CRC

PROCESSING of RNA

David Apirion

CRC PRESS

Processing of RNA

Editor

David Apirion, Ph.D.
Professor of Microbiology and Immunology
Department of Microbiology and Immunology
Washington University School of Medicine
St. Louis, Missouri

CRC Press, Inc.
Boca Raton, Florida

Library of Congress Cataloging in Publication Data
Main entry under title:

Processing of RNA.

Bibliography: p.
Includes index.
Contents: Protein-polynucleotide recognition and the RNA-processing nucleases in prokaryotes/Norman R. Pace — Molecular biology of RNA processing in prokaryotic cells/David Apirion and Peter Gegenheimer — Processing of bacteriophage-coded RNA species/Francis J. Schmidt — [etc.]
1. Ribonucleic acid — Metabolism. I. Apirion, David, 1935- . II. Title: Processing of R.N.A.
QP623.P76 1983 574.87'3283 82-22834
ISBN 0-8493-6510-4

Direct all inquiries to CRC Press, Inc., 2000 Corporate Blvd., N.W., Boca Raton, Florida, 33431.

International Standard Book Number 0-8493-6510-4

Library of Congress Card Number 82-22834
Printed in the United States

PREFACE

"RNA processing comes of age," as Dr. R. Perry so aptly put it in naming a recent article. However, there is still much to be learned about it. As can be attested from the chapters in this book, all organisms from bacteriophage to man can do it. This book attempts to cover the various aspects of RNA processing in the researched biological world. For this reason, the organization is by organism rather than by the kind of RNA processed (tRNA, rRNA, etc.). This means that certain phenomena such as RNA ligation in wheat germ and in yeast, or RNA splicing in yeast and *Tetrahymena* will have to be discussed in different chapters.

In a fast-moving field it is unlikely that articles written more than a year ago would be completely up to date. The purpose of this book is to bring to the nonspecialist an overall view as well as an update on the state of the art as it existed in the beginning of 1982, and to the specialist the opportunity to have a single source of information for how the other organisms do it, and also to enable him to find out the status of the various aspects of RNA processing with which he might not be too familiar. Even if only some of these goals are achieved, all those who labored so diligently to bring about the publication of this book would be more than gratified.

D. A.

THE EDITOR

David Apirion, Ph.D., is Professor of Microbiology and Immunology in the Department of Microbiology and Immunology of the Washington University School of Medicine, St. Louis.

Dr. Apirion received his M.Sc. degree from Hebrew University, Jerusalem, in 1960, and subsequently was awarded the Ph.D. degree in Genetics from Glasgow University in 1963. From 1963 to 1965 he was a Post-Doctoral Fellow at Harvard University.

At Glasgow University, Dr. Apirion was an Assistant Lecturer from 1962 to 1963. He came to Washington University as Assistant Professor in the Department of Microbiology in 1965, becoming Associate Professor in 1970. He has held his present position in the Department of Microbiology and Immunology since 1978. In 1973 he was a Visiting Scholar at Cambridge University, England.

Dr. Apirion has received the Tuvia Kushnir Prize (1960), the Alexander Milman Prize (1961), and the Sir Maurice Block Award.

Dr. Apirion serves on the Editorial Board of the *Journal of Bacteriology,* and holds memberships in the Genetics Society of America, the American Society for Microbiology, and the American Association for the Advancement of Science.

Dr. Apirion is the author of more than 100 publications, specializing particularly in molecular genetics and cellular organization.

CONTRIBUTORS

David Apirion
Professor of Microbiology and Immunology
Department of Microbiology and Immunology
Washington University School of Medicine
St. Louis, Missouri

Giuseppe Attardi
Professor of Biology
Division of Biology
California Institute of Technology
Pasadena, California

Glenn Björk
Professor
Department of Microbiology
University of Umeå
Umeå, Sweden

Thomas R. Broker
Senior Scientist
Cold Spring Harbor Laboratory
Cold Spring Harbor, New York

Robert J. Crouch
Research Chemist
Laboratory of Molecular Genetics
National Institutes of Health
Bethesda, Maryland

S. J. Flint
Associate Professor
Department of Biochemical Sciences
Princeton University
Princeton, New Jersey

Peter Gegenheimer
Research Associate
Department of Chemistry
University of Colorado
Boulder, Colorado

Anita K. Hopper
Associate Professor
Department of Biological Chemistry
Milton S. Hershey Medical Center
Hershey, Pennsylvania

Bernard Moss
Head
Macromolecular Biology Section
Laboratory of Biology of Viruses
National Institute of Allergy and Infectious Diseases
National Institutes of Health
Bethesda, Maryland

Joseph R. Nevins
Associate Professor
The Rockefeller University
New York, New York

Norman R. Pace
Professor
National Jewish Hospital and Research Center and
Department of Biochemistry, Biophysics, and Genetics
University of Colorado Medical Center
Denver, Colorado

Francis J. Schmidt
Assistant Professor of Biochemistry
Department of Biochemistry
University of Missouri-Columbia
School of Medicine and College of Agriculture
Columbia, Missouri

TABLE OF CONTENTS

Chapter 1

PROTEIN-POLYNUCLEOTIDE RECOGNITION AND THE RNA PROCESSING NUCLEASES IN PROKARYOTES

Norman R. Pace

TABLE Of CONTENTS

I. INTRODUCTION

Several recent review articles,[1,2] including some in this volume, consider the properties of the few RNA processing enzymes on which we have information. This paper, as well, will survey the known prokaryotic RNA processing enzymes, but not in exhaustive detail. Rather, the author feels it useful to devote much of the available space to a consideration of the features of polynucleotides with which proteins may specifically interact. The fact is that we know little about the molecular details of any specific protein-polynucleotide complex, and the RNA processing enzymes offer excellent models for exploring these. The collection of references used is not intended to be all-inclusive, but rather, generally, to provide access to this literature.

II. PROTEIN-POLYNUCLEOTIDE CONTACTS

None of the RNA processing enzymes is sufficiently well characterized to encourage even speculation on the detailed character of substrate recognition. We are, however, accumulating a reasonably detailed picture of the sorts of interactions which probably occur. It seems of use, therefore, to draw the discussion of substrate recognition a bit further than simple consideration of nucleotide sequences, even in some folded form. The paradigm offered by the sequence-specific DNA restriction endonucleases may have lulled us into overconfidence regarding our abilities to recognize in RNA the same complex information that proteins do, so the chemical details of possible protein-nucleic acid interactions must be borne in mind as we attempt to ferret out the targets of the processing nucleases.

Protein *surfaces* contact polynucleotide *surfaces*. We therefore must consider a recognition/manipulation process in terms of matrices of complementary contacts between the interacting molecules; the unique geometry possible among multiple contacts provides the overall specificity of the interaction. The important questions to pose, then, are (1) what chemical groups in the nucleic acids are potential binding contacts for proteins; and (2) what is their relationship to the surface, i.e., to an interacting protein? We consider, in passing, DNA as well as RNA because they often are considered to be informationally equivalent molecules. Although the similarities are considerable, these two nucleic acids also offer some strikingly different structural and chemical aspects which are instructive to consider. Moreover, sometimes our knowledge regarding certain aspects of protein-nucleic acid interactions is limited to DNA, so that is presented as exemplary.

Except for tRNA, the most detailed structural information that we have on polynucleotides involves the regular double helices.[3-5] These are of concern, here, because RNA processing sites often are found in regions of high secondary structure.

In general, duplex DNA adopts the familiar "B-form" helix under physiological conditions. Certain deoxynucleotide sequences can assume other helical forms (A, D, Z, etc.), but the bulk of the cellular DNA probably is in the B-form. Somewhat in contrast, largely because of conformational constraints imposed by the ribose 2′-OH groups, duplex RNA assumes the A-form of helix. Space-filling projections of the canonical A- and B-form helices and the positioning of the base pairs about the axes are shown in Figures 1 and 2. For the purposes of this discussion, important points to note regarding these structures are the following: bear in mind that the local structures of the polynucleotides probably are very mushy and readily molded by interacting proteins.

1. The periodicity and orientation of the negatively charged phosphate groups differ in the two forms. DNA-B offers a somewhat narrower profile and about 10 nucleotide pair phosphates per 34-Å turn. Phosphate groups project outward from the helix cylinder in DNA-B, but are more tucked into the RNA-A helix, partly blocking the wide groove (Figure 1).

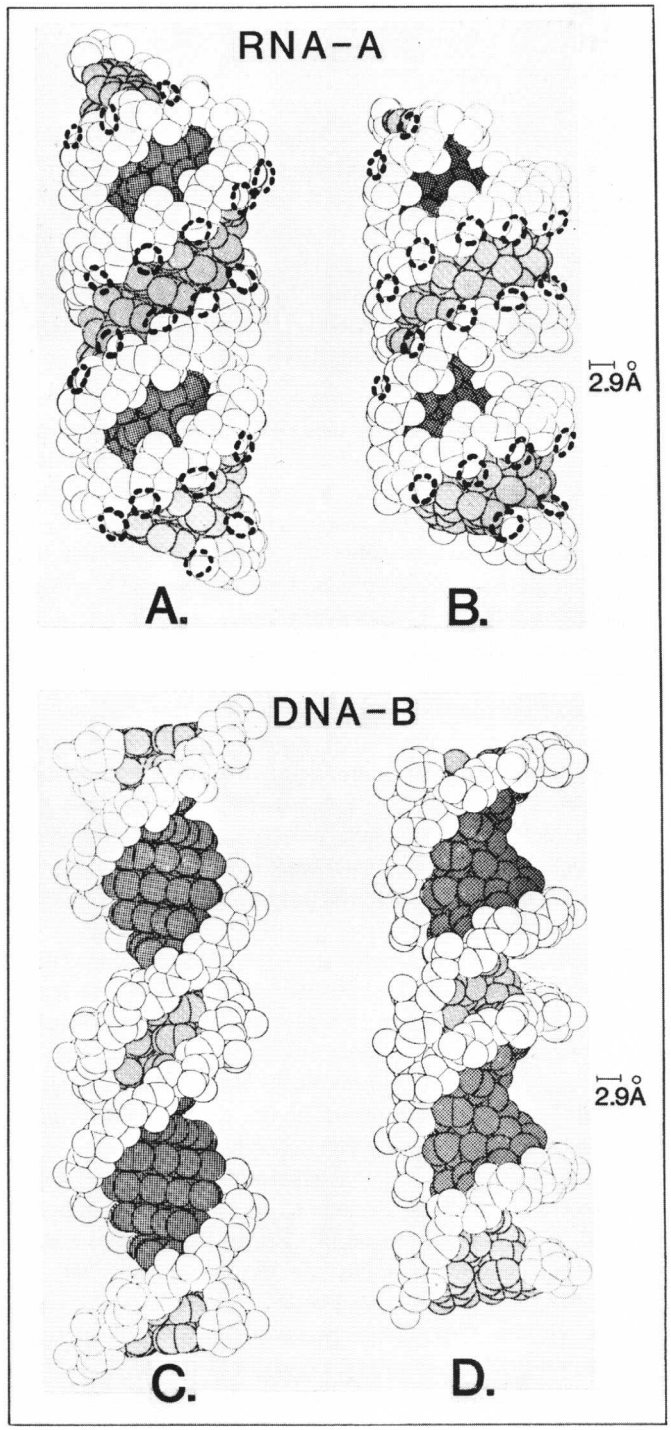


FIGURE 1. Space-filling views of A-RNA (A and B) and B-DNA (C and D) helices. In panels B and D the helices are tilted about 30° to reveal the depths of the helix grooves. Dark shading indicates the wide groove and light shading the narrow groove. In panels A and B, the approximate hydrogen bonding radii of the ribose 2′-OH groups are indicated by dashed circles. The 2.9-Å bar indicates the optimal negative-to-negative center distance for a hydrogen bond contact. (Modified from Alden C. J. and Kim, S.-H., *J. Mol. Biol.*, 132, 411, 1979.)

2. The displacement of the base pairs from the RNA-A helix axis (Figure 2A) defines a very deep, wide groove, which is most evident upon tilting the projection, as seen in Figure 1B. The depth of the wide groove and the barrier offered by the overhanging phosphate groups mean that the base-pair functional groups in the RNA-A wide groove are virtually inaccessible from the surface of the helix. The narrow groove of the RNA-A helix, on the other hand, is superficial (Figures 1 and 2). In DNA-B, because the base pairs are stacked along the helix axis (Figure 2B), both the wide and narrow grooves, in principle, are available to probing groups from an interacting protein.
3. The RNA-A narrow groove is populated by the ribose 2′-OH groups, which are important H-bond donors (Figures 1 and 2). In terms of information content, the 2′-OH groups are a major contrast between RNA and DNA.

Alden and Kim[6] (see also Pullman and Pullman[8]) have provided a more detailed theoretical picture of the information available in the base pairs of the A- and B-form helices by calculating the accessibilities of their various functional groups to "hard-shell" probes of various sizes. Their findings are that, for probe radii greater than 3 Å (amino acids), the helix wide groove contains the most accessible base contacts available in B-form DNA, and only the narrow groove of RNA-A has significant base exposure. The phosphate and 2′-OH groups are freely available.

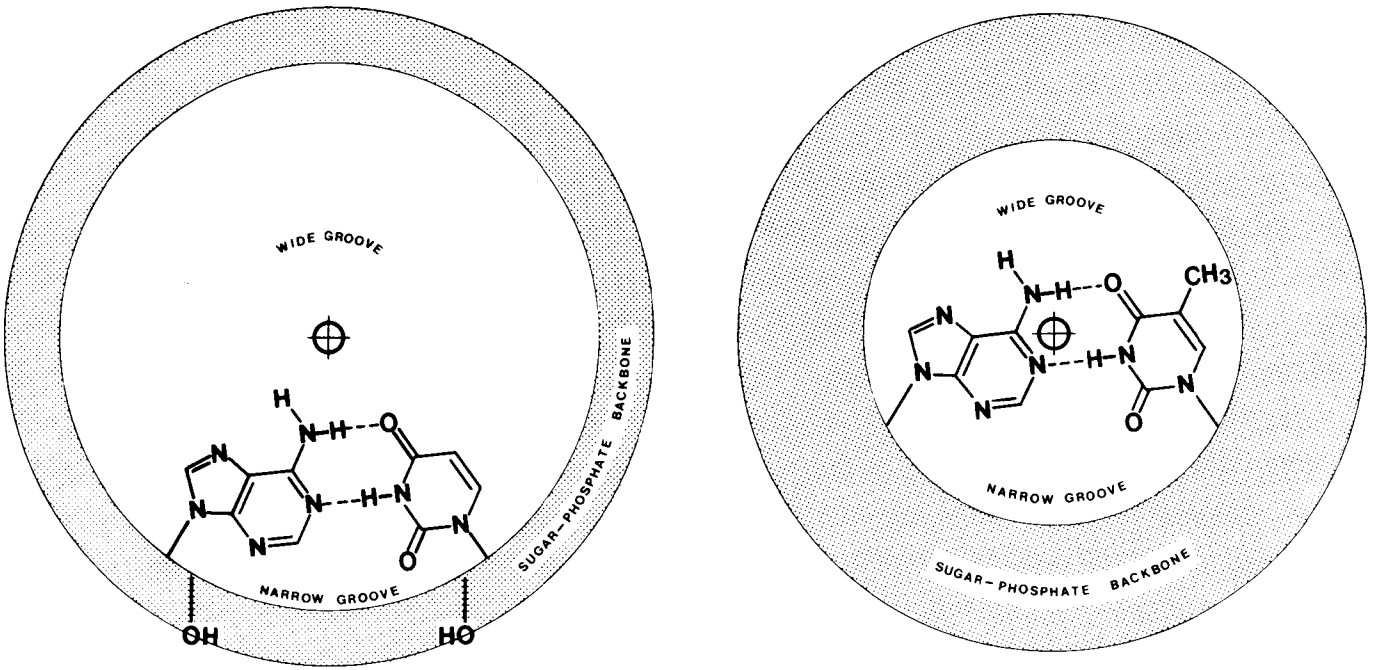
Considerably less attention has been given to crystallographic analysis of "single-strand" polynucleotides than to the duplexes. It is clear from these and other studies, however, that unpaired sequences are far from disordered. Even without the constraints of a complementary pairing, single-strand sequences adopt ordered, helical arrays, stabilized mostly by intramolecular base stacking. Poly A, for example, crystallizes as a right-hand helical structure with about 9 residues per 25-Å repeat;[9] poly C collapses into a 6 base per 18.5-Å repeat helix.[10] Only poly U seems to be substantially unstacked, but is still highly ordered by the conformational constraints of the phosphodiester backbone.[11] In contrast to the regular duplex helices, all of the potential interaction sites for proteins would seem to be freely available in these "single-strand" structures.

Paired sequences containing RNA processing sites often are imperfect complements, containing unpaired or non-Watson-Crick base pairs (G.U, G.A, etc.), or out-of-register complements. At least these latter presumably would yield structures containing bulge loops or extrahelical bases, but little is known regarding their details and how they reflect into adjacent regions. Since protein contacts on the bases seem to be sterically very limited in the regular RNA-A helix, any irregularities within the helices may be important to focus upon in comparative analyses of processing enzyme substrate sites.

Four somewhat overlapping classes of contacts between proteins and polynucleotides can be envisaged.[12-14] These are (1) electrostatic interactions, (2) hydrogen bonding, (3) hydrophobic interactions, and (4) stacking interactions. The most important contributions to protein-polynucleotide binding energies probably are the electrostatic and hydrogen bonds. Stacking and hydrophobic interactions probably are not substantial as proteins confront duplex polynucleotides, but they offer interesting possibilities in considering irregular (non-duplex) nucleic acid conformations or in cases where the nucleic acid conformation is significantly perturbed. Let us consider, now, how these potential contacts are arranged in space in the nucleic acids, and the protein groups which may interact with them.

A. Electrostatic Contacts

Electrostatic interactions are diverse in energy and type; they include the strong ionic contacts afforded by basic amino acids countering the negatively charged phosphate groups in the nucleic acids as well as a hierarchy of weaker and less-defined dipole interactions.



A. RNA-A

B. DNA-B

FIGURE 2. The relationships of the base pairs to the RNA-A and DNA-B helix axes. (Modified from Bloomfield, V. A., Crothers, D. M., and Tinoco, I., Jr., *Physical Chemistry of the Nucleic Acids*, Harper & Row, New York, 1974, 125.)

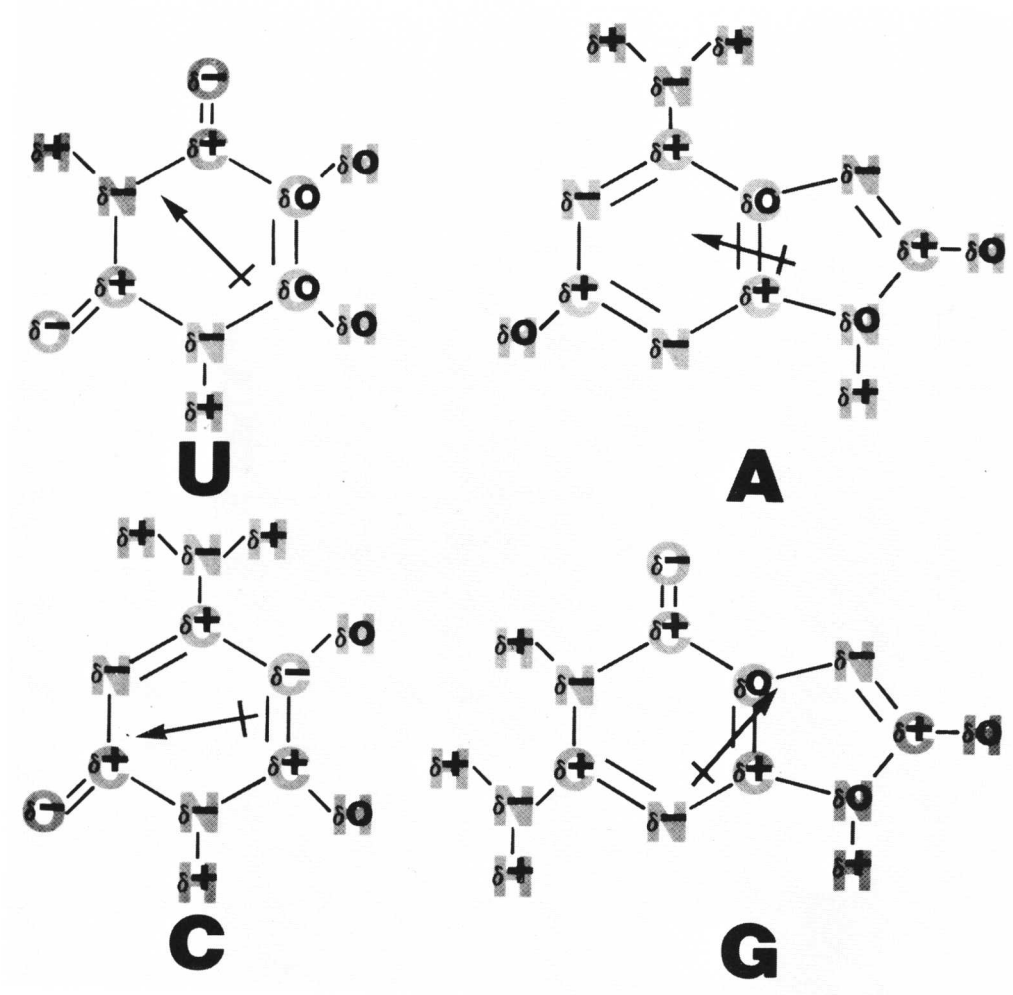


FIGURE 3. The partial charge distributions and permanent dipole moments of the purine and pyrimidine rings. (Modified from Bloomfield, V. A., Crothers, D. M., and Tinoco, I., Jr., *Physical Chemistry of the Nucleic Acids*, Harper & Row, New York, 1974.)

Nucleotide sequences should be viewed not simply as a string of letters, but as some array of partial charges. Figure 3 illustrates the partial charge distribution in the base rings. The intensity of any particular local charge of course is very dependent upon the environment of the base, i.e., its stacking neighbors as well as its association with the solvent, ions, etc.

An important aspect of all the electrostatic contacts is that their interaction energies decay rapidly with increasing distance from the optimum approach. In principle, the interactive energy of two point charges decreases inversely with the square of the distance between them (r); dipole-point charge (e.g., basic amino acid-phosphodiester backbone phosphate) energy decreases with $1/r^3$, dipole-dipole interactions with $1/r^4$, and induced dipole-induced dipole interactions decay with $1/r^6$. In fact, the decay is even more rapid because the interaction energy also is an inverse function of the dielectric constant of the medium. As

water molecules are displaced from the contact sites (water occupies a sphere of about 1.4 Å radius), the dielectric constant of the intervening space decreases 20- to 40-fold and the contact affinity increases correspondingly.

It is obvious, then, that interacting arrays of electrostatic contacts must be precisely in register to be most effective. The strong ionic contacts probably provide the first order of interaction as proteins kinetically nestle onto their nucleic acid targets. Following the primary ionic fit, cooperation from many other and weaker interactions would contribute greatly to the binding energy, if the appropriate matrix of contacts between the molecules is present.

It is easy to see how ionic complementarity between a protein and a polynucleotide can provide considerable specificity. For example, the array of phosphate charges on the RNA-A helix is different from that of the DNA-B helix (Figure 1), therefore selective, strong contacts for proteins are available. Other conformations of course could provide more diverse ionic patterns; the arrays of phosphate contacts which alone could result from RNA tertiary structure seem almost without limit. In our searches for "consensus sequences" associated with RNA processing sites, therefore, it is important to remember that quite different nucleotide sequences can result in a similar matrix of electrostatic (e.g., phosphates) or other groups. The tRNAs offer a good example of this. Also, it is noteworthy that enzymes which operate on structurally different substrates (e.g., RNase P with tRNAs or RNase III with a variety of RNA species, see below) need not always use exactly the same matrix or types of contacts. More intensely negative or positive centers can serve as hydrogen bond donor/acceptor sites (below) as well as dipole contacts; when a dipole interaction involves an in-line proton, it is a hydrogen bond.

In the regular duplex helices, because the bases are tightly stacked, their potential dipole or hydrogen binding contact sites for an interacting protein are available only in the helix grooves, and the sites most diagnostic of the local sequence are in the wide groove.[12] Access to both grooves is no particular problem in the case of B-form DNA or helical, single-strand RNA, but the wide groove in duplex RNA essentially is not available to a protein surface unless it is unwound. The charge center-to-center distance in an effective dipole contact is short, say 2 to 3 Å, and the wide-groove groups in the bases are removed from the RNA-A helix surface considerably farther than that (see Figures 1 and 2). Presumably, then, any base-specific contacts in the intact RNA-A helix must originate from the narrow groove, which essentially is flush with the surface.

The importance of ionic contacts in protein-polynucleotide complexes is evident in their general inactivation at elevated salt concentrations, but oftentimes dependence upon low salt concentrations. This is less well documented for RNA processing or binding proteins than for proteins which bind to DNA, but likely will prove to be a common theme. As an illustration of how substantial these effects may be, the monovalent cation dependence of one rather nonspecific protein-RNA complex is shown in Figure 4.[15] The progressive enhancement of the binding constant as the NA^+ concentration increases to 0.1 *M* is interpreted as a requirement for counterions to shield charged groups (the protein is somewhat acidic) which otherwise would be repulsive, preventing surface contacts by the macromolecules. Then, as the ionic strength of the solution is further increased, the counterion clouds bury the interacting electrostatic pairs and the binding constant of the protein for the RNA is reduced. Record and colleagues,[16] and others, have made elegant use of this phenomenon to distinguish electrostatic and nonelectrostatic contributions to the thermodynamics of protein-polynucleotide interactions.

Appropriate (physiological) solution ionic strengths are demonstrably important to the specificity of processing nucleases. For example, *Escherichia coli* RNase III, which at physiological or higher ionic strength is sufficiently specific to release the rRNAs from their tandem transcript,[17,18] dramatically relaxes its selectivity at low ionic strength (<0.05 *M* Na^+). It then is capable of binding to and hydrolyzing a variety of natural and synthetic

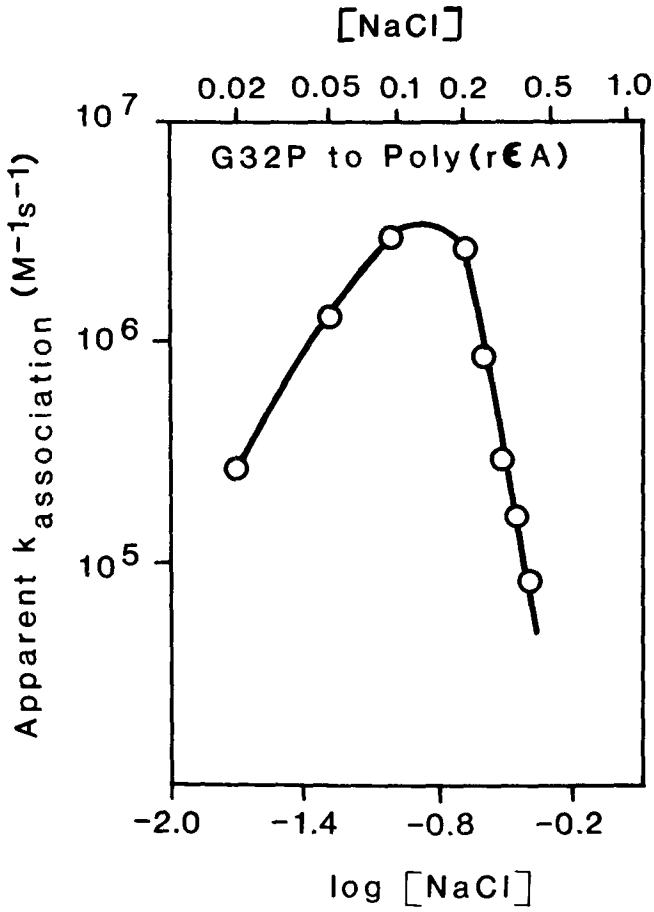


FIGURE 4. The dependence upon salt concentration of the apparent association constant of phage T4 gene 32 protein with poly (1,N^6-ethenoadenylic acid). These measurements with a variety of nucleic acids are described in Reference 15. (Modified from Lohman, T. M. and Kowalczykowski, S. C., *J. Mol. Biol.*, 152, 67, 1981. With permission.)

RNAs.[19] This loss of specificity conceivably is explained by an insufficient cation concentration to shield ionic contacts on the RNA which are not quite in register with basic groups on the protein surface. At higher ionic strengths these spurious contact sites would not participate in the protein-RNA binding, but at low salt concentrations they could mimic, to some extent, a natural cleavage site.

B. Hydrogen Bond Contacts

Hydrogen bonds are the interactions between a proton-containing dipole (H-bond donor) and an electronegative center (acceptor). An important key to thinking of hydrogen bonding interactions, as proteins confront nucleic acids, is their dimension and vectorial nature. The acceptor must be in line (within approximately 20°) with the donor-proton dipole, and the optimal negative center-to-negative center distance for a hydrogen bond is about 2.9 Å. This dimension, relative to the regular helices, is indicated in Figure 1. Much closer approach becomes repulsive, and a bit further is energetically ineffective. Additionally, potential

competition from water molecules means that hydrogen bond donor and acceptor pairs must be positioned very precisely in order to contribute to an interaction.

The polynucleotides offer a variety of potential contributions to hydrogen bonding.[12] Some base nitrogens and the exocyclic base amino groups are potential donors, as are the exocyclic oxy groups, if the base is polarized by its environment (nucleotide sequence or presence of a protein or other ligand) such that the lactim (enol) mode of the base prevails over the normal lactam (keto) form. Additionally RNA, but not DNA, contains the 2′-OH group, a highly conspicuous and important H-bond donor (below). Potential hydrogen bond acceptors associated with the bases are the exocyclic oxygens and the ring nitrogens, and conceivably the exocyclic amino groups in their imino tautomeric forms. The charged backbone phosphate groups are also potential acceptors.

As with the electrostatic contacts, an array of H-bond participants in a polynucleotide can provide a highly specific surface for recognition by a protein surface with a complementary donor/acceptor matrix. The contacts on the nucleic acid may be dependent upon the particular bases involved,[12] but, to reiterate, different nucleotide sequences may provide similar matrices. This is not to say that we should throw up our hands at the prospects of finding "consensus sequences" defining protein action sites in nucleic acids. The local nucleotide sequence certainly determines the detailed, local geometry (helix pitch, base tilt, and alignment relative to the helix axis, etc.) of a potential target. However, even sequence-dependent contacts need not present the same nucleotide sequences along their entire length. Since it seems unlikely (and unnecessary) that interacting proteins wrap extensively around the nucleic acid helices, probably only five or six base pairs per helix turn, even using both grooves, would be in register with a "globular" protein face and available for interaction.

Proteins offer a wealth of contributors to hydrogen bonding. Point contacts are available from the hydroxyl groups of Ser, Tyr, and Thr, the sulfhydryl group of Cys, the amino and amide groups of Lys, Asn, and Gln, and the His imidazole. More complex, somewhat base-specific associations involving the carboxylate anions of Asp and Glu or the Arg guanidinium cation also seem possible (Figure 5).[12,14] Of course, the peptide bond itself offers both an H-bond donor (the amide) and an acceptor (the carbonyl).

As with the in-plane dipoles associated with the base pairs, the DNA and RNA duplexes offer strikingly different aspects in their hydrogen bonding groups which are accessible to an interacting protein. Hydrogen bonding possibilities on the base pairs in duplex DNA are available from both the wide and the narrow grooves, while strictly duplex RNA (because of the deeply recessed A-form wide groove) would seem capable of H-bond contact between the base pairs and proteins only in the narrow groove.

The narrow groove of RNA, however, has information not present in DNA in that it is densely populated by the ribose 2′-OH groups. These undoubtedly play a significant role in many protein-RNA contacts. Carter and Kraut[20] have proposed an attractive model by which proteins might coordinate onto the RNA-A narrow groove, utilizing the 2′-OH groups. Their notion derives from the fact that antiparallel polypeptide β-sheets, which are common components of protein foldings, fit remarkably well to the RNA-A helix, such that alternate carbonyl groups in each of the two peptide chains are in appropriate register for hydrogen bond contact with each 2′-OH group of the nucleotide units. It was envisaged by Carter and Kraut[20] that additional contact might derive by the H-bond bridging of a water molecule between alternate amide protons in the polypeptides and the furanose ring oxygen atoms, although the latter is a questionable H-bond acceptor. Kim and colleagues[21] have pointed out that the antiparallel polypeptide β-sheet also can be fitted to the narrow groove of DNA-B, if alternate amide protons are invoked to form H-bonds with each of the 3′-oxygen atoms in the phosphodiester chain.

In the polypeptide β-sheet, alternate amino acid side chains are on alternate sides of the sheet, so in principle those pointing into the narrow groove could interact with the poly-

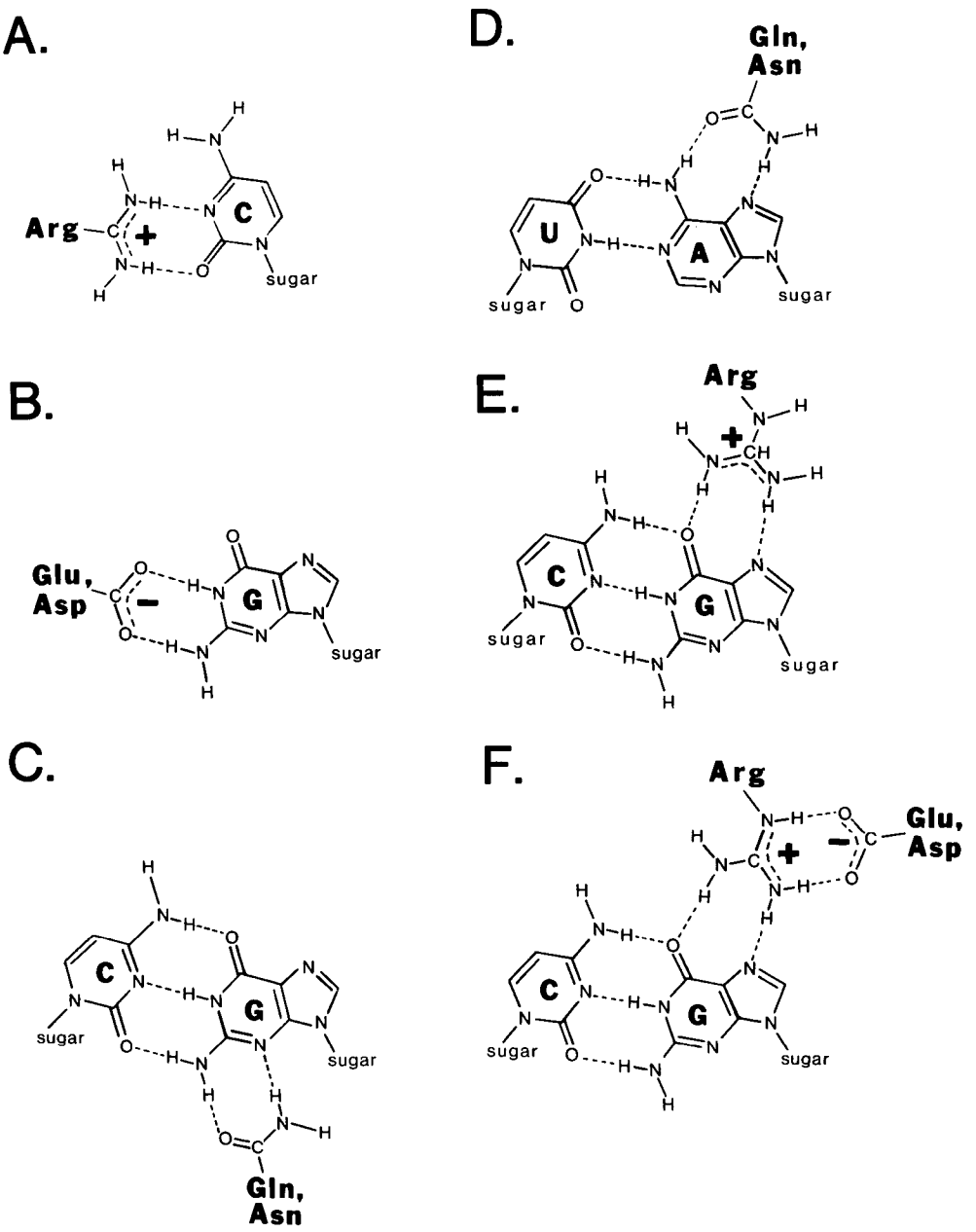


FIGURE 5. Various multiple-point hydrogen bonding associations between amino acids and bases. (Modified from Seeman, N. C. et al.[12] and Helene and Maurizot.[14])

nucleotide base pairs. The narrow groove of DNA, which is somewhat recessed (Figure 1), could accommodate the bulk of most of the amino acid side chains. In RNA-A, however, the narrow groove is essentially flush with the helix surface so would seem to offer steric barrier to the amino acid side chains. Of course, it is possible that the polynucleotide could rearrange somewhat to offer appropriate fit. Imperfections in the Watson-Crick complementarity at or very near the cleaved phosphodiester bonds in double helical processing sites might also offer accommodation to an interacting amino acid side chain.