**Title:** Biopolymers: Proteins and Nucleic Acids

**Author(s):** John H. Richards; John N. Abelson; Peter B. Dervan; Leroy H. Hood; Melvin I. Simon

The scientific work supported by N00014-86-K-0775 focused on the basis of information content in macromolecules and how this information directs various biological processes. Specifically, we dealt with the activity of the two major classes of biological macromolecules, proteins and nucleic acids. We addressed such questions as: (i) how does the base sequence in a region of DNA influence its three-dimensional structure especially in ways that might facilitate the exquisitely specific recognition of a particular site in the DNA by a particular protein? (ii) how does the amino acid sequence of the protein that binds specifically to a particular site in DNA influence this binding and conversely how does this interaction depend on the sequence of the DNA? (iii) how can we use these insights to design proteins that will specifically recognize and cleave DNA? (iv) how does the base sequence in tRNA determine which amino acid will be specifically charged in a process that lies at the heart of the translation of the nucleic acid sequence of a gene into the unique amino acid sequence of the protein that the gene encodes?
FINAL REPORT
9/15/86-9/30/92
March 13, 1995

Biopolymers: Proteins and Nucleic Acids

California Institute of Technology
Pasadena, California 91125

Co-Investigators: John N. Abelson, Leroy E. Hood* and Melvin I. Simon, Division of Biology; Peter B. Dervan and John H. Richards, Division of Chemistry and Chemical Engineering

*Current address: Department of Molecular Biotechnology, University of Washington, Seattle, Washington

Contract: N00014-86-K-0755

Ph.D. Students Supported by Project
Michael Bruist
Thomas Chang
Michael Emerling
Kenneth S. Graham
Herman Hesen
Diane Hollenbaugh
Brent L. Iverson
David Long
David P. Mack
Tadashi (Jack) Mizoguchi
Jennifer Normanly
Martha G. Oakley
Todd Richmond
James P. Sluka
Claire Slutter
Yun Sun

Postdoctoral Fellows Support by Project
Jamie E. Arenas
Wonhwa Cho
Anna C. Glasgow
Kelly T. Hughes
Heon Man Lim

Progress Report

The scientific work supported by N00014-86-K-0755 focused on the basis of information content in macromolecules and how this information directs various biological processes. Specifically we dealt with the activity of the two major classes of biological macromolecules, proteins and nucleic acids. We addressed such questions as: (i) how does the base sequence in a region of DNA influence its three-dimensional structure especially in ways that might facilitate the exquisitely specific recognition of a particular site in the DNA by a particular protein? (ii) how does the amino acid sequence of the protein that binds specifically to a particular site in DNA influence this binding and conversely how does this interaction depend on the sequence of the DNA? (iii) how can we use these insights to design proteins that will specifically recognize and cleave DNA? (iv) how does the base sequence in tRNA determine which amino acid will be specifically charged in a process that lies at the heart of the translation of the nucleic acid sequence of a gene into the unique amino acid sequence of the protein that the gene encodes?
The work has resulted in a number of papers and reviews (33 are submitted with this report) that cover these various issues and will be discussed with relation to the four questions just outlined.

The particular system on which we have focused the work of the first three questions plays a central role in a cell’s response to its external environment, specifically response to concentration gradients of nutrients by *Salmonella typhimurium* as a manifestation of the general phenomenon of chemotaxis. The biological mechanism involves an inversion of a gene that encodes a protein essential for operation of the bacterial flagella. In one orientation the protein is expressed; in the other orientation expression is suppressed.

\[
\begin{array}{c|c|c|c|c|c}
A & A' & B & B' & \text{Hin Recombinase} & A' & B \\
\hline
\text{protein expressed} & \text{expression suppressed}
\end{array}
\]

The system that accomplishes this rearrangement includes a number of proteins operating in concert. The most important of these proteins, Hin recombinase, specifically recognizes base sequences in the DNA labeled A, A', B, B' in the diagram and formally termed *hix* sites.

The fourth question focuses on the translation of a nucleic acid message into a protein. This process requires a complex biological machinery, one of whose essential elements is a family of transfer RNA species each of which accepts a particular amino acid which it then carries to the ribosome where that particular amino acid is inserted into the growing protein chain according to the sequence specified in the messenger RNA. The charging of a particular tRNA with its specific amino acid is accomplished by one of a class of enzymes that recognize certain structural features of a particular tRNA and reliably charge it with the appropriate amino acid. The act of charging each species of tRNA only with its cognate amino acid constitutes the central decoding step in translation of the information from the world of nucleic acids into the sequences of amino acids in a protein. The amino acid sequences of a protein in turn mandate the complex three-dimensional structures and biological functions that characterize proteins.

Together these biological phenomenon provide superb cases whose detailed analysis reveals the general features of the highly specific interactions between proteins and nucleic acids that provide the highly sophisticated recognition between the two most important classes of biological macromolecules.

(i) **Sequence-dependent structure of DNA.** Multi-dimensional magnetic resonance studies of two 14 base pair DNA molecules that contain the recognition sites for Hin recombinase (specifically R. *hixL*; d(GGTTTTGATAAAAG)*d*(CTTTATCAAAAACC) and L. *hixL*; d(GGTTCTTGGAAAACC)*d*(GGTTTCAAGAACC) show that these two sites have remarkably similar structures both of which differ appreciably from classical B-DNA. The refined DNA structures possess a significant bend (25–32°) in the middle of the helices. This bending causes a compression of the major groove at almost exactly the position where the recombinase binds. The DNA molecules were also found to have a deepened and narrowed minor groove near the continuous dA tracts, where the minor groove contacts occur between the N-terminal residues of the recombinase binding domain and the DNA molecules (Figure 1). Such pre-existing structural features of the free DNA molecules are likely to contribute to the specific interactions characteristic of Hin recombinase and the *hix* sites. Even slight changes in the base sequence of the 14 mers causes their structure to revert to that of classical B-DNA (1).
Figure 1.

Final structure of refined DNA R. hixL d(GGTTTTGATAAAG)_d(CTTTATCAAAAAACC). Structures were generated by averaging coordinates in six converged structures. The resulting coordinates have been optimized for force field potentials. A. Side view: bending of the DNA molecule and the positive roll angle are shown. B. Rear view: at the junction between base-pairs T7•A22 and G8•C21, a positive slide has taken place so the interstrand Pur-Pur overlap can be maximized. C. Front view: narrowed major groove is shown.
Three-dimensional x-ray diffraction studies of the protein-nucleic complex have also been initiated (2).

(ii) How does the amino acids sequence of the protein that specifically binds to a particular site in DNA influence this interaction and conversely, how does this interaction depend on the structure of the DNA.

A study of phase variation in Salmonella analyzed the dependence of this biological phenomenon on the in vivo interaction between Hin recombinase and the hix recombination site. The bacteriophage P22-based challenge phase selection was used to characterize the binding of Salmonella Hin recombinase to the wild-type hixL or hixR recombination sites and to repress transcription from an upstream promoter in the challenge phase system. Hin-mediated repression results from Hin associating into multimers either prior to binding or during the cooperative binding process at the hix operator sites. The ability of Hin multimers to repress transcription is eliminated when the hix 13 bp half-sites are rotated to opposite sides of the DNA helix by inserting 4 bp between them. Insertion of 1 bp between half-sites reduces overall repression. Hin also binds one of the hixL half-sites to repress transcription, but only when high levels of Hin protein are present in the cell. Mutations have been identified in the hix sites that impair Hin binding. Five of the 26 bp in the hix sites are critical; sites with base-pair substitutions at these five positions show greatly reduced binding. Three additional base pairs make minor contributions to binding. These results are consistent with the results of binding studies between Hin and the hix sites in vitro (3).

Additional studies using nuclease and chemical protection further defined these interactions. The recombinase of the Salmonella inversion system, Hin, mediates site-specific recombination between two 26 bp inverted repeat sequences (hixL and hixR) which flank a 993 bp DNA segment. We investigated Hin recognition of, and association with, the hix recombination sites. Nuclease and chemical protection studies with linear and supercoiled DNA substrates demonstrate that Hin initially binds hixL and hixR independently of binding of the other protein components of the inversion system, Fis and HU. DNA-binding assays with mutant recombination sites and methylation interference experiments indicate that the critical bases for Hin recognition of its DNA-binding site are within an 8 bp sequence covering adjacent major and minor grooves of the DNA helix in each of the 12 bp half-sites of the hix recombination sites. The nature of the Hin-hix complexes in these binding studies and the results of gel filtration assays with purified Hin suggests that Hin binds the recombination sites as a dimer (4).

Use of DNA sequence variants that have pairs of single base substitutions reveal further details of the specificity of the Hin-hix recognition. The Hin recombinants of Salmonella catalyzes a site-specific recombination event which leads to flagellar phase variation. Starting with a fully symmetrical recombination site, hixC, a set of 40 recombination sites which vary by pairs of single base substitutions was constructed. This set was incorporated into the Salmonella-specific bacteriophage P22-based challenge phase selection and used to define the DNA sequence determinants for the binding of Hin to DNA in vivo. The critical sequence-specific contacts between a Hin monomer and a 13 bp hix half-site are at two T:A base pairs in the major groove of the DNA which are separated by one base pair, and two consecutive A:T contacts in the minor groove. The base substitutions in the major groove recognition protein which were defective in binding Hin still retained residual binding capability in vivo, while the base pair substitutions affecting the minor groove recognition region lost all in vivo binding. Using in vitro binding assays, Hin was found to bind to hix symmetrical sites with A:T base pairs in the minor groove recognition sequences, but not to G:C base pairs. In separate in vitro binding assays, Hin was equally defective in binding to either a G:C or a I:C contact in a major groove recognition sequence. Results from in vitro binding assays to hix sites in which 3-deazaadenosine was substituted for adenine are consistent with Hin making a specific contact to either the N3 of adenine or O2 of thymine in the minor groove within the hix recombination site.
on each symmetric half-site. These results taken with the results of previous studies on the DNA binding domain of Hin suggest a sequence-specific minor groove DNA binding motif (5).

As an aid in further analyses of this recognition, an artificial recombination site that is composed of two identical half sites that bind Hin recombinase was constructed and the effects of DNA mutations on recombination determined. An artificial recombination site hixC composed of two identical half-sites that bind the Hin recombinase served as a better operator in vivo than the wild type site hixL. In vitro binding assays such as gel retardation assay and methylation protection assay demonstrate that Hin binds to hixC as tightly as it binds to hixL, even when the sites are located in negatively supercoiled plasmids. However, hixC served as a poor recombination site when it was subjected to the standard inversion assay in vitro. hixC showed a 16-fold slower inversion rate than the wild type.

A series of biochemical assays designed to probe different stages of the Hin-mediated inversion reaction, demonstrated that Hin dimers bind to hixC have difficulty in forming paired hix site intermediates. KMnO4 and S1 nuclease assays detected an anomalous structure of the center of hixC only when the site was in negatively supercoiled plasmids. Mutational analysis in the central region of hixC and assays of paired hix site formation with topoisomers of the hixC substrate plasmid suggest that Hin is not able to pair hixC sites because of the presence of the anomalous structure in the center of the site. The structure does not behave like a DNA “cruciform” since Hin dimers still bind efficiently to the site. It is thought to consist of a short denatured “bubble” encompassing 2 base pairs.

During the study of mutations in the center of hixC, we found that Hin is not able to cleave DNA if a guanine residue is one of the two central nucleotides close to the cleavage site. Furthermore, Hin acts in a concerted fashion and cannot cleave any DNA strand if one of the four strands in the inversion intermediate is not cleavable (6).

Interestingly supercoiling of the target DNA significantly influences the recombination process. A series of biochemical assays were developed and performed to monitor the molecular events that occur during the Hin-mediated DNA inversion reaction. These events can be divided into five different stages: (1) binding of proteins (Hin, Fis, and HU) to DNA; (2) pairing of Hin-binding sites; (3) inverterosome formation; (4) dDNA strand cleavage; (5) strand rotation and religation. A series of topoisomers of the wild type DNA substrate plasmid [ranging from fully relaxed molecules to those with more than the physiological superhelical density (the physiological superhelical density of pKH336 from E. coli EH10B is ~0.072 in this study)] was generated, and the role of negative supercoiling in each step of the inversion reaction was investigated. We found differences in the dependence of the formation of paired Hin-binding sites and of the inverterosome formation on the superhelical density of the substrate plasmid. Pairing of Hin-binding sites occurs independently from inverterosome formation, and a relatively low degree of negative supercoiling is enough to promote maximal pairing. However, efficient inverterosome formation requires higher levels of negative supercoiling (7).

Minor groove contacts, as well as other interactions, have an important influence on the recognition of DNA by the binding domain of Hin recombinase. Incorporation of the DNA-cleaving moiety EDTA•Fe at discrete amino acid residues along a DNA-binding protein allows one to determine the positions of these residues relative to DNA bases, and hence the organization of the folded protein by mapping using high-resolution gel electrophoresis. A 52-residue protein, based on the sequence-specific DNA-binding domain of Hin recombinase (139-190), with EDTA at the amino terminus cleaves DNA at Hin recombination sites. The cleavage data for EDTA-Hin(139-190) reveal that the NH2 terminus of Hin(139-190) is bound in the minor groove of DNA near the symmetry axis of Hin-binding sites. Six proteins varying in length from 49 to 60 residues and corresponding to the DNA-binding domain of Hin recombinase, were synthesized by solid-phase methods: Hin(142-190), Hin(141-190), Hin(140-
190), Hin(139-190), Hin(135-190), and Hin(131-190) were prepared with and without EDTA at the amino termini in order to test the relative importance of the residues Gly139-Arg140-Pro141. Arg142, located near the minor groove, for sequence-specific recognition at five imperfectly conserved 12-base-pair binding sites. Footprinting and affinity cleaving reveal that deletion of Gly139 results in a protein with affinity and specificity similar to those of Hin(139-190) but that deletion of Gly139-Arg140 affords a protein with altered affinities and sequence specificities for the five binding sites. It appears that Arg140 in the DNA-binding domain of Hin is important for recognition of the 5'-AAA-3' sequence in the minor groove of DNA. Our results indicate modular DNA and protein interactions with two adjacent DNA sites (major and minor grooves, respectively) bound on the same face of the helix by two separate parts of the protein.

(iii) How can we use these insights to design proteins with attached functional groups that specifically bind and cleave DNA. Aspects of this subject have been reviewed in some breadth. High-resolution crystallographic views of protein-DNA complexes reveal the structural complexity of protein-DNA interactions. The combination of direct protein-DNA contacts mediated by multiple hydrogen bonds and sequence-dependent DNA conformational effects limits our ability to make detailed structural predictions, even if a new DNA-binding protein can be assigned to a structural class such as helix-turn-helix, double-barreled helix, zinc-binding finger, or scissor grip-leucine zipper. In the absence of high-resolution crystallographic and nuclear magnetic resonance (NMR) data, solution methods such as affinity cleaving can be used to characterize the topology of protein-DNA complexes and correlate sequence similarities with known structural classes.

The conversion of a sequence-specific DNA-binding protein into a sequence-specific DNA-cleaving protein by covalent attachment of the iron chelator, ethylenediaminetetraacetic acid (EDTA) to a specific amino acid residue creates a class of hybrid affinity-cleaving proteins that are available through chemical synthesis. Moreover, a structural domain consisting of naturally occurring amino acids that bind transition metals and oxidatively cleaves DNA extends this method to recombinant methods for protein synthesis.

The approach of affinity cleavage has been used to achieve this objective. In relation to the Hin recombinase problem this approach has involved synthetic 52-residue peptide based on the sequence-specific DNA-binding domain of Hin recombinase (139-190) that has been equipped with ethylenediaminetetraacetic acid (EDTA) at the amino terminus. In the presence of Fe(II), this synthetic EDTA-peptide leaves DNA at Hin recombination sites. The cleavage data reveal that the amino terminus of Hin(139-190) is bound in the minor groove of DNA near the symmetry axis of Hin recombination sites. This work demonstrates the construction of a hybrid peptide combining two functional domains: sequence-specific DNA binding and DNA cleavage.

A Hin DNA cleaving protein has also been used to reveal the orientation of the putative recognition helix in the DNA binding domains of Hin recombinase. On the basis of sequence similarity with other known DNA-binding proteins, the DNA-binding domain of Hin recombinase, residues 139-190, is thought to bind DNA by a helix-turn-helix motif. Two models can be considered that differ in the orientation of the recognition helix in the major groove of DNA. One is based on the orientation of the recognition helix found in the 434 repressor (1-69) and λ repressor-DNA cocystals, and the other is based on the NMR studies of lac repressor headpiece. Cleavage by EDTA-Fe attached to a lysine side chain (Ser183→Lys183) near the carboxyl terminus of Hin(139-184) reveals that the putative recognition helix is oriented toward the center of the inverted repeats in a manner similar to that seen in the 434 and λ repressor-DNA cocystals.
Nickel has also been incorporated into a Hin cleaving protein to effect cleavage. A 55-residue protein containing the DNA binding domain of Hin recombinase, residues 139-190, with the tripeptide Gly-Gly-His (GGH) at the amino terminus was synthesized by stepwise solid-phase methods. GGH(Hin139-190) binds sequence specifically to DNA at four 134 base pair sites (termed hixL and secondary) and, in the presence of Ni(OAc)$_2$ and monoperophthalic acid, reacts predominantly at a single deoxyribose position on one strand of each binding sites. We find that, upon treatment with n-butylamine, the DNA termini at the cleavage site are 3’- and 5’-phosphate, consistent with oxidative degradation of the deoxyribose backbone. The nickel-mediated oxidation can be activated with peracid, iodosylbenzene, or hydrogen peroxide. The sequence specificity of the reaction is not dependent on oxidant, but the rates of cleavage differ, decreasing in the order peracid > iodosylbenzene > hydrogen peroxide. Optimal cleavage conditions for a 1 μM concentration of protein are 50 μM peracid, pH 8.0, and 1 equiv of Ni(OAc)$_2$. The preferential cleavage at a single base pair position on one strand of the minor groove indicates a nondiffusible oxidizing species. A change of absolute configuration in the GGH metal binding domain from L-His to D-His [Ni(II)$^{2+}$GGH-(D)-H(Hin139-190)] affords cleavage at similar base pair locations but opposite with regard to strand specificity (12).

This approach has also been extended to the λδ-resolvase system as well as the lac operator. The DNA binding domain of λδ-resolvase, residues 141-183, is thought to bind DNA by a helix-turn-helix motif based on sequence similarities with other known DNA binding proteins. Incorporation of the DNA cleaving moiety, EDTA•Fe, at the NH$_2$ and COOH termini of λδ(141-183) allows the positions of these residues relative to the DNA bases at three resolvase binding sites, each consisting of inverted copies of an imperfectly conserved 9 bp sequence, to be mapped by high resolution gel electrophoresis. The cleavage data for EDTA-λδ(141-183) reveals that the amino terminus of the DNA binding domain of λδ-resolvase is bound proximal to the minor groove of DNA near the center of the resolvase binding sites. Cleavage by EDTA•Fe attached to a lysine side chain (Asn$^{183}$→Lys$^{183}$) at the carboxyl terminus of λδ(141-183) reveals that the putative recognition helix is in the adjacent major groove on the same face of the helix, oriented toward the center of the inverted repeats (13).

Lac repressor (LacR) is a helix-turn-helix motif sequence-specific DNA binding protein. Based on proton NMR spectroscopic investigations, Kaptein and co-workers have proposed that the helix-turn-helix motif of LacR binds to DNA in an orientation opposite to that of the helix-turn-helix motifs of λ repressor, λ cro, 434 repressor, 434 cro, and CAP. In the present work, we have determined the orientation of the helix-turn-helix motif of LacR in the LacR-DNA complex by the affinity cleaving method. The DNA cleaving moiety EDTA•Fe was attached to the N-terminus of a 56-residue synthetic protein corresponding to the DNA binding domain of LacR. We have formed the complex between the modified protein and the left DNA half site for LacR. The locations of the resulting DNA cleavage positions relative to the left DNA half site provide strong support for the proposal of Kaptein and co-workers (14).

(iv) The relation between the base sequence of a given tRNA and the amino acid with which it becomes charged during protein synthesis is the central recognition event in translating the nucleic acid sequence in the mRNA into a specific amino acid sequence in the cognate protein. The structural basis for this fundamental recognition event has been studied by examining how changing certain bases in a tRNA can change the amino acid this mutated tRNA recognizes. In this context, changing 12 nucleotides transformed a leucine tRNA into a serine tRNA indicated that a limited set of residues determine tRNA identity (15). Later, a minimum number of 8 changes were found to be required to change a leucine amber suppressor codon into one recognizing serine. The appropriate changes were localized in the acceptor stem and in the D stem of the tRNA (16).
In pursuing these studies, amber suppressor genes encoding various amino acids had to be constructed. Specifically, amber suppressor genes corresponding to E. coli tRNA\textsuperscript{Phe} and tRNA\textsuperscript{Cys} were constructed for use in amino acid substitution studies as well as protein engineering. The genes for either tRNA\textsuperscript{Phe}\textsubscript{UA} or tRNA\textsuperscript{Cys}\textsubscript{UA} both with the anticodon 5' CTA 3' were assembled from four to six oligonucleotides, which were annealed and ligated into a vector. The suppressor genes are expressed constitutively from a synthetic promoter, derived from the promoter sequence of the E. coli lipoprotein gene. The tRNA\textsuperscript{Phe} suppressor (tRNA\textsuperscript{Phe}\textsubscript{UA}) is 54–100% efficient in vivo, while the tRNA Cys suppressor (tRNA\textsuperscript{Cys}\textsubscript{UA}) is 17–50% efficient. To verify that the suppressors insert the predicted amino acids, both genes were used to suppress an amber mutation in a protein coding sequence. Amino-terminal sequence analysis of the resultant proteins revealed that tRNA\textsuperscript{Phe}\textsubscript{UA} and tRNA\textsuperscript{Cys}\textsubscript{UA} insert phenylalanine and cysteine, respectively. To demonstrate the potential of these suppressors, tRNA\textsuperscript{Phe}\textsubscript{UA} and tRNA\textsuperscript{Cys}\textsubscript{UA} have been used to effect amino acid substitutions at specific sites in the E. coli lac repressor (17).

Indeed a wider range of amber suppressor tRNA genes were constructed using synthetic oligonucleotides. In this regard, we have constructed 17 tRNA suppressor genes from E. coli representing 13 species of tRNA. We have measured the levels in in vivo suppression resulting from introducing each tRNA gene into E. coli via a plasmid vector. The suppressors function at varying efficiencies. Some synthetic suppressors fail to yield detectable levels of suppression, whereas others insert amino acids with greater than 70% efficiency. Results reported in the accompanying paper demonstrate that some of these suppressors insert the original cognate amino acid, whereas others do not. We have altered some of the synthetic tRNA genes in order to improve the suppressor efficiency of the resulting tRNAs. Both tRNA\textsuperscript{His}\textsubscript{UA} and tRNA\textsuperscript{Glu}\textsubscript{UA} were altered by single base changes, which generated -A-A- following the anticodon, resulting in a markedly improved efficiency of suppression. The tRNA\textsuperscript{Glu}\textsubscript{UA} was inactive, but a hybrid suppressor tRNA consisting of the tRNA\textsuperscript{Phe}\textsubscript{UA} anticodon stem and loop together with the remainder of the tRNA\textsuperscript{Pro} proved highly efficient at suppressing nonsense codons. Protein chemistry results reported in the accompanying paper show that the altered tRNA\textsuperscript{His}\textsubscript{UA} and the hybrid tRNA\textsuperscript{Pro}\textsubscript{UA} insert only histidine and proline, respectively, whereas the altered tRNA\textsuperscript{Glu}\textsubscript{UA} inserts principally glutamic acid but some glutamine. Also, a strain deficient in release factor 1 was employed to increase the efficiency of weak nonsense suppressors (18).

Bibliography and Papers Published (work referred to in text)


Other Published Papers


