaction energies with GRE nucleotides 17–25 on the sense strand and 34–42 on the antisense strand. (e) Glutamic acid 469 total interaction energies with GRE nucleotides 17–24 on the sense strand and 35–42 on the antisense strand. (f) Glutamine 471 total interaction energies with GRE nucleotides 13–23 on the sense strand and 36–46 on the antisense strand.

The above studies (Harris *et al*., 1990a,b, 1993, 1994, 1995, 1996a,b) and those that have shown (1) a correlation between amino acids' sidechain physicochemical characteristics and the nucleotides of their cognate codons (Jungck, 1978; Pieber and Toha, 1983; Hendry *et al*., 1984; Lacey *et al*., 1984), (2) stereochemical complementarity and structural relationships between amino acids and their cognate codon and/or anticodon nucleotides (Harris *et al*., 1990a,b, 1993; Hendry *et al*., 1984; Lacey *et al*., 1984; Al'tshtein and Efimov, 1988; Hendry *et al*., 1995), (3) a direct in vitro binding preference for amino acids and their cognate codon or anticodon nucleotides (Harris *et al*., 1994, 1995, 1996a,b; Yarus, 1989, 1991, 1998; Miller-Hill and Kolkhof, 1994; Saxinger, C. and Ponnamperuma, 1974; Desjari-

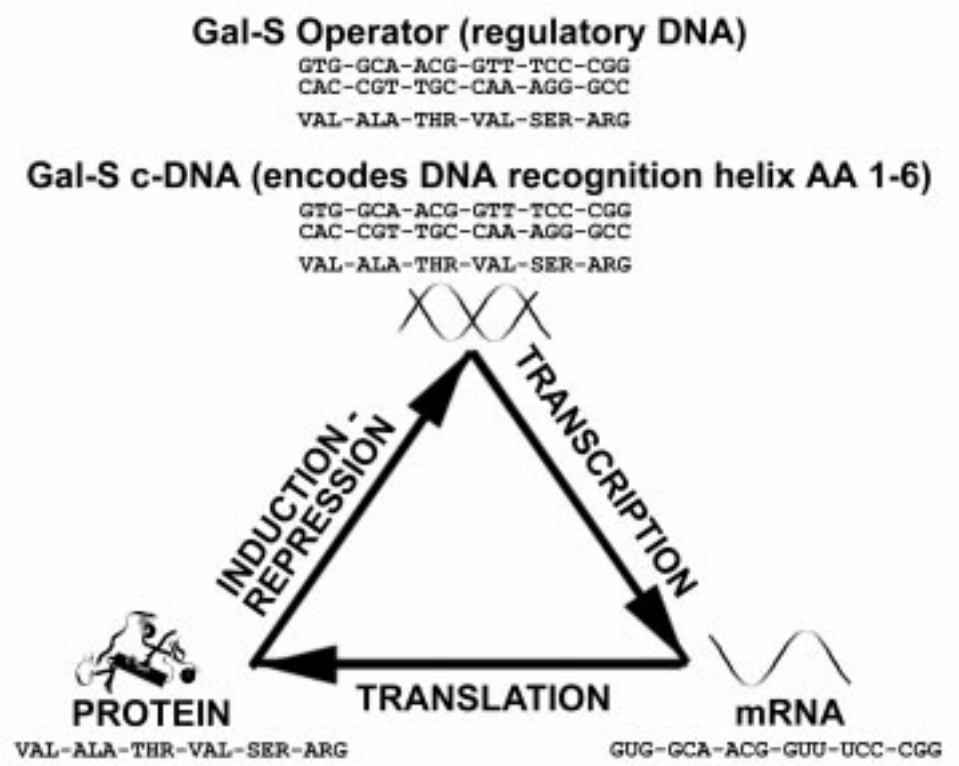

Figure 7. Protein:nucleic acid interactions are directed by genetically conserved sequences in DNA. The c-DNA sequence that directs the synthesis of amino acid residues 1–6 of the DNA recognition helix of the gal/S repressor protein shares identical codon sequence for these amino acids within the gal/S operator sequence to which it specifically binds.

ais and Berg, 1992; Harris *et al*., 1997), provide strong evidence for a directed molecular evolution in the origin of the genetic code (Woese, 1968).

The question of whether the sequences in DNA that are responsible for directing protein interactions (both intramolecular interactions that determine protein shape and intermolecular interactions that determine recognition between different proteins) may be related has been addressed (Blalock and Smith, 1984; Blalock, 1990; Brentani, 1990; Blalock, 1995). It has been proposed that protein folding and protein:protein interactions are determined by the recognition of peptides of opposite hydropathic properties. This idea is based on a relationship between the genetic code and the hydropathic character of amino acids; complementary codon nucleotide sequences specify amino acids that are hydropathically opposites (Blalock and Smith, 1984). Thus, complementary nucleotide sequences theoretically encode interacting peptides (see Figure 8). Support of this notion is provided by reports of sense-antisence peptide recognition (Blalock and Smith, 1984; Blalock, 1990; Brentani, 1990; Blalock, 1995).

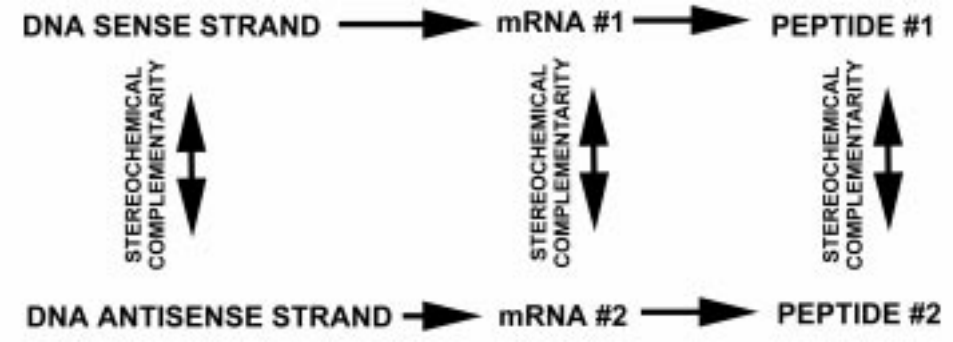

Figure 8. Protein:protein interactions are directed by complementary sequences in DNA. These complementary DNA sequences encode complementary peptide sequences which are hydropathic opposites.

The proposal presented herein states that sequences in DNA coding for nucleic acid:nucleic acid, protein:nucleic acid and protein:protein interactions must be related. As discussed above, the sequences in DNA that direct nucleic acid interactions are related as are many of the sequences in DNA that direct protein:nucleic acid interactions. There is emerging evidence that the sequences in DNA that direct protein: protein interactions are also related. If this proposal is correct, then the relationships between the sequences in DNA coding for these these interactions constitute a life code of which the genetic code is only one aspect of the many related interactions encoded in DNA.

References

Al'tshtein, A. and Efimov, A.: 1988, *Mol. Biol.* (Moscow) **22**, 1411, 1988; English Transl., *Mol. Biol.* **22**, 1133.

Blalock, J. E. and Smith, E. M.: 1984, *Biochem. Biophys. Res. Commun.* **121**, 203.

Blalock, J. E.: 1990, *Trends Biotechnol.* **8** (6), 140.

Blalock, J. E.: 1995, *Nature Medicine* **1**, 876.

Brentani, R. R.: 1990, *J. Mol. Evol.* **31**, 239.

Chou, P. and Fasman, G.: 1978, *Ann'l Rev. Biochem.* **47**, 251.

Crick, F. H. C.: 1966, *J. Mol. Biol.* **18**, 548.

Dayhoff, M.: 1978, *Sequence and Structure*, National Biomedical Research Fdn., Silver Spring MD.

Desjariais, J. and Berg, J.: 1992, *Proteins* **12**, 101, and correction **13**, 273.

Hard, T., Kellenbach, E., Boelens, R., Maler, B., Dahlman, K., Feedman, L., Carkstedt-Duke, J., Yamamoto, K., Gustafsson, J. and Kaptein, R.: 1990, *Science* **249**, 157.

Harris, L. F., Sullivan, M. R. and Hickok, D. F.: 1990a, *Computers Math. Applic.* **20**, 1.

Harris, L. F., Sullivan, M. R. and Hickok, D. F.: 1990b, *Computers Math. Applic.* **20**, 25.

Harris, L. F., Sullivan, M. R,. and Hickok, D. F.: 1993, *Proc. Acad. Sci. U.S.A.* **90**, 5534.

Harris, L. F., Sullivan, M. R., Popken-Harris, P. D. and Hickok, D. F.: 1994, *J. Biomol. Struct. Dynam.* **12**, 249.

Harris, L. F., Sullivan, M. R., Popken-Harris, P. D. and Hickok, D. F.: 1995, *J. Biomol. Struct. Dyn.* **13**, 423.

Harris, L. F., Sullivan, M. R., Popken-Harris, P. D. and Hickok, D. F.: 1994a , *Biol. Struct. Dynam. Proc. 9th Conversation, State Univ. N. Y. Albany, N.Y. 1995*, R. H. Sarma and M. H. Sarma (eds.), Adenine press.
Harris, L. F., Sullivan, M. R., Popken-Harris, P. D. and Hickok, D. F.: 1996b, University of Minnesota Supercomputer Institute Research – UMSI **96**/100.
Harris, L. F., Sullivan, M. R. and Popken-Harris, P. D.: 1997, *J. Biomol. Struct. Dyn.* **15**, 407.
Hendry, L., Bransome, E., Jr., Hutson, M. and Campbell, L.: 1984, *Perspect. Biol. Med.* **27**, 623.
Hendry, L. B., Mahesh, V. B.,Bransome Jr., E. D., Hutson, M. S. and Campbell, L. K.: 1995, The World Wide Web J. Biol. (ww.epres.com/w3jbio) **1**-3.
Hatfield, D., Lee, B. and Pirtle, R.: 1992, *Transfer RNA and Protein Synthesis*, CRC Press, Boca Rotan, Fl U.S.A., pp. 87–112; pp. 113–124; pp. 125–140.
Jungck, J. J.: 1878, *Mol. Evol.* **11**, 211.
Lacey, J., Mullins, D. Jr. and Khaled, M.: 1984, *Origins Life* **14**, 505.
Luisi, B., Xu, W., Otwinowski, Z., Freedman, L., Yamamoto, K. and Sigler, P.: 1991, *Nature* **352**, 497.
Miesfeld, R., Godowaki, P., Maler, B. and Yamamoto, K.: 1987, *Science* **236**, 423.
Muller-Hill, B. and Kolkhof, P.: 1994, *Nature* **369**, 614.
Nelsestuen, G. J.: 1978, *Mol. Evol.* **11**, 109.
Nelsestuen, G.: 1979, *Biochemistry* **18**, 2843.
Osawa, S., Jukes. T. H., Watanabe, K. and Muto, A.: 1991, *Microbiol. Res.* **56**, 229.
Osawa, S.: 1995, *Evolution of the Genetic Code*, Oxford University Press, p. 205.
Payvar, F., DeFranco, D., Firestone, G., Edgar, B., Wrange, O., Okret, S., Gystafsson, J. and Yamamoto, K.: 1983, *Cell* **35**, 381.
Pieber, M. and Toha, J.: 1983, *Origins Life* **13**, 139.
Saenger, W.: 1988, *Principles of Nucleic Acid Structure*, Springer-Verlag, New York, M.Y.
Saxinger, C. and Ponnamperuma, C.: 1974, *Origins Life* **5**, 189.
Scheidereit, C., Geisse, S., Westphal, H. and Beato, M.: 1983, *Nature* **304**, 749.
Scheidereit, C. and Beato, M.: 1984, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 3029.
Sjostrom, M. and Wold, S. J.: 1985, *Mol. Evol.* **22**, 272.
Watson, J. D. and Crick, F. H. C.: 1953, *Nature* **171**, 737.
Woese, C.: 1968, *Proc. Natl. Acad. Sci. U.S.A.* **59**, 110.
Yarus, M.: 1991, *New Biologist* **3**, 183.
Yarus, M. and Christian, E.: 1989, *Nature* **342**, 349.
Yarus, M.: 1998, *J. Mol. Evol.* **47**, 109.