



Lawrence Berkeley Laboratory

UNIVERSITY OF CALIFORNIA

CHEMICAL BIODYNAMICS DIVISION

Computer Modeling 16S Ribosomal RNA

J.M. Hubbard
(Ph.D. Thesis)

April 1990



LOAN COPY
Circulates
for 2 weeks

Bldg. 50 Library

LBL-29070

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

Computer Modeling 16S Ribosomal RNA

By

John Michael Hubbard

Ph.D. Thesis

April, 1990

Chemical Biodynamics Division

Lawrence Berkeley Laboratory

University of California

Berkeley, CA 94720

This work was supported by the Director, Office of Energy Research, Office of General Life Sciences, Molecular Biology Division of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

Computer Modeling 16S Ribosomal RNA

by

John M. Hubbard

Abstract

The rapid advances in computer graphics and in the power and affordability of computers provide us with an opportunity to develop an objective, reproducible, and flexible modeling procedure which is superior to the physical modeling techniques now in use. It was the goal of this research to develop and implement a rational modeling protocol which could be applied to the folding of 16S ribosomal RNA.

Modeling of a DNA dodecamer and comparison with the structures of the molecule obtained by NMR and X-ray crystallography reaffirm that the results obtained by modeling are dependent on the input data set. Energy minimization produces a final structure which reflects the structural characteristics of the small molecules from which the empirical energy parameters are deduced. Distance geometry structures derived from NMR studies, overvalue the few distances which are known at the expense of other molecular properties. X-ray crystallography can only provide us with information about crystalline conditions and these may be very far removed from the native environment of a biomolecule.

Transfer RNA modeling is used to demonstrate the feasibility of the new protocol and to explore what levels of structural shorthand can be successfully applied. Initially, a representation is employed which replaces each all-atom nucleotide with six pseudoatoms. With an alternate replacement scheme which emphasizes the helical substructures, it is possible to reduce the size of the model to be manipulated by more than twenty-fold and still obtain good results. The modeled structures clearly occupy the same area of

conformational space as that of the tRNAs which have been determined by X-ray crystallography.

The 16S ribosomal RNA at the heart of the small subunit of the ribosome has been extensively studied by physical, chemical, and phylogenetic means. Once the problems produced by the size of this molecule have been addressed, applying the new computer modeling protocol is straightforward. A closely related set of conformers is consistently obtained and these structures are comparable to models which include data from outside the computer parameter set.

ACKNOWLEDGEMENTS

First and foremost I must acknowledge the generous support of my research director, John Hearst. The continuing financial support for both personal and computing needs is the most obvious and not to be underrated. But John was especially generous with his trust in my judgement and that is most uncommon. I also want to thank him for an almost inexhaustible supply of hanging rope.

Bob Glaeser went far beyond the minimal effort that is required of a committee member and were it not for the direction and encouragement that he contributed, I might never have made it.

Steve Holbrook and I had many conversations of theoretical and practical computing. Steve made it clear by example that it is possible to deal intimately with computers and still remain sane. He also managed to acquire the Trojan Horses that kept my programming skills up to par.

Jason Kahn and Sam Lipson made life as a computer hacker much easier to bear by sharing some of the computer load.

Finally I am eternally indebted to Effie Shu, Suzanne Cheng, Pete Spielmann, and Joe Monforte for getting me out of the last minute crack into which I climbed.

Chapter 1. INTRODUCTION	
Ribosome overview	2
Modeling premise	9
References	14
Chapter 2. MODELING SURVEY	15
Modeling Rationale	16
Empirical algorithmic modeling	20
Distance geometry modeling	24
Graphical modeling	28
The unified protocol	33
References	36
Chapter 3. DNA OLIGIONUCLEOTIDE MODELING	
Introduction	39
Oligonucleotide structure	
Modeling rationale	
Materials and Methods	41
Materials	
Software	
Hardware	
Structural data	
Methods	
Amber modeling	
Graphics modeling	
Results	43
NMR solution structure	
Computer generated DNA duplex	
Computer generated DNA hairpin	
Structure characteristics	
Discussion	53
Structures of similar oligonucleotides	
AMBER shortcomings	
Protocol problems	
References	56

Chapter 4. TRANSFER RNA MODELING

Introduction	58
Molecule description	
Modeling rationale	
Materials and Methods	61
Materials	
Software	
Hardware	
Structural data	
Primary structure	
Secondary structure	
Tertiary structure	
Methods	72
Conventional modeling	
Amending the protocol	
Initial reduced atomic representations	
Pseudohelical modeling constructs	
Results	82
Early five pseudoatoms per nucleotide runs (5mer)	
Helical pseudoresidues with 8 constraints per helix (dyn)	
Helical pseudoresidues with 16 constraints per helix	
Structures created with 5 long range constraints (vt)	
Structures created with 12 long range constraints (bad)	
Additional structures made with 5 long range constraints (ext)	
AMBER results	
Discussion	104
Program adjustments	
Alternative pseudoatom trials	
Refinement protocol	
Protocol analysis	
References	118

Chapter 5. 16S RIBOSOMAL RNA MODELING

Introduction	121
16S rRNA function	
16S rRNA shape	
Materials and Methods	125
Materials	
Software	
Hardware	
Structural data	
Primary structure	
Secondary structure	

Tertiary structure	
Phylogenetic relationships	
Psoralen crosslinks	
Ultraviolet crosslinks	
GbzCynAc crosslinks	
Nitrogen mustard crosslinks	
Methods	
Physical modeling	
Computer modeling	
Data preparation	
Distance geometry	
Atomic reconstruction	
Molecular mechanics	
Raster displays	
Results	150
Computer generated structures	
Partial Folds	
Folds with 8 constraints per helix (nt)	
Folds with 16 constraints per helix (last)	
Physical model (byhand)	
Discussion	180
Modeling difficulties	
Refinement problems	
Display adjustments	
Evaluation of models	190
vs physical data	
vs other models in general	
last74 vs byhand model	
vs physical models	
Wollenzien model	
Brimacombe model	
vs computer models	
Noller model	
Protocol analysis	
References	205
Chapter 6. 30S RIBOSOMAL SUBUNIT MODELING	
Introduction	209
IEM map	
Protection map	
Neutron map	
RNA/protein crosslinks	

Materials and Methods	211
Materials	
Computing resources	
Structural data	
16S computer models	
Protein map	
RNA/protein crosslinks	
Methods	215
RNA/protein superposition	
Raster graphics	
Results	219
30S ribosomal subunit models	
RNA/protein crosslink evaluations	
RNA/protein relationships	
Additional ribosomal relationships	
Discussion	237
Data evaluation	
IEM data	
RNA protection data	
Neutron scattering protein map	
Mechanistic implications	
Conclusions	244
Final considerations	
Future directions	
References	248

Chapter 1

INTRODUCTION

Ribosome Overview

Perhaps the most interesting of the large macromolecular complexes is the ribosome. The ribosome is the self-assembling molecular automaton which catalyzes the production of all proteins. It synchronizes the interplay of cofactors, precursors, messenger RNA copy of the DNA, aminoacylated transfer RNAs, and provides the catalytic surface. The ribosome is vital and common to all living organisms as it regulates the informational phase change from nucleic acids to proteins. The translation of the information contained in DNA into proteins is the central catalytic event of all life. The ribosome is present in every cell and sometimes in large quantities. The ribosome appears to be evolutionarily older than all living things and is a good yardstick for interspecies comparison. The ribosome is a quaternary complex of RNAs and proteins. The ribosome also seems to be connected with all the most interesting processes which occur in the cell including RNA processing, RNA catalysis, protein synthesis, proofreading, cooperative self assembly, feedback control and regulation, RNA/protein interactions, RNA/RNA interactions and protein/protein interactions. This simple listing demonstrates that the ribosome will necessarily be a very complex system that will be difficult to study. But for the same reasons investigation of the ribosome has continued despite the size of the problem it poses.

Originally called the microsome, the ribosome was one of the first discoveries made with the electron microscope. Its size (approx. 250 Angstroms) was at the limit of what could be directly imaged in the electron microscope. It appeared as a solid, dark body at the tantalizing edge of perceptibility. The resolution and analysis of electron micrographs have been improved as a direct result of attempts to decipher the complex structure of the ribosome. A typical bacterial cell has 10,000 ribosomes and may contain as many as 25,000 ribosomes during periods of rapid growth. The much larger eucaryotic cell may

contain ten million ribosomes and the RNA components of these automata will comprise more than half of the total cellular RNA. It is these vast numbers that allow the ribosome to be harvested in sufficient quantity for physical chemistry experiments. Gradient sedimentation can be used to isolate a ribosome containing fraction of the cytoplasm of lysed cells. A simple analysis revealed that these intracellular granules contained more ribonucleic acid than proteins. Thus the microsome was renamed the ribose (RIBO) containing body (SOME). When this fraction was mixed with radioactively labeled amino acids, radiolabeled proteins were formed implicating the ribosome in protein synthesis.

Persistent study in many labs by many researchers over the years have determined the components of the ribosome of the most commonly used bacteria, *E. coli*. The 70S ribosome consists of the noncovalent association of two asymmetrical subunits of unequal size. The larger ribosomal subunit has a sedimentation coefficient of 50S and the smaller subunit has one of 30S. The use of the sedimentation labels persists as a reminder of the first quantitative facts derived about the ribosome. The larger subunit is more symmetrical than the smaller and resembles a flattened snowman with a head, a body, and two outstretched arms (fig. 1). Reconstructed images from low resolution X-ray scattering indicates that there is a hole in the body of the 50S subunit through which it is believed the growing peptide chain exits (Yonath et al., 1987). The small subunit is more conical, with a platform-like structure that makes the overall shape resemble a 'Y' (fig. 2). The intersection of the platform and the body form a cleft. The association of the two subunits forms the 70S ribosome (fig. 3). The large subunit has a mass of 1.8 million daltons and the small subunit a mass of 0.9 million daltons.

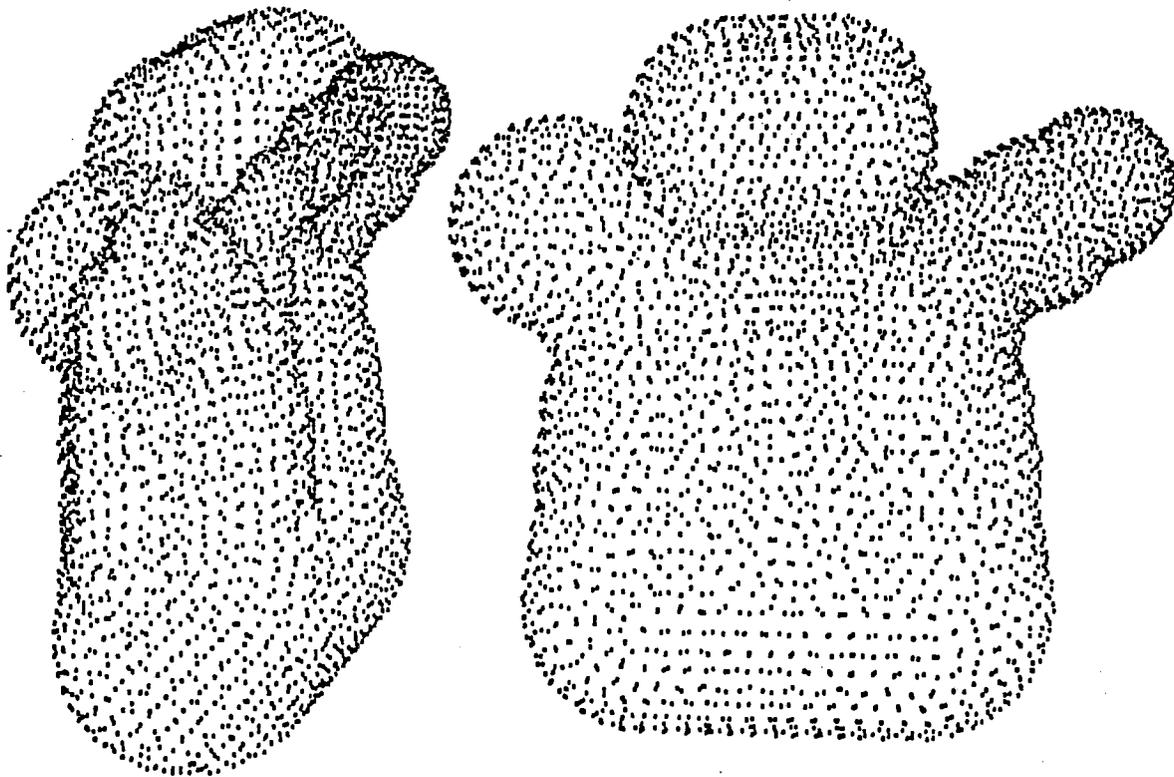


Figure 1. The large subunit of the ribosome. The longest dimension is approximately 250 angstroms. The left side view is rotated by ~ 90 degrees to reveal the curve of the 30S subunit interface.

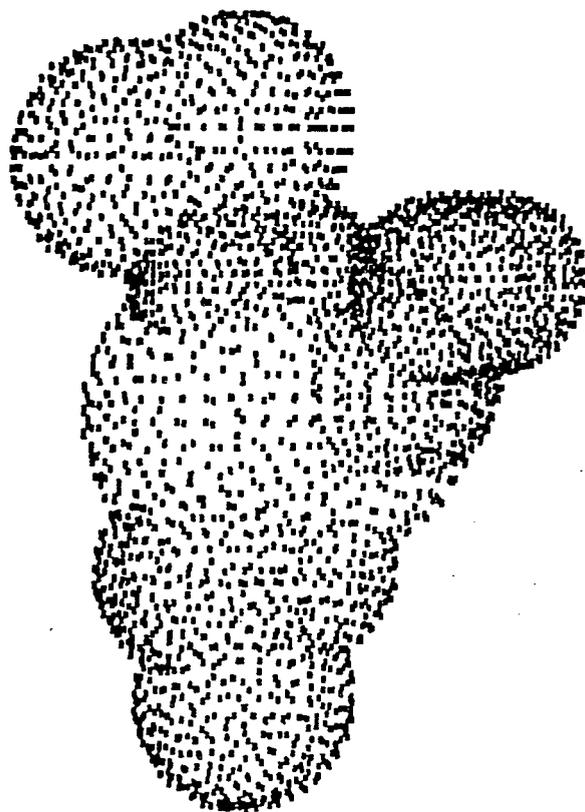


Figure 2. The 30S subunit of the ribosome (size ~ 55 X 220 X 220 angstroms). The 50S interface side is shown.

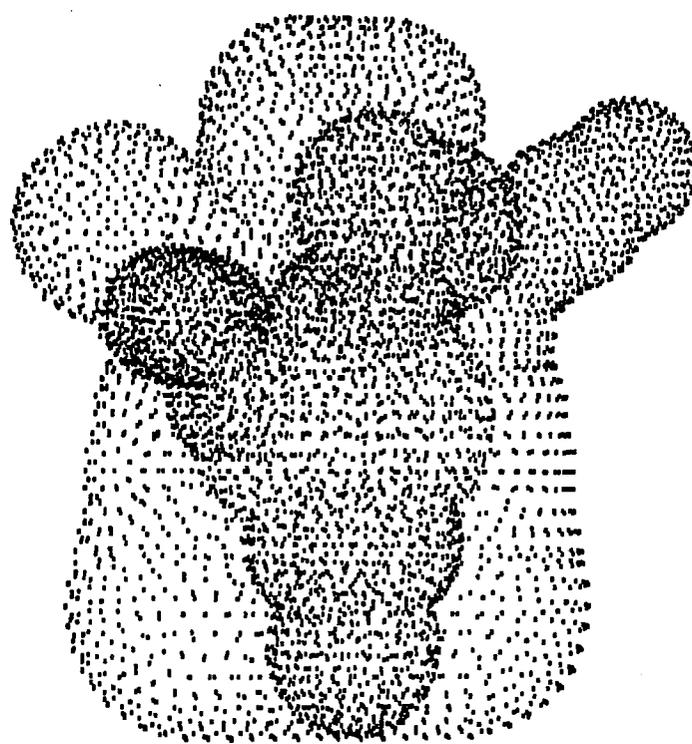


Figure 3. The small subunit is shown interfaced with the large subunit.

One of the most amazing facts about the ribosome is that it is held together with noncovalent bonds. By placing the ribosome and its parts in successively weaker ionic solutions it can be induced to dissociate without damage to the proteins or RNAs it contains. In this manner it was determined that the large subunit consists of 34 proteins, one 5S RNA (120 nucleotides), and a 23S RNA (2904 nucleotides). The small subunit contains 21 proteins and one 16S RNA (1542 nucleotides). The 23S RNA has a mass of 1.1 million daltons and the 16S RNA a mass of 0.55 million daltons. In all ribosomes, RNAs are the major components. In the cell the subunits assemble spontaneously and in exactly the proper ratios without the participation of any cofactors. The complete ribosome is formed when the small subunit binds a messenger RNA, the initial transfer RNA, and the necessary initiation factors. The key discovery which has made it possible to study the ribosome in a rational manner, were the experiments that demonstrated how the ribosome could be reassembled in the test tube (Traub et al, 1971). This breakthrough has led to the development of in vitro assays for protein synthesis of reconstituted ribosomes.

A simple listing of the parts and functions of the ribosome is not equivalent to understanding how the ribosome works and it works very strangely in comparison to protein enzymes. Protein elongation has been measured at 15 amino acids per second in bacteria and 7 amino acids per second in rabbits. No additional proteins, elongation factors, or GTP is required for very slow but accurate in vitro translation (Gutell et al., 1985). It is also possible to reconstitute working ribosomes using ribosomal proteins from different species (Higo et al., 1973). The large size of the ribosome when compared to protein enzymes and the flexibility of the active site suggest that the ribosome may present us with a unique or especially primitive example of biocatalysis.

More than half of the mass of each ribosome comes from the 16S RNA and the 23S RNA which form the core of the small and large ribosomal subunits respectively. At first

these RNAs were believed to be mainly structural and this view was reinforced by the rigid nucleotide approximation used in early DNA modeling. The discovery that certain inhibitors of translation specifically target the ribosomal RNAs (Noller et al., 1987) and the absolute necessity of the RNAs for reconstitution of the ribosome from its constituents, has coincided with our recognition of the flexibility of nucleic acids and suggest that the RNAs play a direct catalytic role in ribosomal functioning. The ribosomal version of the chicken and egg paradox stems from the idea that if all proteins are made in ribosomes and if the ribosome depends on proteins for catalysis, where did the proteins come from to make the first ribosome. The other vital parts of the translation process are all RNAs as well. The messenger RNAs that specify the order of amino acids in the protein chain may contain intervening sequences which must be removed. It is now known that in some cases the processing is self-cleaving and catalytic (Cech, 1981). The mRNA may also contain sequences which provide translation control by forming hairpins which influence the binding of the ribosome and its cofactors directly (Brierley et al., 1989). Transfer RNAs fold into a stable structure (which has been determined by x-ray crystallography) that presents the anticodon at one extreme and the corresponding amino acid at the other. Although a pure RNA ribosome, functionally independent of any proteins, has been lost in the evolutionary past, we should expect that complete understanding of ribosomal function cannot be achieved without an understanding of the role of the ribosomal RNAs.

The ribosomes of archebacteria, eubacteria, and eucaryotes are highly similar and the RNAs which form the heart of this molecular machine have large amounts of conserved primary and secondary structure. The absolute dependence on proper ribosomal function which all living organisms share, places profound constraints on the mutability of the ribosome. In vitro experiments have shown that the ribosomal molecular complex is so highly conserved that ribosomal proteins from evolutionary distant organisms can be substituted during reassembly of the ribosome and still produce a translationally competent

organelle (Higo et al., 1973). This demonstrates that it is the three dimensional shape of the proteins which is most highly conserved since the size and sequence of these proteins vary. Several of the most commonly used antibiotics, for example, puromycin, tetracycline, and streptomycin, target the bacterial ribosome and this reinforces the belief that the ribosome is a central and irreplaceable organelle. The very high conservation of ribosomal processes and its central position in the flow of genetic information suggest that it may be the oldest known constituent of life. As such this RiboNucleic Acid/Protein complex could contain fossil clues to the nature of protolife. The mechanistic schemes of this complex particle may well demonstrate themes and approaches to biochemistry which are repeated throughout the living kingdoms. The most recent phylogenetic analyses of evolutionary relationships among all life forms were developed and depend on the highly conserved nature of the ribosome. The 16S RNA of the small subunit of the ribosome is used as the basis for building universal evolutionary trees. At present the construction of these trees is based on comparison of the primary structures, but if the three dimensional structure of the ribosome were known comparison of the most conserved sites and functions could be used to extend the comparisons and better resolve genetic relationships.

Modeling Premise

Sorting out the translational process would be a lot easier if we had an X-ray crystal structure of the ribosome. As the ribosome is an extremely large, asymmetric complex of proteins and RNAs, none of which have known structures, the structure of the assembled particle will not be published in the near future. Perhaps the structure of the ribosome can be assembled from the shape of its parts. The minimum first step would be the determination the structure of the RNA of the smaller subunit which, based on size alone, should dominate the structure. The large database of structural information and the highly conserved nature of 16S ribosomal RNA makes it the prime candidate for structural study.

But 16S RNA contains almost 50,000 atoms and cannot be crystalized into a form which is analyzable by X-rays.

Paralleling the rise in the understanding of the chemical building blocks of biological systems has been the increase in the capacity and speed of computers. Ribosomal research has always been closely linked with advances in technology and it seems entirely logical that the construction and interpretation of the three dimensional shape of the ribosome will be intertwined with advances in computer technology. The modern minicomputer is now powerful enough to predict the chemistry of simple molecules from first principles and empirical approximations can be profitably applied to much larger systems. Although the computer cannot produce original independent ideas, it can do the tedious collation and cross-checking of facts and then present the results in a concise and intelligible form. Advances in graphics have made the computer into an ideal tool for the modeling of complex molecules. Central to biological catalysis is the exact fitting of the substrate bond to be altered and the functional groups of the enzyme. The mechanisms of lysozyme and hexokinase were determined from the crystal structures of the enzymes with and without bound substrates. Entire papers were devoted to explaining how they work but a few well chosen pictures can convey the same information with much greater clarity and impact. A graphical representation of the enzyme superoxide dismutase revealed not only mechanistic details but showed how a protein with an overall net negative charge could attract and bind a negatively charged substrate. Even an enzyme system like the ribosome, which doesn't appear to be as finicky as a protein enzyme, must be able to create the local, specialized chemical environment that allows living systems to perform complicated chemistry with speed and specificity and without strong chemicals or harsh conditions.

At present many of the most interesting catalytic macromolecules remain a mystery because they do not form analyzable crystals or their size prevents interpretation of their

structure. In some cases a substantial amount of structural information short of the complete molecular shape is available. Using the computer to compile these pieces of the puzzle and displaying them in a manner which can be visually integrated may facilitate the determination of the three dimensional conformation in those cases where waiting for an advance in X-ray crystallography seemed to be the only option. Computer modeling is the attempt to determine the structure, functioning and underlying rules of biology at the molecular level.

The ability to completely denature proteins and nucleic acids and then restore the correct folded structure by adjusting the pH and concentrations of salts means that all the necessary information for the folding and function of a biological molecule is contained in its primary structure. It is the ultimate goal of structural biology to predict the salient structures and mechanisms of an RNA or protein from the DNA sequence. The ideal process would move up through the levels of increasing complexity, building on the structures of the preceding stages. With the development of rapid DNA sequencing technology, the accumulation of primary structure has become routine although the problems associated with intervening sequences remain.

The prediction of secondary structure from primary sequence on a theoretical basis is partially successful. More progress has been made in determining RNA secondary structure because it has only two basic type of residues, purine and pyrimidine, and only two types of secondary structure, helix and coil. Based on thermodynamic parameters it is possible to predict the basepairing of an RNA with some consistency. When combined with the clues provided by phylogeny and secondary structure probes, the double-stranded regions of an RNA can be determined with fair accuracy. The protein secondary structure problem is more difficult because of the 20 different peptidyl residue types and the three general classes of structure, helix, sheet, and coil.

Tertiary structure prediction is still an unfulfilled dream. Theoretically nucleic acid tertiary structure should be simpler as the sequence dependent elements are hidden in the middle of the helix. A protein helix made from glycines will have a radically different external character than that of one which consists of lysines. Steady progress has been made at the atomic and molecular levels of structure thanks to the wedding of computers to X-ray crystallography. The library of structures for nucleic acid oligomers is rapidly expanding and the possible structures of polynucleotides has been greatly extended beyond that of the Watson and Crick B-DNA helix to the A-helices of DNA and RNA and even the exotic Z-helix. It is possible to predict some RNA tertiary interactions phylogenetically but as a practical matter it is more profitable to look for tertiary structure with chemical crosslinkers. Forecasting the tertiary structures of proteins is still very difficult due to the significant variation in the alpha-helical or beta-sheet structure classes. Short of a complete structure determination by X-ray crystallography or two dimensional NMR, protein tertiary structure remains a mystery.

Only the most rudimentary theories concerning the quaternary structure of a multicomponent molecular complex have been presented. In the case of the ribosome, the size of the particle is large enough to make it possible to directly probe the structure with physical chemistry techniques. A fair amount of information has been assembled which establishes semi-quantitative relationships for the positions of the proteins, for RNA/protein positioning, and for placement of the RNA and the proteins within the general outline of the ribosome. In this instance it may be possible to use information from the highest level of structure to determine the structure of a lower level.

The most basic reason for modeling is because the size or the difficulty of the problem is too great for any other approach. In modeling the ribosome there is the additional desire to understand and correlate the structures of all ribosomes, not just that of *E. coli*. Even if the crystal structure of the ribosome were available, modeling would still be

an important adjunct to such a static snapshot. The ribosome has so many important roles and embodies so many processes which occur in different systems, that the ability to reconstruct the self-assembly, regulation, and processing of the ribosome will not be explained by any one picture. The present research strategy envisions the construction of models which can be quantified and subjected to experimental testing. The goal is to construct a rational and objective model which can be reproduced, exported, compared, and used as a guide for designing new experiments, organizing and coordinating data, and expediting the understanding of the functioning of the ribosome.

References

- Brierley, I., Digard, P., and Inglis, S.C. (1989) *Cell* 57, 537-547.
- Cech, T.R., Zaug, A.J., and Grabowski, P.J. (1981) *Cell* 27, 487.
- Crick, F.H.C. (1968) *Journal of Molecular Biology* 38, 367-379.
- Gutell, Robin R., Weiser, Bryn, Woese, Carl R., and Noller, Harry F. (1985) *Progress in Nucleic Acids Research and Molecular Biology*. 32, 155-216.
- Higo, K., Held, W., Kahan, L., and Nomura, M. (1973) *Proceedings of the National Academy of Science* 70, 944-948.
- Noller, H.F., Stern, S., Moazed, D., Powers, T., Svensson, P., & Changchien, L.-M. (1987) *Cold Spring Harbor Symposia on Quantitative Biology* 52, 695-708.
- Traub, P., Mizushima, S., Lowry, C.V., & Nomura, M. (1971) In *Methods in Enzymology*, Vol. 20, pp. 391-407, Academic Press, New York.
- Yonath, A., Leonard, K.R., & Wittmann, H.G. (1987) *Science* 236, 813-816.

Chapter 2

MODELING SURVEY

Modeling Rationale

Computer modeling is the technological extension of traditional modeling and is performed for many of the same reasons. The major reason for modeling is that the size of the problem to be studied is too large to follow in detail. In chemistry, modeling is the attempt to predict which are the major, as opposed to the important but minor, physical forces which produce interesting effects such as structure or reactivity. Based on the dominating forces, a simplified replacement structure is then created and major structural motifs or foldings which are the consequence of these assumptions are predicted. A good model must be simple, elegant, and powerful. It should omit irrelevant detail and concentrate on conveying the maximum amount of significant information with minimum effort.

Linus Pauling used the basic chemical structure of the protein backbone and the supposition that compact units might form which were stabilized by intra-unit hydrogen bonds to model protein structure (Pauling et al., 1951). To justify this hypothesis, a free energy value of seven kilocalories per mole was used for the formation of a hydrogen bond. While not unreasonable for donors and acceptors in a vacuum, this value is far too large for proteins in water solution. The existence of a helical structure of specific pitch and height, the alpha-helix, was correctly predicted despite this inaccuracy. The beta-sheet structure was predicted in a similar manner. Serendipitously the first protein for which the crystal structure was solved, myoglobin, is essentially the compact folding of seven alpha-helices (Kendrew et al., 1960). The determination of the structure of myoglobin was greatly accelerated by prediction of the alpha-helix. The existence of alpha-helices and beta-sheets in many of the protein structures that have been determined validates the intuitive leap that emphasized hydrogen bonding.

The next great success in structure prediction was the modeling of the DNA double helix (Watson & Crick, 1953). The basic chemical connectivity of the polynucleotide strand had been determined and the fact that replication was semi-conservative, implying that one strand could reproduce the other, was also known. The association of guanosine with cytosine and adenosine with thymine was suspected since these bases always occurred in equal amounts. The X-ray diffraction pattern of DNA strands which suggested a helical form with a repeat of ~ 34 angstroms was the final clue. From these data and the supposition that the complementary base interaction would be based on hydrogen bonding, Watson and Crick were able to construct a physical model of the B-form DNA helix (Watson & Crick, 1953). The model expedited the cracking of the genetic code and is the seed from which molecular biology and biophysical chemistry have grown.

By ignoring the variable amino acid side groups and introducing the rigid carbonyl/peptide bond linkage, Pauling was able to reduce the dimensionality of protein folding to the level where pencil and paper modeling became possible. The ten base pair repeat of the DNA helix made it possible for Watson and Crick to construct a five centimeter per angstroms scale model of one complete turn. In the absence of similar simplifications, the huge number of variables in RNA structure will frustrate any attempts at modeling it. Even with the assumption that helical regions can be considered as rigid tubes, how are the single-stranded regions to be handled? All of the problems associated with physical modeling are magnified by the size of 16S ribosomal RNA. A physical model will be costly in terms of time and material. Any scaling ratio which attempts to preserve the space-filling characteristics of RNA will have problems with size and gravity. Complex physical models are also fairly immobile and any modifications of the structure often require a complete rebuilding process. A physical model is inherently qualitative and most often reflects the bias of the investigator who built it. Finally a physical model is passive and cannot provide the continuous consistency checking which would prevent the

construction of 'impossible' structures. Only with computer modeling is there any possibility of overcoming these shortcomings.

Moving up through the levels of structure complexity, from a given primary sequence, to a predicted secondary structure, and finally to a theoretically folded tertiary form, would be the ideal modeling process. Computer programs based on empirical thermodynamic data have achieved some success at predicting the secondary structure of small RNAs. Predicting the secondary structure of a unique RNA based solely on thermodynamic principles is still a goal, not a reliable scientific technique. By using a combination of empirically derived energy values and folding rules, researchers have been able to massage 80% of the known tRNA sequences into a cloverleaf and 60% of the 100 known 5S rRNA sequences into the 'Y' shape predicted by phylogeny (Papanicolaou et al., 1984). Improvements in secondary structure prediction make it possible to place the phylogenetically determined secondary structure among the best 10% of the possible structures (Jaeger et al., 1989). These algorithms operate by assuming that the nearest neighbor interactions and base/base stacking will predominate and that tertiary interactions are a fine tuning of the fully folded form. Sequence similarity and alignment programs similar to those used for DNA analysis are used to find the secondary structure of RNA and suggest tertiary interactions from inter-species phylogenetic evidence (Haselman et al., 1989). Phylogeny has proved to be a reliable predictor of secondary structure in those cases where we have a substantial number of sequences from diverse species.

The prediction of nucleic acid tertiary structure is proving to be far more difficult. Computers have always been used to compile and correlate nucleic acid structural data. Examples range from plots which summarize the most frequently observed atomic structure characteristics to programs which attempt to evaluate the free energy of nucleic acid conformations (Olson, 1982). Nucleic acid helices have been analyzed with computers and a new modeling rationale which emphasizes the importance of base stacking has begun to

yield interesting models for bending and kinking in double-stranded DNA (Srinivasan et al., 1987). Phylogenetic analysis of RNA sequences has been used to identify a few tertiary interactions in some cases, but three dimensional structure prediction by theoretical means is as yet impossible. Still it may be possible to create practical RNA models by combining several computer algorithms into a procedure for welding all the structural information that is available into a coherent whole.

Empirical Algorithmic Modeling

The treatment of nucleic acids as a combination of atomic orbitals would be the most exact representation of a molecule. These ab initio methods make the absolute minimum number of approximations and have been successfully used to study the characteristics of molecules as large as ammonia. Slightly larger molecules require significant amounts of CRAY time to examine only a few picoseconds of chemistry. This is clearly not a practical approach to the modeling of nucleic acids and further approximations which reduce the size of the problem will be necessary.

Empirical energy algorithms are less exact but still attempt to retain atomic resolution. The individual electrons are ignored and the molecules are analyzed on the basis of their bonding characteristics. These programs approximate the free energy of a molecule as a sum of the dominating interactions. Terms for bond lengths, bond angles, dihedral angles, electrostatics, van der Waals contacts, and hydrogen bonds are included. The bond length and bond angle terms are steep parabolic functions that depend on the square of deviation of a bond from its ideal equilibrium value ($\sum_{\text{bonds}} K_r (r - r_{\text{eq}})^2$, $\sum_{\text{angles}} K_\theta (\theta - \theta_{\text{eq}})^2$). The equilibrium bond lengths and angles for a particular atom type are derived from the crystal structures and microwave and infrared spectroscopy of many molecules. Dihedral angle energies are defined as one plus the cosine of the modular phase of the dihedral angle ($\sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)]$). The modulus coefficient determines the number of rotational minima. The manner in which the first three terms are defined means that the absolute minimum energy that could be contributed by these terms to the overall energy function is zero. If a conformation is to have a negative free energy it will result from the contributions of the other terms. The van der Waals contacts are treated as nonbonding 6-12 interactions ($\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6}$). The electrostatic free energy is the product of the partial charges of the atoms divided by a dielectric constant multiplied by the distance between them ($\frac{q_i q_j}{\epsilon R_{ij}}$). The partial charges for each atom type are calculated from quantum mechanics. Unless explicit water atoms are included the dielectric constant is usually

replaced with the value of the atomic separation to mimic the shielding of a bulk solvent. Hydrogen bonding is the weakest interaction to be included and is only calculated for those special atomic classes that are identified as hydrogen bond acceptors or donors. As a 10-12 function $\left(\sum_{\text{Hbonds}} \left[\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}} \right] \right)$, hydrogen bonding is treated as a sort of weak electrostatic interaction without any directionality. The scaling coefficients for all of the interactions are iteratively tuned so that the program will reproduce known structures. For example, variations of one angstrom in a carbon-carbon single bond would contribute 337 kilocalories to the free energy sum. In contrast, hydrogen bonds have energy wells of only 0.5 kilocalories because much of the free energy contribution of such interactions have already been accounted by the other terms.

The sum of all these terms over all atoms in the structure produces a numeric evaluation of the free energy of the structure as a whole. As all of the terms can be expressed in terms of the interatomic vectors, it is possible to determine the derivative of this analytical function. This makes it possible for the program to alter the atomic coordinates so as to minimize the free energy. The simplest minimization technique is the Newton/Rapheson method. The initial set of coordinates is used to produce a new set of coordinates which improve the ratio of the function to its derivative. The next set of coordinates equals the previous set minus the value of the function divided by the derivative of the function. This process is repeated until the the change in coordinates between steps drops below a specified target or until a preset number of steps is reached. Conjugate gradient refinement is a more sophisticated and less abrupt approach to structure minimization. This method determines the gradient of the function and moves the structure down the conformational energy hypersurface. The magnitude of the adjustment to the structure is controlled and the direction of change is coordinated with that taken in the previous step. The process can be continued until the RMS deviation between steps approaches zero or until arbitrary limits on time or number of steps have been reached. Both of these refinement techniques can become trapped in a local minimum. As the

number of local minima increases exponentially with the number of degrees of freedom, the chances of finding the global minimum approaches zero for large structures. Therefore strictly abstract modeling requires that you begin the modeling process in a conformation which closely resembles the global minimum.

Attempts at improvements in the molecular mechanics modeling approach include molecular dynamics and/or distance penalty function additions to the energy equation (Clore et al., 1986). Molecular dynamics allows the atoms of a molecule to move around in the structure in response to the energetics of its bonding environment. By evaluating the Newtonian equations of motion, the atoms are allowed to jiggle around their equilibrium values just as they do in the real world. But molecular dynamics calculations are computationally expensive and are therefore confined to conformations within a few nanoseconds of the starting structures. Simulated annealing is a related technique which, in effect, reverses the minimization process and heats up the molecule. The changes in structure that this introduces can allow portions of the molecule which are trapped in poor conformations to rearrange. Molecular dynamics is then performed to allow the system to find a new equilibrium. Further minimization may then be able to find a new local minimum. In another approach, a penalty function can be added to the empirical energy function to enforce particular atomic interactions that have been determined by physical means. But adding three dimensional constraint terms to the empirical energy equation is a distortion of the original reasoning underpinning such algorithms and will require that the scaling multipliers be readjusted. Monte Carlo methods can be used to sample conformational space in a wide-ranging and random manner, but their underlying rationale insures that a large amount of time must be expended to examine and discard worthless structures. Attempting to search a large conformational space manually is also impractical and cannot guarantee that the best structure will ever be found. Determining a good starting structure for empirical energy refinement remains the weak point of this approach.

Empirical energy programs operate in xyz-space and express all functions in terms of an atomic separation, $r = \text{atom2}(x,y,z) - \text{atom1}(x,y,z)$. As an alternative, attempts have been made to build structure sequentially by evaluating dihedral angles along the primary backbone of a protein (Vasquez & Scheraga, 1988). This transition from xyz-space to dihedral angle space results in a modeling process that resembles the attempts to find secondary structure in nucleic acids and can suffer from space/time limitations due to the large number of intermediate structures. The process could also be used as a definition for chaos since small changes in the first part of the folding process may produce completely different end results. Another approach regards DNA as a stack of planar bases of regular nucleotides, invoking yet another level of abstraction and is a distinct approach to the problem (Srinivasan et al., 1987). By summing the pairwise stacking preferences of a DNA sequence, it is possible to predict that certain double-stranded sequences will form bent or curved shapes which may explain their unusual electrophoretic gel mobilities. The success of this last method clearly demonstrates the paramount importance of base stacking over hydrogen bonding. This research also indicates a possible direction for future abstractions that might be appropriate for other levels of structure.

The original modeling protocol envisioned the transformation of a manually constructed physical model into a computer representation. This initial conformation would be evaluated and adjusted by an empirical energy protocol. Interactive graphics or molecular dynamics would then be employed to search for improved conformations. But in practice this approach has serious problems in creating structures which can only be addressed by amending the protocol.

Distance Geometry Modeling

Modeling RNA structure can be viewed as the folding of a set of rigid helical tubes that are tied together with single-stranded linkers. The flexibility of the single-stranded linkers is length dependent. Physical clues derived from the native conformation provide the modeling constraints and specify which portions of the tubes or single-strands must be where. This is similar to the Traveling Salesman Problem. The traveling salesman must visit all the cities in his territory and would prefer to find a route that connects them which minimizes his path length and energy expenditure. Such problems are members of the combinatorial optimization class. Mathematicians using graph theory have determined that this problem is a member of the NP-hard class of problems which is not solvable in a finite time (Hoffman & Wolfe, 1985). Therefore it is not possible to prove that the best solution has been found except in the simplest of cases. A theoretical solution to the RNA folding problem does not appear possible, but there is a practical method for producing reasonable structures.

Distance Geometry is an attempt to build molecular structures based on interatomic distances. From the x, y, and z coordinates of two atoms we can easily determine the distance between them as the length of the vector which connects them. Given the crystallographic coordinates of a molecule it is easy to transform the data set from three dimensional space to the distance space constructed from all of the unique interatomic distances. But given the set of distances is it possible to perform the inverse transform and reproduce a unique set of coordinates in xyz-space? Mathematicians have been unable to answer this question, even when the problems caused by chirality are excepted. This problem is further complicated in biochemical structures by the fact that substantially less than the complete set of interatomic distances is available and the distances that are known are not invariant. The situation is not as hopeless as it might seem at first glance. It is not necessary to know all the atomic separations in order to substantially restrict the number of possible molecular conformations. When the number of atoms (n) is three or four, the

number of degrees of freedom in xyz-space ($3n-6$) equals the number of interatomic distances. For larger molecules the number of interatomic distances ($n(n-1)/2$) will vastly exceed the degrees of freedom.

Distance geometry begins the process of creating a structure by building an $N \times N$ matrix of interatomic distance bounds (Crippen, 1981). The initial matrix can be seeded with arbitrary upper bounds (e.g. 1000 angstroms) and lower bounds made equal to direct van der Waals contact. The bounding values can then be adjusted to correspond to reflect the chemical structure of the molecule. The upper and lower bounds will be made equal to the same standard value for those atoms that are directly bonded. The bounds for atoms which are bonded to a common atom can also be converted into an exact distance by assuming that the variation in bond angle will be minimal. The first and last atoms of a dihedral angle do not have a fixed separation. But rotation about the central bond will establish the upper bound on the distance between these atoms in the trans conformation. The lower distance constraint will be determined by the cis conformation. Exact interatomic distance can also be determined for the atoms of a planar ring system. Atomic relationships that are based on secondary or tertiary structure are specified by the researcher and added to the bounds matrix. Once the chemical characteristics of the molecule have been converted into distance constraints, further narrowing of the separation between the upper and lower bounds can be achieved by applying the triangle inequality. If the maximum distances from atom A to atom B and from atom B to atom C are known, then maximum possible distance between A and C is equal to the sum of their separations from B. Repeated application of the triangle inequality will smooth the upper bounds and produce a more uniform, self-consistent matrix. Application of a corollary of the triangle inequality to the lower bounds completes the process of wringing every possible bit of information out of the input parameter set. The quality of the bounds matrix will depend on the total number of bounds where the upper and lower bounds are equal and the average difference between the upper and lower bounds for every interatomic distance in the structure.

Embedding the N-space structure that is described by the list of atomic separations into a three dimensional structure which satisfies all of the distance constraints is the next step. An N X N interatomic distance matrix is built on the basis of the bounds matrix. Gaps in the bounds matrix are seeded with a random distance which falls between the upper and lower bound. Now that the matrix contains a single value for every interatomic distance in the molecule it can be diagonalized. The three largest positive eigenvectors are used as the starting referents for reducing the dimensionality of the N-space structure until a three dimensional solution is obtained. This conformation is a projection of the higher dimensional structure onto three dimensions of xyz-space and the 'squashed' result of this process must be minimized against the bounds matrix. The same minimization techniques employed by empirical energy programs are used. Although there are no energy terms per se in distance geometry, there is a relationship between the distance geometry and empirical energy functions since both evaluate and minimize structures based on interatomic distances. Thus the conjugate gradient refinement of the structure can be based on the bounds violation function and the cgr error function at any point in the refinement will be proportional to the sum of squares of the bound violations ($\text{cgr error} \propto \sum_{\text{bounds}} (\text{distance}_{ij} - \text{bound})^2$). This is similar to the bond or angle terms of empirical energy minimization, lacking only the scaling constant ($\sum_{\text{bonds}} K_r (r - r_{\text{eq}})^2$). In the final analysis, a distance geometry program assesses structural deviations from the bounds matrix in terms of angstroms. It would only be necessary to plug these lengths into the appropriate expression for bond length, bond angle, etc. to obtain an empirical energy value. Comparison of superimposed distance geometry structures can be made with the usual RMS deviation which is equal to the square root of the sum of the squares of the difference vectors between identical atoms. A simpler average superposition fit error can be found by summing lengths of the difference vectors and dividing by the total number of atoms.

Distance geometry is presently used to obtain three dimensional structures from NMR data. By combining the primary structure information and the implied atomic bond distances, angles and dihedrals with the distance constraints derived from the through bond (COSY) and through space (NOESY) transference of NMR energy, the structures of several oligonucleotides and proteins have been solved (Wuethrich, 1989). As the distance dependence of NMR data is so severe ($1/r^6$) and the NMR spectrum so crowded, this approach is most applicable to molecules which are small when compared to the large biologically active molecules and molecular complexes. The comparison of protein structures produced with distance geometry from NMR data with those found by X-ray crystallography has established that this as a reliable technique which can be superior to crystallography in the study of solution or dynamic structures (Heidorn & Trewella, 1988).

Distance geometry was used to find a configuration for the proteins of the small subunit of the ribosome from a minimal data set (Kuntz & Crippen, 1980). Although this research did extend distance geometry into much larger physical dimensions with much longer distance constraints, it attempted to model the conformation of only 21 objects. Adapting DG to fold 16S RNA will be a unique attempt to apply this technique to the much larger realm of global folding patterns. It will require the inclusion of distance constraints of varying quality from different distance domains. The hardest problem in adapting this algorithm to the modeling program may be to devise a reduction scheme that will allow the program to run on the available computer resources.

Graphical Modeling

Chemistry is a three dimensional process in which size and shape play a vital role in reaction rate and specificity. Most biochemistry takes place through directed contact, as opposed to the mass action principles of benchtop solution chemistry. To avoid low reaction cross sections and low reaction rates, the exact alignment of functional groups and creation of the proper microenvironment is absolutely essential. Large portions of the human brain are dedicated to processing visual information and interactive graphics workstations attempt to utilize this innate wetware. A single, well-chosen representation can expedite the comprehension of a chemical process by reminding us of a similar process in the macroworld. Docking a substrate in an enzyme active site may be as easy as catching a fly ball. Combining innate visual judgement with the scientific training of the chemist concerning reactive moieties is the underlying rationale of intelligent drug design. Advances in computer graphics have made the synergistic combination of human and computer abilities an attainable goal.

The computer is the interpreter who stands between the raw experimental data and the human observer. The modern digital computer is an electronic creature with unique capabilities and limitations. A bit is the most basic computer element. It is a binary, on/off data switch and when grouped into a set of eight it forms a byte. The present standard encoding scheme uses one byte for every character of a text file. The byte is the most common measure of computer memory and magnetic disk storage. Engineering and manufacturing advances have made it possible to equip each computer with capacities in the millions of bytes (MByte) and in so doing, have made it possible to apply the computer to molecular modeling. Of course the increase in computer power has made it possible to tackle much larger problems, but the major improvement that faster, smarter computers allow is in visualization of the results. Using computers to show graphically what is unseen or hard to see brings human visual skills into the modeling process. Vision is an instinctive

ability and it is language and phrasing independent. Therefore computer graphics will be applicable in many areas and will improve scientific communication.

A pixel (PIXture ELEment) is the smallest addressable dot on the computer screen. The resolution of a video display is equal to its number of distinguishable pixels. An American TV has a resolution of 512 pixels horizontally and 432 pixels vertically. A video drawing is made from a collection of finite, discrete dots and relies on the human eye to integrate the dots into lines and shapes. The jagged stairstep effect that appears most often in diagonal lines, is the result when the finite resolution of the screen inadequately mimics the continuous lines of the real world. There are three types of video terminals and graphics workstations and some microcomputers have been designed to emulate any of these types of terminals. The most common is the ordinary computer terminal, which can only display alphanumeric characters in a fixed format. Vector graphic display terminals draw only the points and lines that are specified by computer command. This approach keeps the amount of information which the computer must calculate and update at a minimum. To draw a line which spans the screen and may contain hundreds of pixels requires only the start and end points. This makes real time manipulation of a line image practical. Since the actual line is drawn, the jaggies are minimized. Raster graphic displays use the same approach as convention television. The value of every pixel is specified for every view displayed. Creating a single black and white image requires the calculation and transfer of 27,648 bytes of information to fill one TV screen. A color raster picture may require more than a MegaByte of data. Most computer raster displays have a higher resolution than TV but even at these higher pixel densities, the jaggies can be a serious problem unless computationally expensive compensation is performed. These large data requirements make interactive raster graphics impractical.

Despite their computational overhead and limitations, raster graphics displays are becoming more common because of their ability to produce pictures that appear three dimensional. Several techniques are used to engage the human sense of depth perception.

Depth cueing includes decreasing the brightness and size of an object with increasing distance from the viewer. Objects can be made to appear nearer or further from the viewer by interposing one in front of the other or by using a geometric perspective. Distance from an unseen light source can be simulated with shading or shadows. A specular highlight which appears to be a spot reflection of the light source can also serve to orient the viewer and increase the sense of depth. When stereo views are created by making separate pictures for the left and right eye that have been rotated slightly relative to one another, the effect can be startlingly effective. With the latest in ray-tracing computer programs, pictures can be created that are difficult to distinguish from photographs of real objects (Pool, 1989).

At even the earliest stages of computer modeling one is faced with the difficult decision as to which representation of the molecules being studied should be used. A vector line drawing which traces the main backbone is the simplest and will therefore be the easiest to manipulate. Unfortunately this will ignore many of the most important interactions and can give one a false sense of accessibility and mobility. In a prime example, hexokinase, a backbone representation or even a secondary structure model will fail to reveal a very obvious active site and reaction mechanism. A crude surface model which ignores hydrogens and is constructed from a simple union of van der Waals radii spheres, is vastly superior. Comparison of these kinds of displays of hexokinase with and without bound substrate, reveals in a single glance how the enzyme is able to exclude water from the active site with an tight molecular fit.

Since hydrogen bonding plays a vital role in nucleic acid structure, any model which ignores hydrogens must be somewhat inadequate. The presence of aromatic bases in every nucleotide also guarantees that the molecules cannot be accurately modeled by simple spheres as the electronic distributions of these planar groups will be highly asymmetric. The earliest molecular surface representations considered the effective molecular surface to be the center of a spherical solvent molecule which was rolled across the van der Waals contact surface of the atoms (Lee & Richards, 1971). This is an inappropriate approach

because the hydrophobic bases in the center of a nucleic acid helix are not uniformly solvated. The most accurate surface representation is created by the Analytical Molecular Surface (AMS) program (Connolly, 1983). It constructs the surface as a union of the polygons formed by the contact surface of a probe sphere, usually a solvent molecule, and the covalent radius of a molecule's atoms. The advantage of this approach is that it reveals the shape and docking surface of a molecule which may interact with a substrate or probe so strongly as to exclude solvent. It also fills in the gaps between atoms which are too small to admit the probe. Comparing the surface of an oligonucleotide duplex made from a simple union of van der Waals spheres and that produced by AMS demonstrates the differences. The van der Waals surface is easier to produce and can be done independently of a mainframe computer by a graphics workstation with the proper hardware. The AMS surface requires the user to assemble several input files and determine the correct view mathematically. The calculation and display may then take the majority of an hour of computer time for even a small oligonucleotide. The van der Waals surface appears bulky and knobby with gaps between the stacked bases. The AMS surface is smoother, more compact and solid. The functional groups in the major and minor grooves are more accessible and distinct patches of color indicating a common chemical nature can be discerned. The van der Waals surface closely resembles the physical constructs made with CPK models and will therefore be more familiar to the scientist. The AMS surface displays the molecule which a nucleic acid binding protein would 'feel' as it searched for its binding site.

Of course either of these representations would be inappropriate in some instances. For example, an NMR spectroscopist might prefer a stick figure which would allow him to see the hydrogen bonding and sugar puckering of the nucleotides. Color is a particularly useful tool for emphasizing particular portions of the complex picture formed by even a short oligonucleotide. Simple two color schemes can be used to distinguish between the phosphate/sugar backbone and the central base stack or the charge variations along a

polynucleotide. The selection of the visual representation to be used should be based on conveying the maximum amount of information without obscuring the important structural features. It is also possible to overload a graphic with so much data that interpreting the picture would require yet another research project. The ability to produce images which vary in responsiveness and detail will make it possible to display the modeling results to best effect.

The Unified Protocol

In the early stages of 16S RNA modeling, a crude physical model was constructed and converted into computer coordinates. Straightforward attempts to improve the model through interactive graphics modeling proved to be beyond the capabilities of the available computing resources. An evaluation of empirical energy modeling showed that it lacked the necessary discrimination. Distance geometry can be used to objectively fold small molecules, but will require the use of special modeling constructs. Combining these various approaches will produce a new modeling protocol.

The original research plan was to retain as much of the full representation as possible and make simplifying structure reductions in a stepwise fashion only when the modeling process became unwieldy. The secondary structure posited by Harry Noller and coworkers would be used as the basis for forming helical subunits. As the cleavage data indicate that these regions are persistent and stable, the flat, parallel drawings of the secondary structure shown in most papers would be transformed into the RNA A-form helices. When the numbers of atoms becomes unwieldy, it is these regions which would be replaced with group representations. A careful substitution preserving the phosphate backbone and hydrogen bonding pattern would make it possible to easily reinsert all the atoms. These helical regions would then be linked together with the appropriate single strands. Finally the entire molecule would be assembled with the general relationships of the domains to be taken from the crude physical model. This final structure can then be minimized at a level of abstraction which is practical for the computer resources available. The minimizations would be done with a series of empirical energy programs. The results of these analyses and the three dimensional structure of the molecule would then be displayed in stereo for visual integration and analysis by the researcher.

But attempting to find the global shape by hand is tedious and is not certain to search all of conformational space. It is also a highly subjective process that can be tainted by poorly defined relationships and hunches. An objective search for global conformation

can be done with distance geometry, constrained molecular dynamics, or dihedral angle search. The latter can be ruled out as not well adapted to nucleic acids which have far too many degrees of freedom. Using empirical energy dynamics on a molecule the size of 16S RNA is not practical, especially considering the quality of computing available. Additionally the introduction of pseudoatoms into the empirical energy routines would be very time consuming and require additional assumptions. The recent development of distance geometry algorithms has made it possible to objectively produce folded structures. In effect this new approach makes it possible to search conformational space broadly while energy minimization can be used to find the best 'local' conformation. Pseudoatoms can be simply meshed with distance geometry as this algorithm automatically abstracts the problem into distance space and really does not deal with molecules. Distance geometry is widely used to determine the structure of molecules from NMR experiments. NMR data is very short range in nature and the accumulation of errors throughout the modeling process can lead to structure variations which exceeds the accuracy of the data. Applying long range data to the structure determination averts the potentially chaotic process of building up structure from local considerations alone.

The new computer modeling protocol has been designed to incorporate various forms of data, several computer programs, and human judgement into a coordinated method for constructing three dimensional models of RNA. This protocol will allow the large RNAs, which play so vital a role in living organisms, to be modeled. In the developmental stages transfer RNA is used as an authentic sample. Based on the phylogenetically determined hydrogen bonding pattern which forms the secondary structure of an RNA, A-form helices are introduced in these double-stranded regions as the first step in forming a three dimensional structure. Proceeding beyond this step, the traditional modeling problems of subjectivity are encountered. Distance geometry is introduced to fold the molecule objectively into a fully three dimensional conformation. Empirical energy modeling programs can then be used to make sure that the basic rules that have been

deduced from smaller molecules are not violated. Appropriate graphical representations are used to facilitate the comparison with other data sets and models. As additional data becomes available or existing data is further refined, the process can be repeated. Relationships which cannot be quantified may be introduced into the model through interactive graphical manipulation.

References

- Clore, G.M., Bruenger, A.T., Karplus, M., & Gronenborn, A.M. (1986) *Journal of Molecular Biology* 191, 523-551.
- Connolly, M.L. (1983) *Journal of Applied Crystallography* 16, 548-558.
- Crippen, G. (1981) In *Distance Geometry and Conformational Calculations*, Chemometrics Research Studies Series, (Bawden, D., ed.) pp 1-58, Research Studies Press, New York.
- Haselman, T., Camp, D.G., & Fox, G.E. (1989) *Nucleic Acids Research* 17, 2215-2221.
- Heidorn, D.B. & Trehwella, J. (1988) *Biochemistry* 27, 909-915.
- Hoffman, A.J., & Wolfe, P. (1985) In *The Traveling Salesman Problem* (Lawler, E.L., Lenstra, J.K., Rinnooy Kan, A.H.G., and Shmoys, D.B. eds.) pp 1-15, Wiley & Sons, New York.
- Jaeger, J., Turner, D.H., & Zuker, M. (1989) *Proceedings of the National Academy of Science* 86, 7706-7710.
- Kendrew, J.C., Dickerson, R.E., Strandberg, B.E., Hart, R.G., Davies, D.R., Phillips, D.C., & Shore, V.C. (1960) *Nature* 185, 422-427.
- Kuntz, I.D. & Crippen, G.M. (1980) *Biophysical Journal* 32, 677-695.
- Lee, B. & Richards, F.M. (1971) *Journal of Molecular Biology* 55, 379-400.
- Olson, W.K. (1982) In *Topics in Nucleic Acid Structures: Part 2* (Neidle, S., ed.), pp 1-79, Macmillan Press, London.
- Papanicolaou, C., Gouy, M., & Ninio, J. (1984) *Nucleic Acids Research* 12, 31-44.
- Pauling, L., Corey, R.B., & Branson, H.R. (1951) *Proceedings of the National Academy of Science* 37, 205-211.
- Pool, R. (1989) *Science* 244, 1438-1440.
- Srinivasan, A.R., Torres, R., Clark, W., & Olson, W.K. (1987) *Journal of Biomolecular Structure & Dynamics* 3, 459-496.

Vasquez, M. & Scheraga, H.A. (1988) *Journal of Biomolecular Structure & Dynamics* 5, 705-756.

Watson, J.D. & Crick, F.H.C. (1953) *Nature* 171, 737-738.

Wuethrich, K. (1989) *Science* 243, 45-50.

Chapter 3

DNA OLIGONUCLEOTIDE MODELING

Introduction

Hydrogen bonding came to be seen as the underlying driving force in the folding of proteins and nucleic acids. The successful prediction of the protein alpha-helix and the DNA B-form helix based on hydrogen bonding, caused it to be overrated. This problem is compounded by the role of hydrogen bonding plays in the genetic code. But the importance of hydrogen bonding in the transmission of genetic information does not imply that it is the driving force in helix formation. In fact the interactions of the individual bases is dominated by hydrogen bonding only in nonpolar solutions. In aqueous solution bases interact by stacking, while the hydrogen bonding groups interact with the solvent. Physical chemistry studies have confirmed the obvious, that intramolecular hydrogen bonding in a water solution is far too weak a force to be responsible for the folding process. Rather hydrogen bonds are vital to the maintenance of a stable structure where the interior is commonly hydrophobic. It has become clear that the entropically favorable segregation of nonpolar moieties away from water is the main driving force in the formation of folded protein and nucleic acid structures. This view is confirmed by recent computational explorations which indicate that nucleic acid helix formation and local structure variation are a direct outgrowth of the sequence dependent base stacking (Haran & Crothers, 1989). The crystal structure of transfer RNA reflects this as well since 71 of 76 bases are stacked and two of the nonstacked bases are nonaromatic modified bases, while only 42 bases are involved in Watson-Crick basepairing (Saenger, 1984). The character of the folding forces is further reflected in the differing natures of protein and nucleic acid helices. Naturally occurring amino acids have 20 different functional side chains which vary in size, charge, and hydrophobicity. In the formation of a regular helix these groups are placed on the outside of the helix while the central helical core is stabilized by interresidue backbone hydrogen bonding. In nucleic acids both the purine and pyrimidine bases are highly aromatic groups which minimize contact with the aqueous solution by stacking in the interior of the helix. Thus each protein helix will have a highly sequence-dependent exterior character, while nucleic acid helices will be much more uniform, particularly in the short helical segments of less than a full turn that are found in RNA.

The DNA oligomer, CGCGAATTCGCG, contains the recognition site for the restriction enzyme EcoRI and its structure was determined by X-ray crystallography (Wing

et al., 1980). This oligomer forms a base-paired duplex and crystallizes as a B-form helix. As a clear and direct confirmation of the structure predicted by Watson and Crick, the Dickerson dodecamer has been extensively analyzed and is the basis for predicting the local variation in DNA structure embodied by the Calladine Rules (Calladine, 1982). The concept of continuously polymorphic DNA is in contrast to the completely uniform structures constructed from the X-ray diffraction data (Arnott et al., 1973). Of course the diffraction data reflects the average conformation of large quantities of DNA fibers, while the oligomer crystal reflects the structure of a specific sequence under very different solution conditions. Although a crystal may contain upwards of 70% water, the conditions therein diverge greatly from that of a simple solution and even more so from that which exists inside a living cell. Two dimensional NMR studies can probe the solution behavior of oligonucleotides and computer modeling can extrapolate from known data to give us a sense of the behavior of a variety of oligonucleotides under other conditions. But each step away from the atomic resolution of an X-ray crystal structure compounds the uncertainty of the results. Therefore many researchers have studied systems that are very similar to the Dickerson dodecamer. Questions about the variation in DNA structure caused by mismatches and hairpins lend themselves quite naturally to this sort of approach and data concerning the relative stability of these various forms will be vital in prediction of the secondary structure of nucleic acids (White & Draper, 1989).

By replacing the adenosine residues of the Dickerson dodecamer with thymines, an oligonucleotide is created which should form a hairpin in preference to a regular duplex. Two dimensional NMR and distance geometry have been used to determine the structure of this molecule in solution (Hare & Reid, 1986). As the eventual goal is to develop RNA models based on helical substructures, modeling this oligonucleotide should be a good test of the protocol which combines manual model construction with empirical energy minimization. Both a duplex bulge-loop molecule and a hairpin based on regular B-form helices were built for comparison. All three structures vary greatly from that of the X-ray structure of a closely related oligonucleotide (Chattopadhyaya et al., 1988). This initial modeling study indicates that the results are highly dependent on the experimental conditions and that improvements in the protocol will be necessary.

Materials and Methods

Materials

Software

The molecular modeling package of programs, AMBER circa 1984, was obtained from Peter Kollman at UCSF. The interactive graphics modeling program, GRAMPS, written by T.J. O'Donnell for the National Resource of computation in Chemistry at LBL, was used for vector display. Molecular surface calculations and raster pictures were generated with the modeling programs written by Mike Connolly as distributed by the Scripps Research Institute in San Diego. The program to convert the raster images into black and white drawings was written by the author in FORTRAN.

Hardware

The calculations were performed on a VAX 11/780 equipped with four MBytes of main memory. An Evans & Sutherland MultiPicture System connected to the VAX was used for interactive graphics. Raster graphics were displayed on a VT240 DEC terminal.

Structural Data

The Dickerson dodecamer, CGCGAATTCGCG, can basepair with another nucleotide of the same sequence to form a completely hydrogen bonded, double-stranded helix. With the development of the automated DNA synthesizer, it became possible to generate oligonucleotides of a predefined sequence in sufficient quantity and purity to begin to study the conformations of nucleic acids in solution with NMR. When the central adenosines of the Dickerson dodecamer are replaced with thymines, an irregularity is introduced into the helix. The hairpin that is formed when the 5' and 3' ends of the sequence form intramolecular basepairs has as many Watson/Crick hydrogen bonds as a duplex formed by two of these nucleotides. The backbone strain that is required to bend the unpaired thymines of the hairpin into a loop is compensated for entropically, by the greater ease with which the bases to be paired can find each other. The structure of this hairpin was determined by distance geometry from two dimensional NMR data (Hare & Reid, 1986).

The coordinates of the DNA hairpin were obtained from Brian Reid via David Wemmer. The parameters for creating standard DNA helices are included in the NUCGEN module of AMBER and are derived from X-ray diffraction by DNA fibers.

Methods

The B-form DNA helix was created from Arnott parameters with the NUCGEN module of the AMBER distribution. The computer version of the hairpin was created from B-form CGCG DNA oligonucleotide built by NUCGEN. The end of the helix was spanned by calculating phosphate and C4' positions in a crude arc from the 5' to 3' ends of the helix such that the phosphate to phosphate distance was approximately 6 angstroms. AMBER was then used to reinsert the all atom representation of these thymine residues based on its standard conformation library.

Structure minimization with AMBER was allowed to proceed until the RMS change in a structure from one step to the next was less than 0.1 angstroms.

Analytic molecular surface areas and volumes were calculated with the AMS and VAM modules of the Connolly programs respectively. The van der Waals surface areas were calculated by the Molecular Surface program written by Mike Connolly. The van der Waals volumes are calculated by the ANALYSIS module of the AMBER programs.

Line drawings are copies of MPS vectors displays created by GRAMPS on a Tektronix thermal printer. The stick figure raster drawings were created with the ball-and-stick option of the AMS program written by Mike Connolly. The raster displays of the structures created by the RAMS Connolly program were converted into black and white figures on a VT240 terminal with REGIS graphics. Each figure was directly output to a dot matrix printer attached to the terminal.

Results

The NMR hairpin has strong stacking of all the bases except for the thymine in position six which is swung out of the loop into solution (fig. 1). Although the hairpin stem shows some helical twisting, it is of a much more gentle and extended character than that of the regular B-form helix. It has been suggested that this more extended structure is an artifact of the distance geometry program and may not accurately reflect the true solution structure (Metzler et al., 1989). Regardless, AMBER finds no major conflict between its structural database and the conformation of the molecule as determined by NMR. There are some minor adjustments to the stacking of the thymines in the loop and some fraying of the C1:G12 terminal basepair. Although the AMBER minimization improved the free energy of the structure by some 60 kilocalories, the RMS difference between the initial and final structures is small. The fact that these changes required the least amount of computer time and cycles of the three structures indicates that the alterations were facile. The NMR structure coordinates had already been minimized with respect to the geometry of template nucleotides derived from small molecule X-ray crystallography. AMBER just found a slightly different local energy minimum which reflects the geometries from which its empirical coefficients were derived.

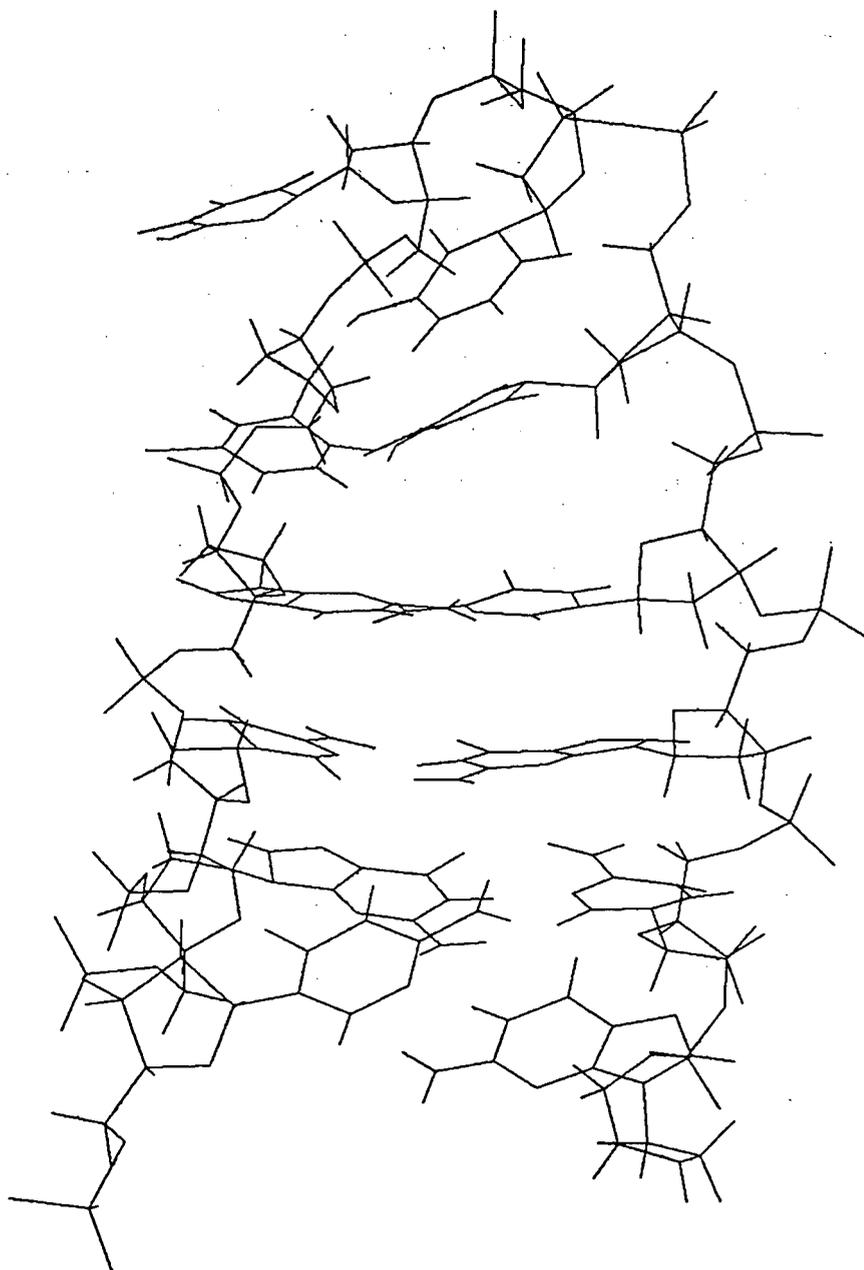


Figure 1. The DNA oligonucleotide hairpin as determined from NMR data and modified by AMBER minimization.

There were no major changes made in the structure of the DNA duplex (fig. 2). Minor adjustments in sugar puckering are made and the bases are tilted and rolled in a manner suggestive of the Calladine rules. The difference between the initial free energy of 4.4 kilocalories and the final value of -0.974 kilocalories reflects the similarity between the structural parameters derived from fiber diffraction and the sources of the energy coefficients for AMBER. Since the AMBER parameters have been tuned to reproduce the standard DNA geometries any other result would indicate a mistake in running the program. When comparing the results for the duplex to the other structures it must be remembered that the values for the duplex are for two strands of DNA. Thus the energy per oligonucleotide for the duplex falls between the NMR and the computer generated hairpin.

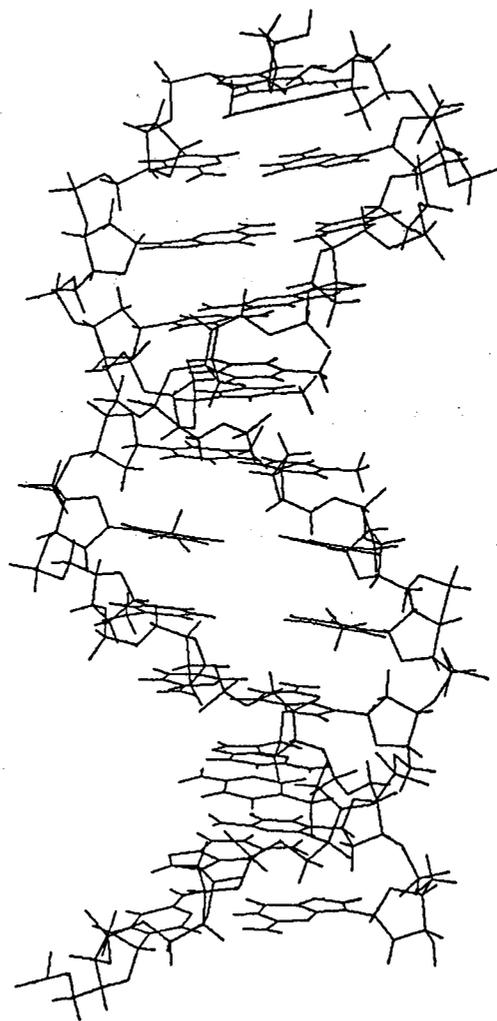


Figure 2. Double-strand DNA duplex with four basepair mismatch created and minimized with AMBER.

The crowding of the bases in the loop of AMBER generated hairpin caused by the ad hoc manner in which the 5' and 3' sides of the CGCG basepairs were added is the cause of its substantially larger initial free energy (fig. 3). AMBER relieves this van der Waals crowding by rotation of the thymine in the sixth position about the phosphate backbone until the base is completely clear of the hairpin. The overall size of the loop is not increased. A human modeler might adjust the bonds of the entire phosphate backbone of the loop region in order to maintain base stacking. AMBER chooses to alter the less energetic dihedral rotation about a single bond, than the more complex series of adjustments that would be required to lengthen the phosphate backbone of the entire loop. Even though the computer chose the path of least resistance it still took significantly longer to minimize the hairpin structure. The smaller RMS change for the structure as a whole disguises the radical repositioning of the thymine in the sixth position. It is interesting that AMBER chose to alter the same residue that is unstacked in the NMR structure. This could be the result of the manner in which the computer generated hairpin was built. It would be necessary to repeat the experiment several times to discover if there is an underlying structural reason for choosing this nucleotide.

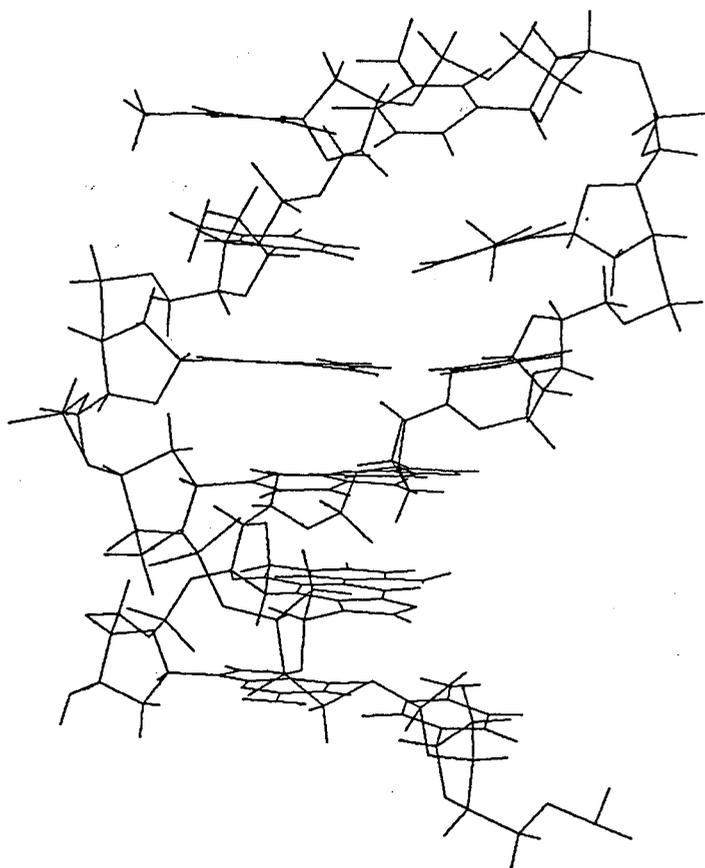


Figure 3. DNA hairpin constructed by arbitrarily bridging a four basepair B-form helix.

When the three structures are placed next to each other in the same scale, it is apparent that they have very different conformations (fig. 4). Despite these obvious differences, AMBER finds these structures to be roughly equivalent when judged on the basis of minimized free energy. The radius of gyration may not be a particularly sensitive indicator of the differences but it does much better than the free energy. If the surface area or volume of the molecules could be measured directly in solution it would be possible to tell whether the hairpin or duplex was present but it would be much more difficult to distinguish the two hairpin forms on either of these bases.

	<u>Summary of Results</u>		
	NMR	Duplex	Hairpin
AMBER run (#steps)	608	810	1103
CPU time (hours)	1:41	5:11	2:50
RMS (angstroms)	1.43	0078	1.01
G (kcal) initial	59.8	4.4	260.5
final	-0.494	-0.974	-0.444
Rg (angstroms) initial	8.77	13.41	9.27
final	8.66	13.63	9.43
Surface area (vdW)	1859	2859	1886
in a**2 (Connolly)	640	1283	654
Volume (vdW) initial	2829	10102	3340
in a**3 final	2716	10599	3509
Volume (Connolly)	3160	6563	3211

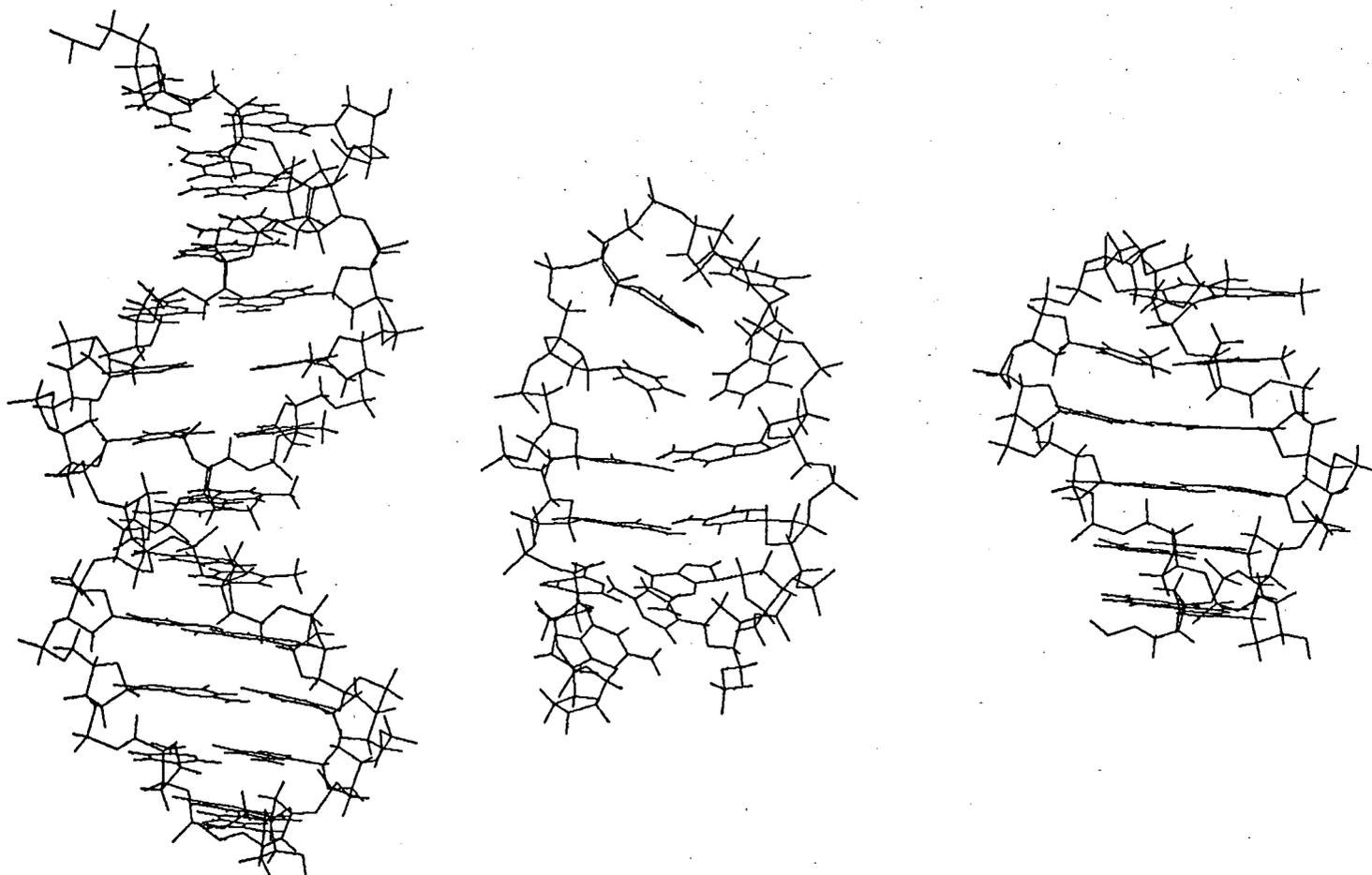


Figure 4. The DNA duplex, NMR hairpin, and computer generated hairpin to the same scale.

The differences between the van der Waals and molecular surface representations are dramatic. The surface areas of all three structures is almost a third less when the actual contact surface is used. The differences in the volume calculations are not as uniform and are much more informative. The volume of the tightly wound computer generated hairpin is essentially the same in both representations. The calculated volume for the NMR hairpin is actually larger when the direct molecular contact surface of the molecule is used instead of a simple union of van der Waals spheres. The reason for this can be seen in the rotational series of stick figures (fig. 5). The NMR hairpin in the bottom row can be seen to be a relatively flat figure when turned edge on. When rotated so that the molecule is at its widest it appears that there is significant empty space between the bases. The van der Waals surface of this molecule does appear to have gaps which would allow particles to pass through but the molecular contact surface created by rolling a water molecule sized sphere over the structure is a solid with no gaps. The volume of the duplex as determined by molecular contact is approximately twice that of either hairpins. As the duplex would be similar to two hairpins stacked on one another this makes good sense. The van der Waals volume of the duplex is more than three times the size of the van der Waals volume of the NMR hairpin. As can be seen in the rotational series of the duplex, this is due to the thymines in the middle of the structure. Although the van der Waals spheres may fill the interior of this regular duplex, the gap between the thymines which can be seen is not an illusion. The pyrimidines are significantly shorter than the purines to which they should be basepaired. The molecular contact surface algorithm recognizes this fact and is able to roll the probe sphere through the duplex. Therefore the van der Waals surface of a molecule must be used with great caution as it may produce representations that are both larger and smaller than they should be.

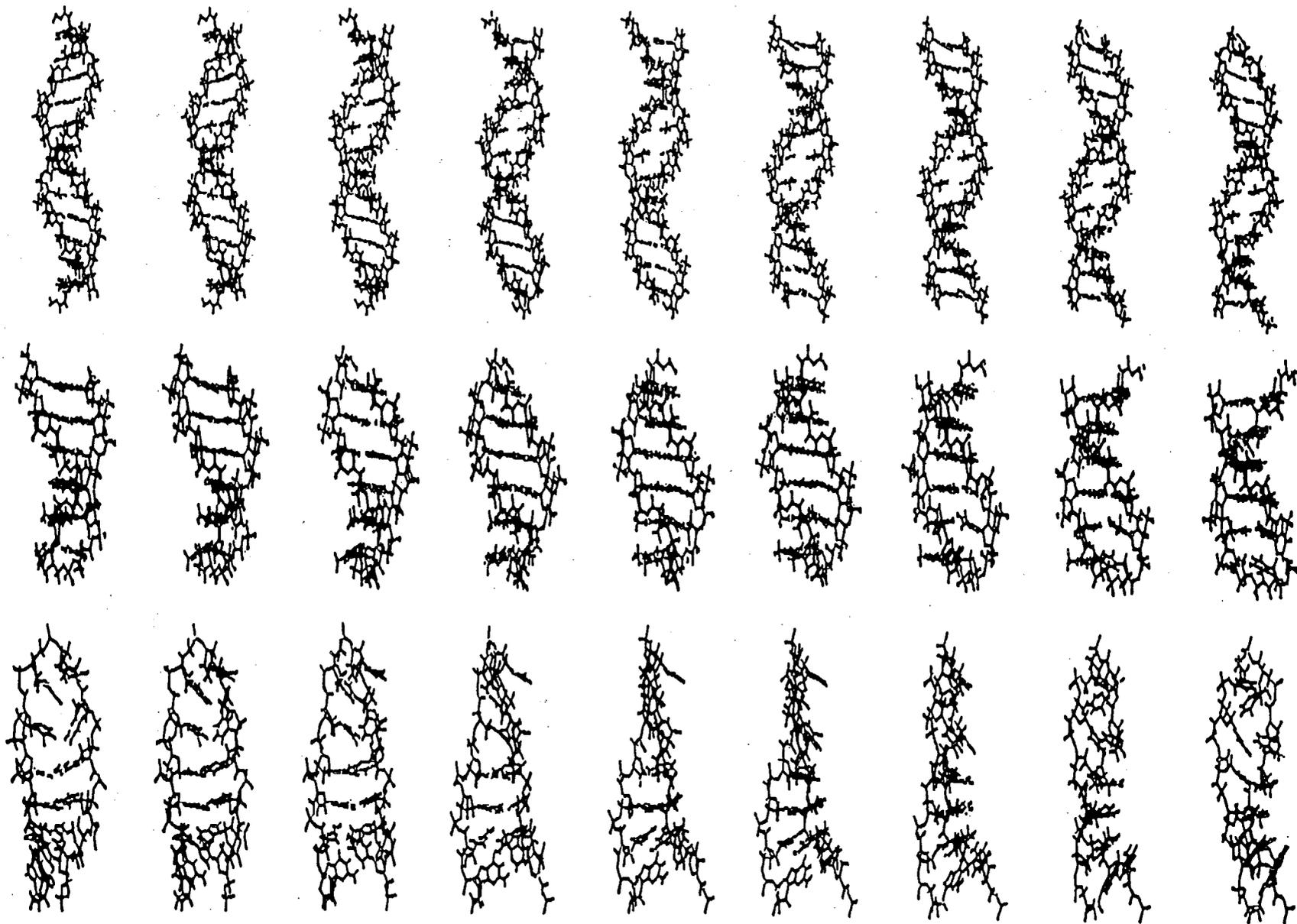


Figure 5. The DNA duplex (top), NMR hairpin (middle), and computer generated (bottom) hairpins are shown at 30 degree rotational increments. The molecules are not identically scaled.

Discussion

The continuing study of DNA oligomers has confirmed that these molecules have a wide range of possible conformations and that the structure formed depends upon the molecular environment. In one case an oligonucleotide was crystallized as an A-form helix while NMR indicated that it formed a B-form helix in solution (Reid et al., 1983). A very recent NMR study of the hairpin formed by the sequence, CGCGTTGTTTCGCG, was used as the basis for the construction of an energy minimized structure (Williamson and Boxer, 1989). The structure was built as a B-form stem of four G:C basepairs with a five base loop and satisfies all the NMR distance constraints. It very strongly resembles the computer generated hairpin constructed for this study. The crystal structure of the oligonucleotide, CGCGCGTTTTTCGCGCG, has also been published (Chattopadhyaya et al., 1988) It differs from the hairpin studied here only in having a G:C basepair added to the stem. Yet the structure of this molecule has been determined to be a Z-form helix. The extremely high salt conditions that were necessary to form crystals are known to favor Z-form DNA. Two of the four thymines in the loop are forced out into solution by the tighter Z-form helix of the stem. The bases are stacked on bases from a symmetry related hairpin. Earlier NMR studies of the same hairpin under lower salt conditions did not indicate any unusual Z structure (Ikuta et al., 1986). Clearly crystal structures may not reflect biologically important conformations if they are formed in exotic conditions. As the parameters for many of the modeling and minimization programs rely heavily on crystallographic data this is a cause for concern.

The results of all these studies illustrate what those involved in wet chemistry have always known but that theoreticians sometimes forget, there are a large number of structures available to nucleic acid polymers and they coexist in a dynamic equilibrium. That one particular conformation dominates to the extent that it is the only one detectable by NMR or crystallography, does not exclude other structures which may be distant in conformational space. Consequently the ability of biological systems to distinguish

different structures must be based on direct three dimensional complementarity and not the free energy (i.e. stability) of the target molecule.

The AMBER results demonstrate that empirical energy algorithms will have difficulty in choosing or converting between nucleic acid conformations. Physical characteristics like the radius of gyration may indicate if a structure is radically wrong but most structures will appear to be acceptable. The bulge duplex appears to be an energetically comparable conformation because the distance dependent dielectric option of AMBER, which avoids the use of explicit water molecules, does not introduce a penalty for the 'vacuum' which exists between the mismatched thymines in the middle of the structure. Introduction of water molecules into these spaces would disrupt the helical base stacking and destabilize the duplex just as in real solutions. As empirical energy calculations are unable to detect large differences which are easy to see, attempts to demonstrate a molecular preference for a particular ribose ring pucker or similar details based solely on such calculations are guesses. A defensible modeling approach is to fix all the sugars to that found in the model B-DNA structure. It can be argued that since the base/base interactions are the dominant feature, individual sugar pucker can be ignored so long as the overall B-form geometry is maintained. As the energy barrier for pseudorotation of the sugar ring is approximately one kilocalorie, nucleic acids will be rapidly alternating between fairly equivalent structures: AMBER does fray the ends as seen in NMR work but the minimization process makes noticeable changes only in those basepairs at the end of the helix. The same results could be obtained with an octamer or with a longer nucleotide and this suggests that reduction schemes based on helical subunits will be possible.

As the answer obtained by energy minimization is very dependent on the initial starting conformation, manual modeling can introduce an important random variable. Unfortunately descriptions of the construction process similar to 'the oligonucleotide was adjusted to get good basestacking interactions and to avoid bad contacts with bases' are far too vague and open to interpretation. Such structures can be difficult to reproduce even for

the same researcher. Additionally there is no clear consensus on something as common as the conformation of a nucleic acid loop. Energetically 5', 3', and centrally stacked bases have equivalent structures (Haasnoot et al., 1985). Most researchers who attempt to use manually constructed models compensate for this by doing multiple runs from similar but distinct starting structures. Searching conformational space in this manner is an onerous task which is possible only for the smallest molecules and may not be very informative about the dominant conformational variant.

In addition to the theoretical problems associated with empirical energy modeling, some ominous practical limitations began to appear as well. In the computer run times it took a proportionally longer time to perform a larger number of minimization steps for the hairpin structures. The duplex molecule required an intermediate number of minimization cycles to reach convergence but at a much greater cost in computer time. The duplex has exactly twice as many atoms as a hairpin but the time used seems to have increased as the square of the number of atoms. The amount of time required for a large RNA may be cosmological. Similar time and size problems were encountered with AMS in the generation of molecular surfaces. Unfortunately this program is not easily expanded and the more accurate representations that it produces will have to be foregone when dealing with larger molecules.

The looming problems of size, the inability of AMBER to discriminate between nonequivalent structures, and the certainty that the final structure will be almost identical to the initial structure, mean that the modeling protocol must be improved if the structure of 16S RNA is to be attempted.

References

- Arnott, S., Hukins, D.W.L., Dover, S.D., Fuller, W., & Hudgson, A.R. (1973) *Journal of Molecular biology* 81, 107-122.
- Calladine, C.R. (1982) *Journal of Molecular Biology* 161, 343-352.
- Chattopadhyaya, R., Ikuta, S., Grzeskowiak, K., and Dickerson, R.E. *Nature* (1988) 334, 175-179.
- Haasnoot, C.A.G., DeBruin, S.H., Hilbers, C.W., van der Marel, G.A., and van Boom, J.H. (1985) *Proc. Int. Symp. Biomol. Struct. Interactions, Suppl. J. Biosci.* 8, 767-780.
- Haran, T.E. and Crothers, D.M. *Biochemistry* (1989) 28, 2763-2767.
- Hare, D.R. & Reid, B.R. (1986) *Biochemistry* 25, 5341-5350.
- Ikuta, S., Chattopadhyaya, R., Ito, H., Dickerson, R.E., and Kearns, D.R. *Biochemistry* (1986) 25, 4840-4849.
- Metzler, W.J., Hare, D.R., and Pardi, A. (1989) *Biochemistry* 28, 7045-7052.
- Reid, D.G., Salisbury, S.A., Bellard, S., Shakked, Z., & Williams, D.H. (1983) *Biochemistry* 22, 2109-2025.
- Saenger, W. (1984) *Principles of Nucleic Acid Structure.* (Cantor, C., Ed.) Springer-Verlag, New York.
- White, S.A. & Draper, D.E. (1989) *Biochemistry* 28, 1892-1897.
- Wing, R., Drew, H., Takano, T., Broka, C., Tanaka, S., Itakura, K., & Dickerson, R.E. (1980) *Nature* 287, 755-758.
- Williamson, J.R. and Boxer, S.G. *Biochemistry* (1989) 28, 2819-2831.

Chapter 4

TRANSFER RNA MODELING

Introduction

The ability of single-stranded RNA to form intramolecular hydrogen bonds gives it a much greater conformational variability than double-stranded DNA. This versatility combined with the size of cellular RNAs presents us with a formidable problem as we attempt to probe the form/function relationships of RNA. There are very few well-established RNA structures. The average conformational RNA A-form helix is known from the fiber diffraction data (Arnott et al., 1973). Recently the structures of two RNA oligomer duplexes have been determined (Dock-Bregeon et al., 1989; Happ et al., 1988). They generally conform to the A-form helix with local variations in stacking similar to that seen in the structures of DNA oligomers. Most significantly, the structures of phenylalanine, aspartic acid, glycine, and initiator f-methionine transfer RNAs from yeast, and f-methionine tRNA from *E. coli* have been determined by X-ray crystallography (Sussman et al., 1978; Moras et al., 1980; Schevitz et al., 1979; Woo et al., 1980; Wright et al., 1979). As a biologically important molecule, tRNA is the touchstone for all RNA modeling and it is reassuring to find that base stacking and the A-form helix are the predominant rules in this RNA structure. Any attempt to predict the tertiary structure of RNA molecules must be able to predict the form of transfer RNA if we are to place any credence in the protocol.

Transfer RNA sequence (primary structure), Watson-Crick hydrogen bonding pattern (secondary structure), and compact folded conformation (tertiary structure) have been the focus of extensive research. With the vast improvements in DNA sequencing technology it has become trivial to determine the primary sequence of an RNA by locating it on the parent gene. The number of catalogued tRNA primary sequences is approaching 1500 (Sprinzl et al., 1989).

Using the thermodynamics of base stacking, (Tinoco et al., 1973) it is possible to evaluate the possible secondary structures which an RNA may form. Improved empirical parameters and computer programs now make it possible to produce RNA foldings which

correspond to the native hydrogen-bonding patterns with some accuracy (Jaeger et al., 1989). By comparing the sequences of similar RNA molecules of different species we can often determine which pattern of basepairing is preferred. This phylogenetic approach proved to be a very reliable predictor of the secondary structure of transfer RNA. When the possible secondary structures of various tRNAs were first compared at the Cold Spring Harbor meeting of 1966, it was apparent that the 'cloverleaf' pattern of hydrogen bonding was common to all the tRNAs known at that time (Zachau et al., 1966). The primary sequences of all known tRNAs, including the longer eucaryotic transfer RNAs, can be arranged into this phylogenetically constructed cloverleaf (Sprinzl, 1989).

Phylogeny may also indicate that some of the bases are involved in Hoogsteen or noncanonical basepairing. Folded RNA molecules can be probed for tertiary interactions with a variety of chemical and enzymatic techniques. Given the constraints that such relationships would impose on the three dimensional conformation of tRNA and some knowledge of the tertiary interactions in tRNA, several researchers tried to predict the folding of transfer RNA (Ninio et al., 1969). Michael Levitt attempted to model the three dimensional structure of tRNA using a combination of spacefilling and wireframe physical models followed by empirical energy minimization (Levitt, 1969). He and Robert Langridge even used one of the first interactive graphics systems to display his model (personal communication). But no one successfully predicted the structure of tRNA that was revealed by X-ray crystallography in 1973 (Kim et al., 1973).

The ultimate goal is to find and understand the structure of the catalytic and ribosomal RNAs. The initial plan was to replace traditional modeling techniques with a computer protocol and proceed up through the levels of structure, replacing a flat secondary structure model or physically modeled helical region with computer generated helices. Interactive graphics modeling would then be used to dock these helical subunits with the appropriate single-stranded connectors. Tertiary folding as dictated by long range interactions would then produce the final model and energy minimization should insure a

reasonable structure. This is very reminiscent of the work done by Levitt. Since graphical modeling is not inherently different from physical modeling, it can only increase the ease of modeling, not the quality. DNA oligonucleotide modeling shows that while the fine details of energy minimization and computer modeling have improved, they alone cannot yield the necessary quantitative discrimination required to distinguish among folded conformations (Srinivasan & Olson, 1987). And although some techniques (e.g. NMR or X-ray crystallography) can supply the density of information necessary to produce unique structures for small molecules, the amount of tertiary data for larger molecules remains very sparse. Therefore the original modeling approach must be modified to include distance geometry techniques if better results than those obtained some twenty years ago are to be obtained.

Distance geometry has been successfully applied to the folding of molecules based on NMR data (Wuethrich, 1989). Including this technique in the modeling protocol should provide the required objective structure construction that is missing from other modeling approaches. This algorithm has been adapted to the folding of larger RNAs with distance constraints not derived from a crystal structure or NMR data. As attempting to work with an all atom version of an RNA larger than transfer RNA is neither practical or possible, pseudoatom substitutes for the RNA nucleotides have also been developed. After a set of exploratory runs using a 6-fold reduction of each RNA residue, a series of tRNA foldings using a more drastically reduced pseudoatom set which is based on helical substructures was employed. For the majority of these foldings, only the primary structure, the phylogenetically deduced hydrogen bonding, and five long range interactions that were found independently of the crystal structure were utilized. In a control set of foldings, all the interactions listed by Levitt in 1969 were included. In all cases the addition of distance geometry to the modeling process leads to a correct prediction of the global conformation of transfer RNA.

Materials and Methods

From the beginning of this project it was decided to use programs which were already in common use whenever possible. This means that we will be able to draw on the experience and databases of previous researchers and can avoid attempting to reinvent the wheel. This should also make it easy for other scientists to implement similar studies of other systems. Lastly, thanks to the spirit of publicly funded academic research, most programs are available for the cost of the medium on which it is transferred. The molecular modeling package of programs, AMBER version 3.0, was obtained from Peter Kollman at UCSF. Standard geometry A-form helices were generated with the NUCGEN module of AMBER. The distance geometry program, DSPACE (versions 1.3 and 2.1), written by Dennis Hare and Robert Morrison was obtained from Hare Research, Inc. The interactive graphics modeling program, GRAMPS, written by T.J. O'Donnell for the National Resource for Computation in Chemistry at LBL, was used for black and white vector drawing although several more common packages (e.g. INSIGHT or FRODO) would have worked as well. Color raster pictures were made with the modeling package of programs written by Mike Connolly as distributed by the Scripps Research Institute in San Diego.

The programs used to transform files from one format to another were written in FORTRAN. PDBDS converts Protein DataBank files into the 5mer residues of DSPACE format. DSPDB converts pseudoatoms in DSPACE format into conventionally named atoms in PDB format. EXPAND adds the ability to superimpose previously created helices onto pseudohelices while converting the format from DSPACE to PDB. The program, MPSRAM, converts the translations and rotations needed to produce a particular interactive graphics view into a matrix suitable for use with the Scripps raster picture module (RAMS). RAMPAR, transforms global rotations in degrees about x, y, or z into a RAMS matrix. MATRIX transforms a PDB file by translating and/or rotating it about any axis and in a specified order. This is necessitated by the inconsistent use of pre- and post-multiplication of matrices by the various graphics devices and graphing programs.

Materials - Hardware

Circumstances conspired to force this project to utilize a number of very different computer resources. The transfers from one operating environment to another increased the amount of time it took to complete the work but it reinforced the decision to use less specialized software packages. Initial exploratory runs were done on a CRAY-XMP with 4 MWords of storage which the University had purchased. Under the UNICOS operating system, this left a maximum program size of 2.5 MWords. By tailoring the arrays, DSPACE could be configured to use 2.2 MWords, which was barely large enough for 5mer version of tRNA. As the CRAY does not have virtual memory capability and main memory cannot be expanded, even more drastic reductions in the number of atoms would still leave other RNAs far too big for this system. Further trials were run on a VAX 8800 with 2 processors, 32 MBytes of memory and an additional 50 MBytes of virtual memory. This was meant to be the VMS operating system machine for the Campus Computing Center and accounts for both it and the CRAY were obtained from a Campus Computing grant. When the VAX 8800 was removed, some work was attempted on the VAX 11/780 in the Laboratory of Chemical Biodynamics. As this proved to be too big a burden for the available resources, tRNA runs were transferred to the VAXstation II with 10MB of main memory and 10MB of virtual memory which was purchased by John Hearst.

Interactive graphics were performed on an Evans & Sutherland Multi-Picture System (MPS) using a VAX 11/780 as host. This real-time vector display system was originally purchased by LBL for the National Resource for Computation in Chemistry but was relocated to LCB due to lack of use. The MPS system was later migrated to the specially configured VAXstation II host. The line drawing illustrations are screen dumps of MPS displays or of DSPACE drawings on terminals emulating Tektronix video displays.

Materials - Structural Data

Primary Structure

Yeast Phenylalanine Transfer RNA Sequence

5'p-GCGGAUUUAG CUCAGDDGGG AGAGCGCCAG ACUGAAYAΨC
 UGGAGGUCCU GUGTΨCGAUC CACAGAAUUC GCACCA-OH

As tRNA's can be charged with the appropriate amino acid and competently participate in translation without the modified nucleotides (Sampson et al., 1989), standard RNA residues will be used instead.

Unmodified Transfer RNA Sequence

5'p-GCGGAUUUAG CUCAGUUGGG AGAGCGCCAG ACUGAAGAUC
 UGGAGGUCCU GUGUUCGAUC CACAGAAUUC GCACCA-OH

Secondary Structure

The secondary structure depiction of transfer RNA is formed by displaying the 76 residues of the yeast phenylalanine tRNA in a threeleaf cloverleaf arrangement. The helix formed by the 5' and 3' ends is called the aminoacid acceptor stem. The DiHydroUridine stem, the riboThymidine stem and the anticodon stems derive their names from the unique, conserved bases found in their loops (Fig. 1). The four basepaired stems contain 42 of the nucleotides leaving 34 single-stranded residues. It is from the double-stranded regions that the helical constraints are derived. The assumption is that these regions will form stable A-form RNA helices that persist and dominate the folded structure. In an A-form helix there are 10.8 bases per turn of the helix and 3.4 angstroms rise per base. The NUCGEN module of AMBER is used to create helical subunits and reference helices from which average distances and constraints are derived for secondary structures.

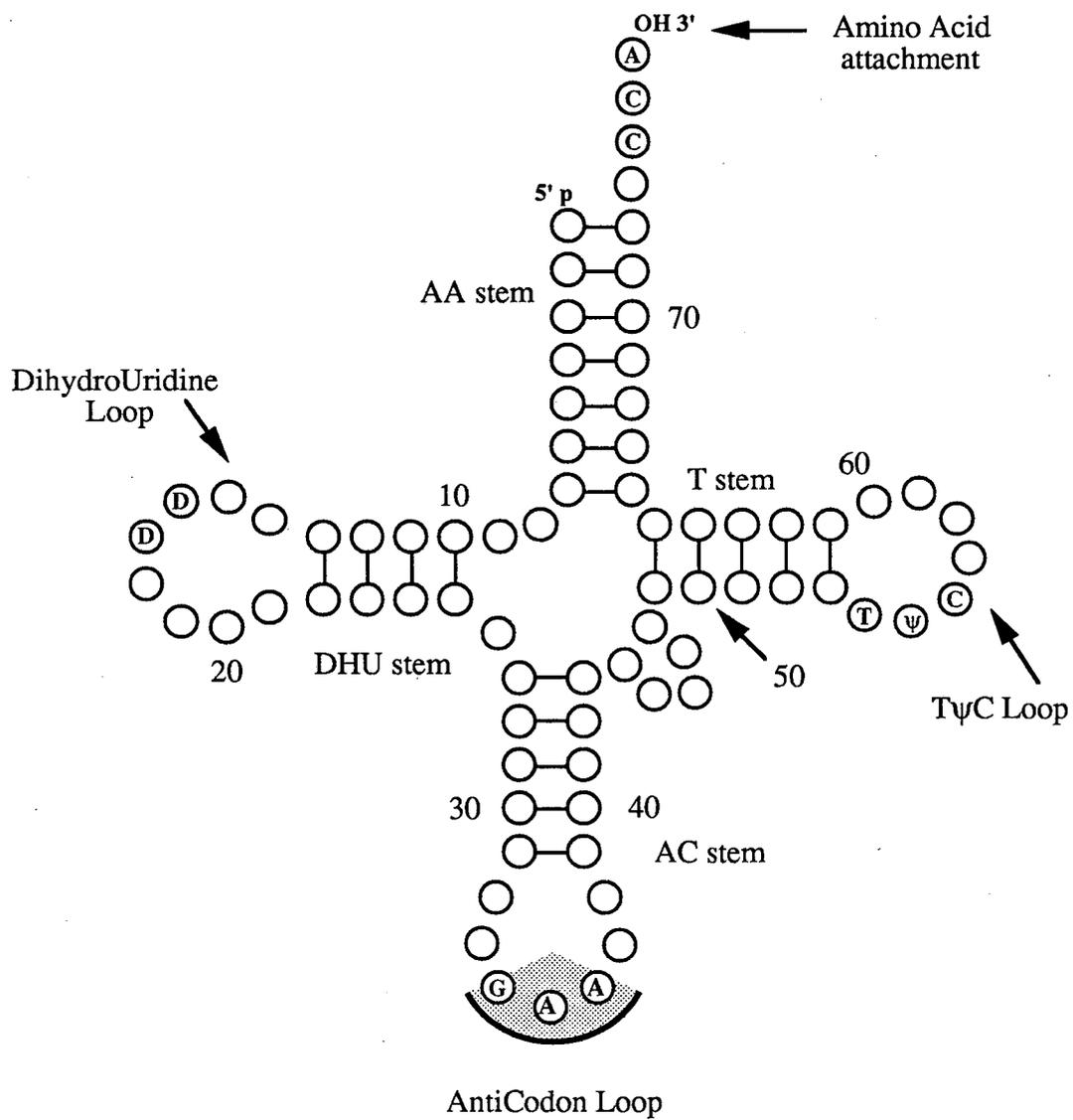


Figure 1. Transfer RNA cloverleaf secondary structure map.

Tertiary Structure

Tertiary structure information is generally hard to come by and is derived from a variety of sources. To reflect this reality, it was essential to use data that varied in kind and precision. Similarly it would not be particularly informative to use data that was derived from the crystal structure.

In some tRNAs the eighth position is occupied by the modified nucleotide thioUridine. When irradiated with ultraviolet light this residue may become crosslinked to the cytosine at position 13 (Yaniv et al., 1969) (Fig. 2). These UV induced crosslinks are the result of bond formation between the sulfur of the thioUridine and the cytosine base. As these sulfur-carbon bonds cannot exceed 2.0 angstroms, a crosslink requires that the linked bases be in direct contact. But there is no requirement that the bases be helically related. Constraints for the bases were devised which would force van der Waals contact. The constraints for the sugar-phosphate backbones of the crosslinked nucleotides, varied from van der Waals contact as a minimum to the greatest extension possible with all bonds leading to the crosslink being in an all-trans configuration.

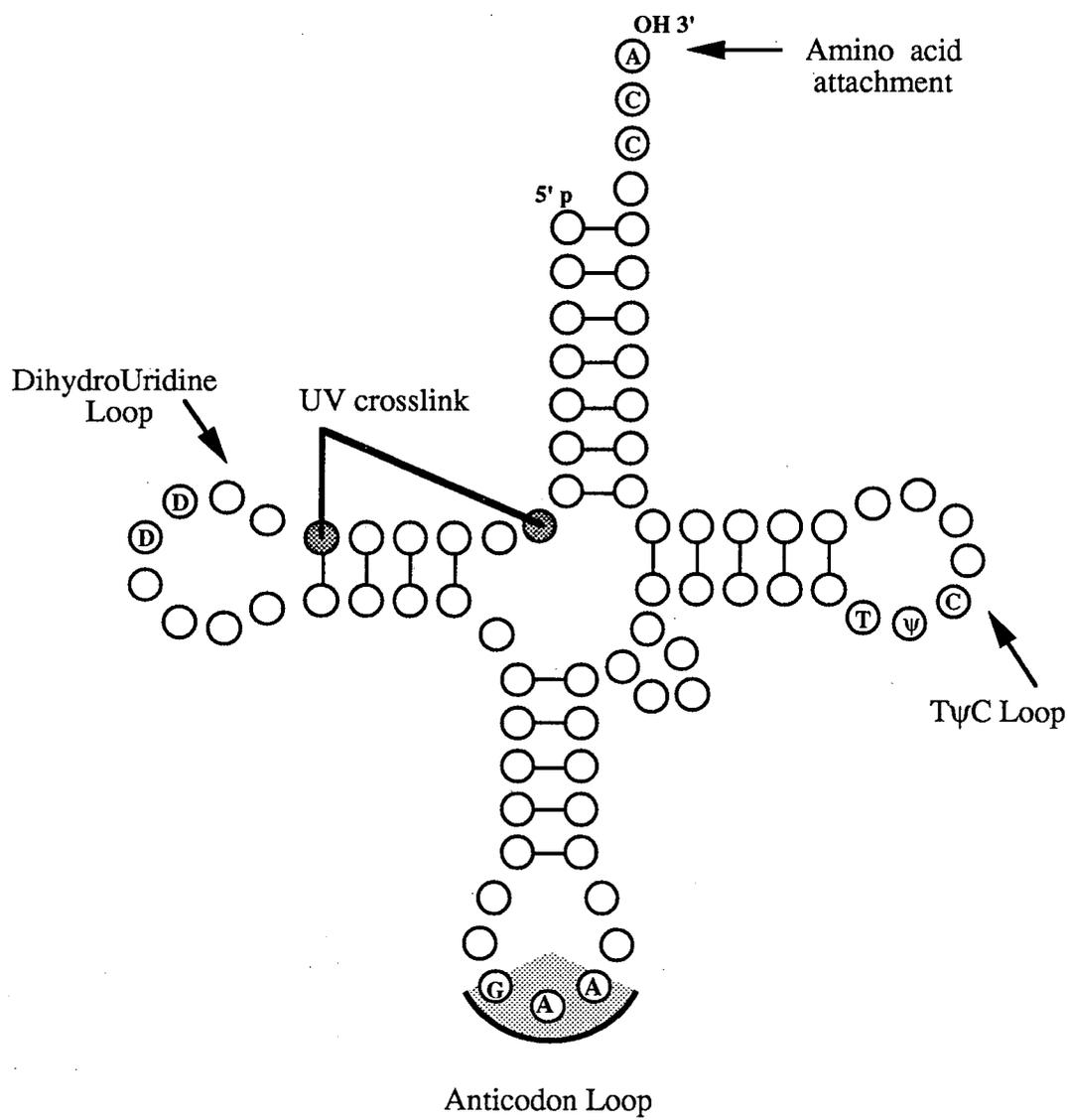


Figure 2. Ultraviolet crosslink site for tRNA.

When transfer RNA is treated with psoralen and UV radiation five crosslinks are formed (Garrett-Wheeler et al., 1984). Four of the crosslinks are in the stems of the cloverleaf structure. The fifth crosslink between the uridine in the eighth position and the cytidine in the forty-eighth position is due to the coaxial stacking of the aminoacid acceptor stem and the riboThymidine stem (Fig. 3). Psoralen is a rigid heterocyclic molecule which will intercalate between the stacked bases of a nucleic acid just as ethidium bromide does. When exposed to two quanta of UV light it will crosslink pyrimidines on opposite strands if the geometry is properly aligned. Usually psoralen crosslinks are found in helical structures but other base stacking geometries can also be crosslinked and may indicate tertiary interactions. The cyclobutane rings which it forms are highly strained and the resultant geometry fixes the distance between the C5-C6 bonds of the crosslinked bases at approximately 7 angstroms. Thus the hydrogen bonding groups must be in virtual contact while the backbones distances should vary narrowly about the constraints for antiparallel helices.

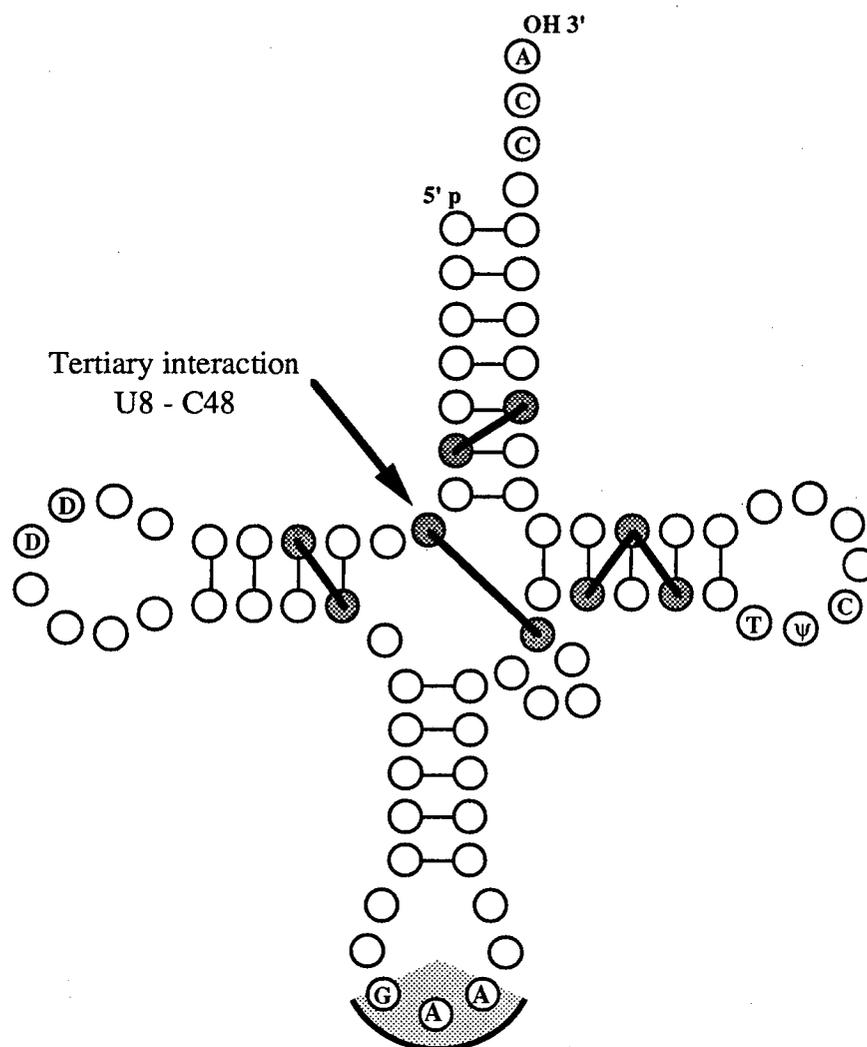


Figure 3. Psoralen crosslinks in transfer RNA.

Tertiary phylogenetic relationships are deduced based on the covariance of associated bases among differing species. This assumes some sort of hydrogen bonding is involved. Upper and lower constraints for distance geometry are based on an A-form helix. Thus the corresponding H-bonding groups are allowed to be in van der Waals contact (3.5 angstroms) at a minimum or separated by 4.0 angstroms at a maximum. Phosphate to phosphate distances vary between 17.5 and 18.5 angstroms. Three phylogenetic correlations are used as this approximates the level of tertiary relationships per nucleotide which is known for 16S rRNA. The three tertiary links, G15-C48, G18-psi55, and G19-C56, were selected from the set used by Levitt (Levitt, 1969) based on the confidence placed in them by researchers at the time and the fact that they are confirmed by the crystal structure (Fig. 4). This is justified considering the improvement in our ability to statistically detect phylogenetic relationships (Haselman et al., 1989). The remaining seven additional relationships listed by Levitt, A9-U12, A21-T54, C25-G57, C32-psi38, A44-G57, psi55-A58, and A73-A76, are used in the generation of the control group of DSPACE structures (Fig. 5).

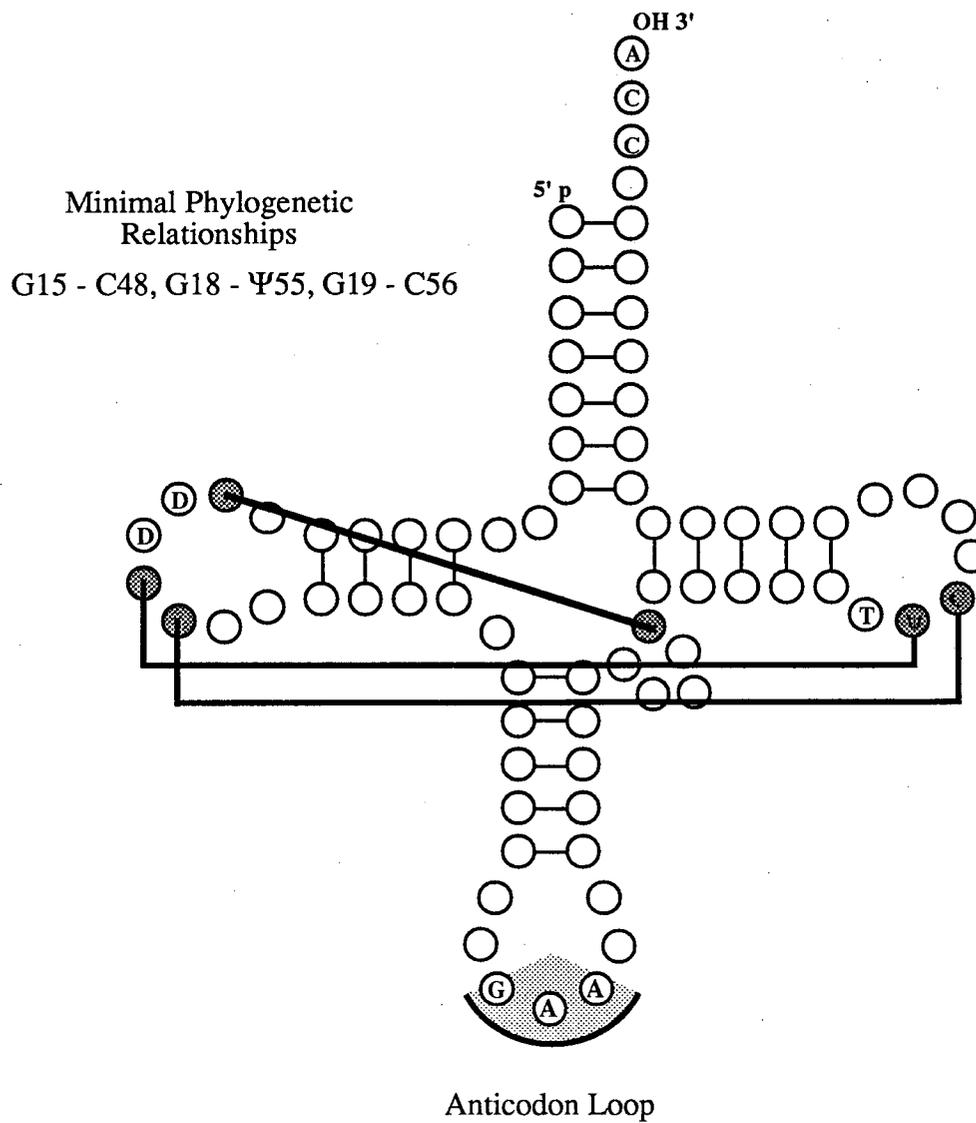


Figure 4. Suspected phylogenetic relationships that are confirmed by the crystal structure of transfer RNA.

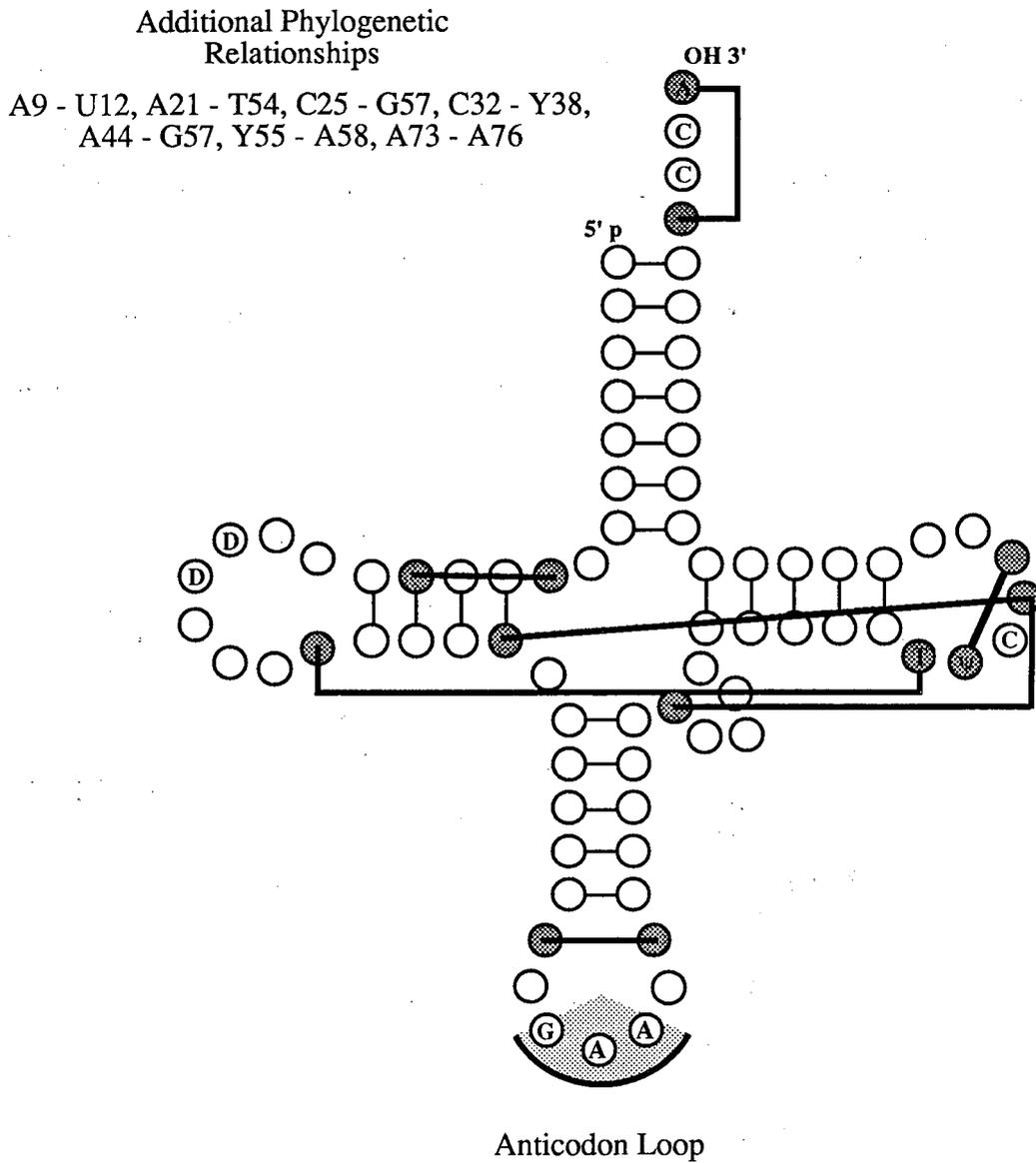


Figure 5. Other suspected phylogenetic relationships circa 1969.

Methods

Conventional Modeling

The crystal structure of yeast phenylalanine transfer RNA as further refined in 1978 (Sussman et al., 1978) was taken from the 1984 magnetic tape release of the Brookhaven Protein Data Bank (Fig. 6). As the standard of comparison, the crystal structure of transfer RNA was taken through the AMBER protocol and minimized. The names of the modified nucleotides were changed to that of their closest common residue with a text editor and the structure was minimized until the root mean square change in the structure after a minimization cycle was less than 0.1 angstroms.

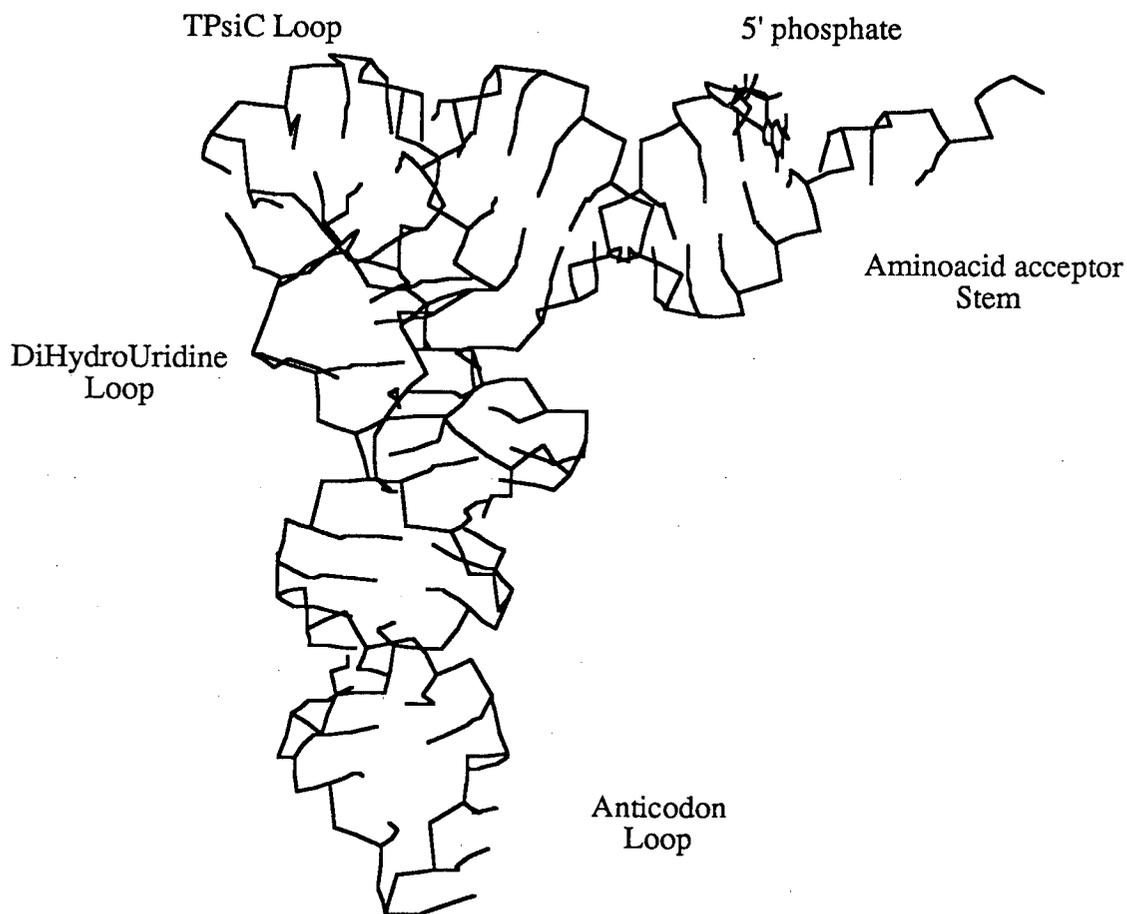


Figure 6. The crystal structure of Yeast Phenylalanine transfer RNA.

The construction of a model of the cloverleaf structure was used as the test case for the traditional modeling procedure. Substituting A-form helices for the double-stranded regions is the first step in forming a structure. A three dimensional version of the classical cloverleaf representation of tRNA was created by building the helices and single strand connectors separately using the LINK and NUCGEN modules of AMBER. The pieces were docked interactively on the MPS graphics display. Using the translations and rotations necessary for the visual docking, the original coordinate files were transformed with the MATRIX program and the separate files concatenated to form a single structure. The resultant tRNA conformer was minimized with AMBER (Fig. 7).

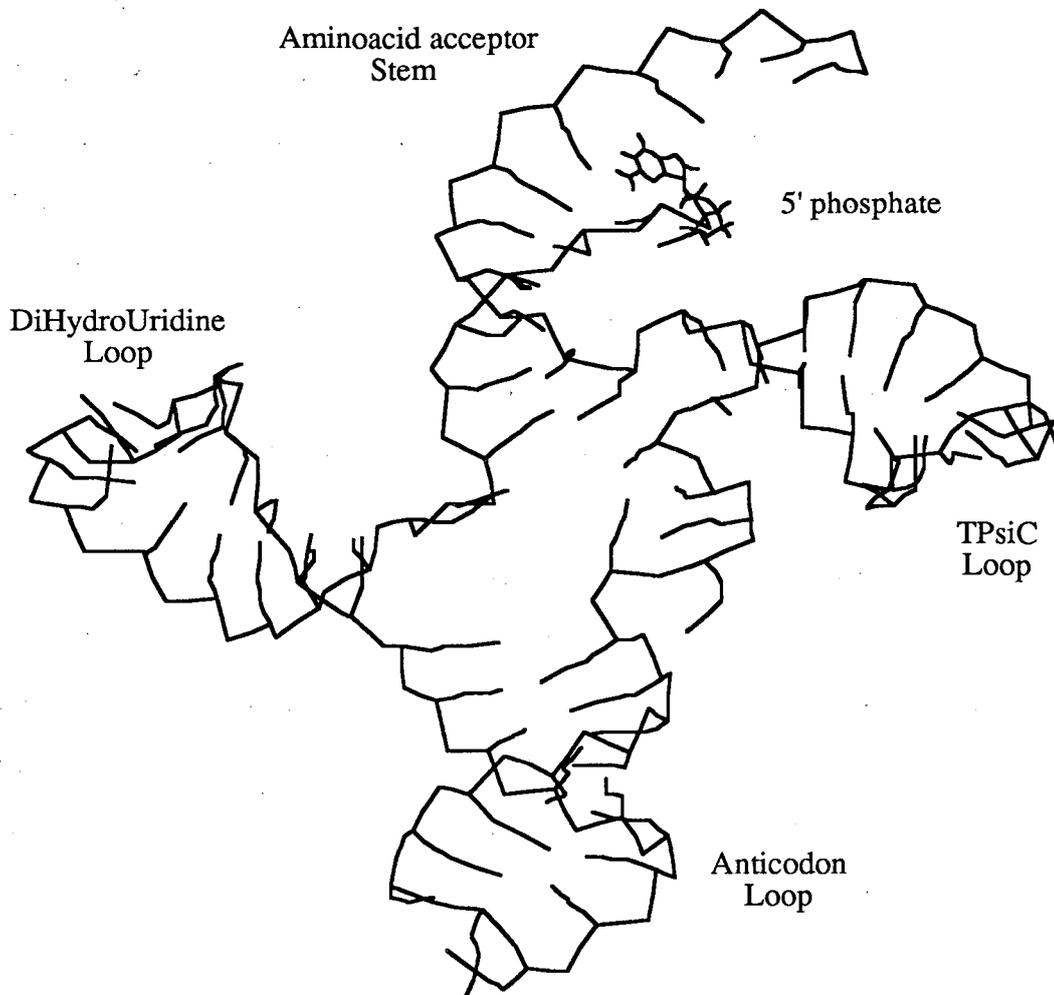


Figure 7. The tRNA cloverleaf with basepaired regions twisted into A-form helices.

Amending the Modeling Protocol

Folding the extended cloverleaf, so as to satisfy the long range interactions, is the next step in the protocol and brings us to the traditional problems caused by inadequate models and modeling subjectivity. The major problem with physical models is that they poorly mimic the molecular characteristics which predetermine the final structure. Folding nucleic acids should be easier than folding proteins since the helices are relatively stable and uniform subunits in which the differing functional groups are hidden in the interior of the helix. Therefore as a first approximation, constructing a three dimensional nucleic acid model can be considered to be a packing of rigid cylinders which are linked by flexible single strands. Interactive graphical modeling is superior to physical modeling as a digital electronic representation does not have any weight or space limitations. Additionally the changing atomic environment of small molecules can be followed quantitatively. But RNAs are not small molecules and interactive folding remains a highly subjective process which is dependent on the judgement and preferences of the modeler. It is at this stage that we introduce the DSPACE program to automatically fold the molecule. At the heart of the distance geometry algorithm is the distance matrix. It contains an entry for the distance from every atom of a structure to every other atom of the structure and consequently dominates the memory requirements of the computer program. Since the ultimate goal is to predict the structures of much larger RNAs, the level of detail that can be used to represent the tRNA molecule must be considered. Attempting an all-atom approach to the 1542 nucleotides of 16S RNA, as opposed to the 76 of tRNA, would exceed the capacity of the largest computer. Even introducing helices into the secondary structure map of 16S as was done with the tRNA cloverleaf cannot be easily done. Therefore some simplifying approximations are necessary.

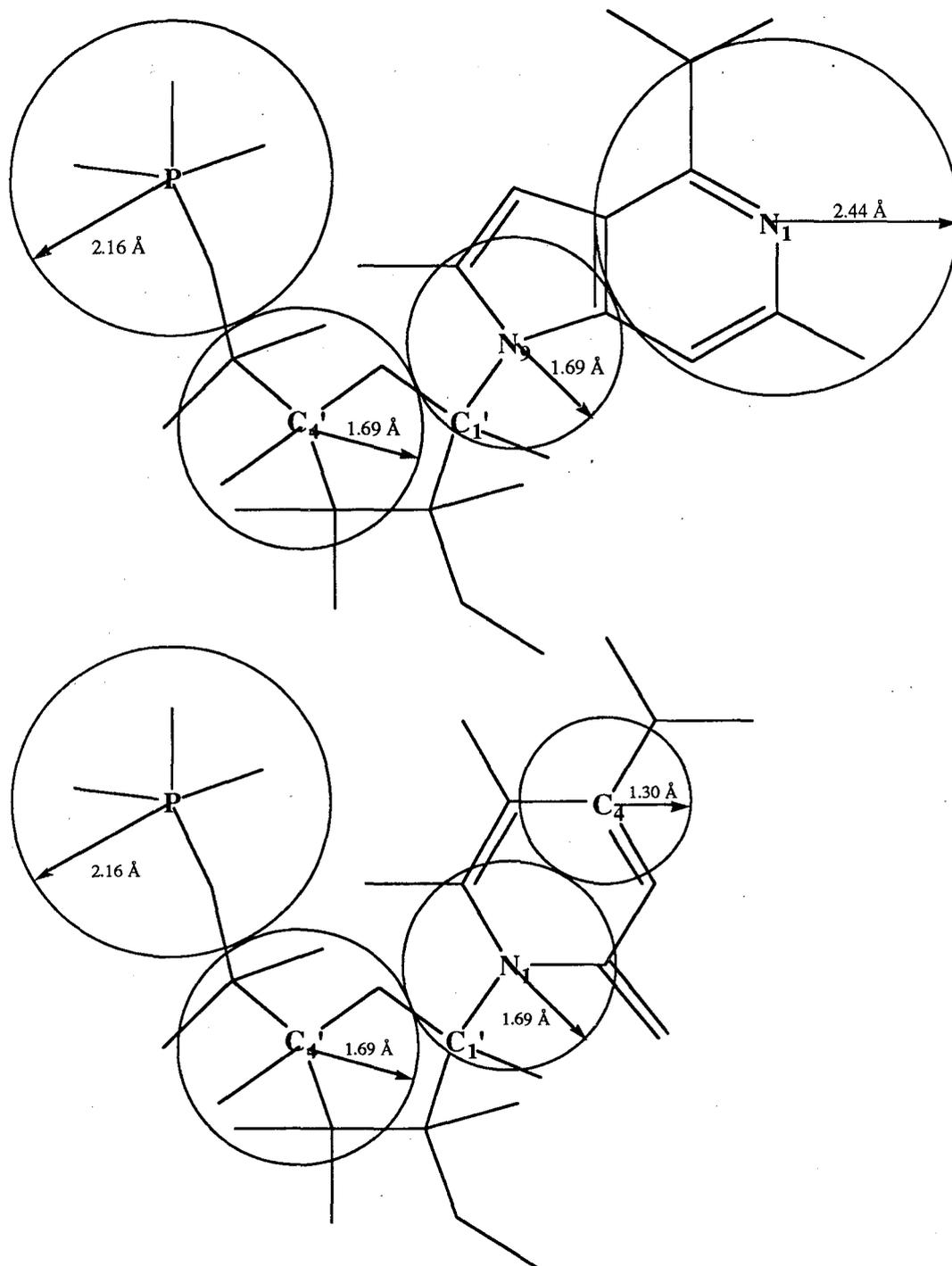


Figure 8. The six-fold reduction scheme for adenosine (top) and cytosine (bottom).

Initial Reduced Atomic Representations

The first residue of the sequence is retained in an all-atom representation as this makes it easy to find the 5' end of the molecule when it is drawn on a graphics terminal and it also facilitates the reintroduction of the all-atom version of the other nucleotides after the distance geometry manipulations are completed.

As illustrated in the diagram of the adenosine and cytidine monophosphates (Fig. 8), a 6-fold reduction in the size of the problem can be achieved by using only 5 pseudoatoms of appropriate radii. This will retain the essential space-filling characteristics of the purines and pyrimidines. The radii of these groups will also allow the close approach of other bases that is required for hydrogen bonding and base stacking. The C1' pseudoatom is given a negligible radius to avoid conflict with the necessary increase in the size of its neighbors. From a spacefilling viewpoint it could be eliminated altogether, but it is retained in order to provide a schematic tracing of the nucleotidyl geometry. The P, C4', and C1' atoms (b, j, k pseudoatoms) form the common sugar/phosphate backbone of each nucleotide. The N9 and N1 atoms (l and e pseudoatoms) of purines or the N1 and C4 atoms (l and f pseudoatoms) of pyrimidines were used as the atomic level replacements for the base attachment and hydrogen bonding groups of the bases. At this level of reduction there is no difference between adenine and guanine or uracil and cytosine. The library of residue level replacements constructed from these pseudoatoms were derived from the structure files created with the NUCGEN module of AMBER. For clarity, the names of the pseudoresidues were created by preceding the normal DSPACE residues names with an 'x'. For example the single-stranded molecule pApUpGpC could be constructed from the library simply with the DSPACE sequence arna, urna, grna, crna. Each nucleotide is constructed from the library substructures po4, ribose, base. The reduced sequence is called with xade, xura, xgua, xcyt and these residues call up the reduced library structures xa, xu, xg, and xc respectively. A library definition specifies the number of atoms in a residue, the number of internal and external bonds, and the bonding connectivity. The file

also specifies the names and ideal xyz coordinates of atomic constituents including the expected names and locations of atoms external to the residue. It is from these geometries that the bond lengths, bond angles, and dihedral angles are determined. From these standards and any experimental constraints, DSPACE constructs the distance matrix which is used to create three dimensional structures and for conjugate gradient minimization. For a pseudoresidue structure the b to b bond length is 5.65 angstroms, the b-b-b backbone angle is 150.3 degrees, and b-b-b-b backbone dihedral angle is 21.3 degrees. The pyrimidine base angle to the backbone will be 84.0 degrees (b-b-f) and the purine/backbone angle will be 94.7 degrees (b-b-e). The DSPACE atomic radii in angstroms are 2.16(b), 1.69(j), 0.10(k), 1.69(l), 2.44(e), 1.30(f), 1.9(p), 1.5(c), 1.3(n), 1.3(o), and 0.9(h).

To force the double-stranded regions into a helical conformation with distance constraints would be a large and onerous task. It would require the specification of interatomic distances from each pseudoatom to at least several of the atoms on the opposite strand. With DSPACE it is possible to define each helical segment as a residue made up of atoms or functional groups. This automatically provides extra constraints for creating a particular local geometry. But this advantage is not free. Each helical residue definition must vary as the sequence of a strand varies. Therefore I grouped the individual residues of a helical strand into a single library definition file based on a geometry derived from the standard Arnott A-form RNA helix. The 5 bases of the 5' strand of the anticodon stem are replaced in the primary structure definition file with the library name, ac5. This in turn reads in the definition acode5 in which the 25 pseudo atoms are one strand of a helix. In this manner the program is forced to consider the helicity of such a region with the same weight that it considers the planarity of the all-atom version of a purine. Having assured the helicity of the individual strands, constraining the basepaired residues should reproduce the double-stranded form.

The original runs with 5 replacement atoms per nucleotide were used to explore the feasibility of this approach and to determine the range for upper and lower bounds, tertiary

constraints, and flexibility of other program variables. No special weights were given to the helical regions. Only the hydrogen bond donor to acceptor or the phosphate to phosphate distances were specified for base paired nucleotides. At this level of abstraction a transfer RNA has 408 atoms in 42 residues. The structures generated by DSPACE were minimized using a conjugate gradient method until the change in error function between two successive steps was less than 0.1 angstroms.

Pseudo-helical Modeling Constructs

Since the size of the program increases as the square of the number of atoms and computation times increase at a rate proportional to the cube of the number of atoms, it is necessary to reduce the number of pseudoatoms even further in order to attempt the folding of larger RNAs. In fact any reduction scheme which does not reduce the number of pseudoatoms required to describe 16S RNA below the total number of residues in 16S RNA (~1500) will produce a program which is too large to compile or run on the MicroVAX (>50 MBytes). In devising such a reduction scheme there are two major problems which must be considered. Foremost is the necessity of maintaining enough structural information so that the resultant model will still resemble the molecular folding in a significant way. Second, it must be possible to recover the full molecular structure uniquely and unambiguously. As a first step single-stranded residues might be replaced by a single pseudophosphate. It is an unfortunate consequence of this or any other extreme reduction scheme, that the spacefilling nature of the model cannot be maintained simply by increasing the size of the pseudophosphates. The required radius would extend so far to the side opposite the base as to prevent the close packing which must be a feature of a compact conformation. Constructing a helix from such simple residues would also be problem. As the helical secondary structure is the basis of this modeling protocol, it is only logical to devise some scheme based on them.

Simple replacements for the residues in protein alpha-helices are possible since the helices are single stranded and the mass of the backbone is distributed along the center of the helix axis. B-form DNA would be harder to model because the helices are double-stranded. But at least the center of a Watson/Crick base pair is close to the helical axis and the plane of the base pair is perpendicular to the axis. The base pair plane of A-form RNA is substantially tilted with respect to the helical axis. Additionally the axis is displaced almost five angstroms into the major groove of the basepair, producing a helix in which the mass is distributed on the surface of a hollow cylinder about the axis (Saenger, 1984). Therefore a scheme which attempted to replace a helix with simple spheres centered on the helical axis could not accurately represent the mass distribution, bonding characteristics for single strand links, or the simple branched tree structure of an RNA. In computer data structures, a tree is formed from nodes. Each node has a link on one side to its father or predecessor in the tree and links on the other side to its sons or successors. Proper definition of the structure allows all the nodes to be visited or evaluated uniquely and unambiguously. Replacing a double-stranded segment of an RNA with a single artificial construct is incompatible with a simple tree structure since the construct will be linked to both predecessors (5' nucleotides) and successors (3' nucleotides) on each side.

A helical segment can be represented by pseudoatoms which replace the phosphates of only the first and last residues of the helical segment. These residues will also have pseudoatoms for the C4 atoms of pyrimidines or the N1 atom of purines. In this manner the helix length and twist will be specified and preserved in the residue definition. The double-stranding can then be specified as distances to the corresponding pseudophosphate and basepairing group of the opposite strand. Implicit in such a drastic reduction scheme is the phylogenetic reasoning that a helix and not its specific sequence is most important. Under this scheme replacing a helix of four base pairs requires as many pseudoatoms as a single-stranded region. Helical segments of five or more bases will require the same four pseudoatoms to replace a greater number of residues. Replacing short helices of three

basepairs or less would take more pseudoatoms than a single-stranded representation, apart from any scientific judgement as to how believable such helices are. To insure the proper orientation of such schematic strands, the base pairing constraints will be expanded to include not only the H-bond donor and acceptor pseudoatoms, but also the pseudophosphates and the pseudophosphate to hydrogen bonding pseudoatom of the opposing residue.

Of course not all helices will be perfectly regular. The twelve basepairs of the amino acid acceptor stem and the riboThymidine stem form a single, stacked unit which is clearly an A-form helix. But while the 3' strand consists of a single contiguous unit (bases 61-72), the 5' half is formed from two unequal strands which are distant from each other in primary sequence (bases 1-7, 49-53). The transition from the first part of the amino acid stem to the 5' strand of the DiHydroUridine stem is particularly sharp with only two bases (U8 and A9) spanning these two structures. Yet the helical structure is maintained and the bases U7 and C49 are stacked. Attempting to define the two halves the 5' strand of the amino acid acceptor stem as a single residue would require a very convoluted redefinition of the primary structure of tRNA and major modifications in the logic of the programs. This would be a major task and the results would be difficult to apply to other molecules. Some modification the 3' strand definition must be considered. By including the pseudophosphate and base for any residue which is involved in basepairing to a junction of some sort on the opposite strand, the necessary links for double-stranded constraints will be provided. The addition of four more pseudoatoms per junction weakens the rationale behind the new reduction scheme. It also introduces new bonds, angles, and dihedrals which may be distorted in the folding process. Therefore it will allow irregular helices, kinks, and bulges to appear as required by the interplay of structural elements and secondary and tertiary constraints. Considering the state of our understanding of RNA structure this is an improvement in the model.

In this manner the 2500 atoms in 76 residues of yeast phenylalanine tRNA are reduced to 99 pseudoatoms in 42 residues. To minimize the overcompaction problems seen in the 5mer minimizations, DSPACE conjugate gradient refinement was restricted to 64 steps. Each of the structures generated was visually examined. The structures were evaluated on the basis of their cgr error functions and how well they could be superimposed on the crystal structure. The best structures were then taken through the final steps of the modeling protocol. The severely reduced helical residues were replaced with ideal A-form helices of the appropriate sequence and the xyz coordinate file converted to PDB format (EXPAND). Given the all atom representation of the first nucleotide, AMBER, automatically reinserts any missing atoms in succeeding nucleotides by referring to its library of standard residues. Energy minimization is used to resolve any structural conflicts which violate the rules that have been deduced from smaller molecules. These refined structures were used to generate spacefilling and schematic raster pictures to facilitate the visual analysis of the final results.

Results

1) Early five pseudoatoms per nucleotide runs (5mer)

Once the pseudoresidues had been defined, creating the distance constraint files was very easy and could even be done interactively. Running the program on the CRAY or VAX 8800 could be done in real time without inordinately monopolizing the resources of either the computers or the user. This was not feasible on the VAX 11/780 or VAXstation II and DSPACE runs were conducted in the background of the multiuser operating systems as batch jobs. The early versions (1.3) of DSPACE included the ability to display line drawings on a Tektronix 4014 terminal. Finding an informative view with sequential x, y, or z rotations that are possible on such terminals requires experience, but from the very first structure constructed it was apparent that this method had succeeded in finding the proper global fold for tRNA (Fig. 9). The basic bent shape and proper DHU loop and T Ψ C loop stacking were always present. Equally obvious was the fact that the double-stranded regions were not being forced into the proper helical configuration. It was also clear that these structures were too compact and had serious van der Waals conflicts. Some of the problems can be attributed to the reduced residue library as is clear from the large bounds violation values for both the crystal structure and the minimized cloverleaf. When the



xt4



wt5



yt5



zt5



ytx2



yt4

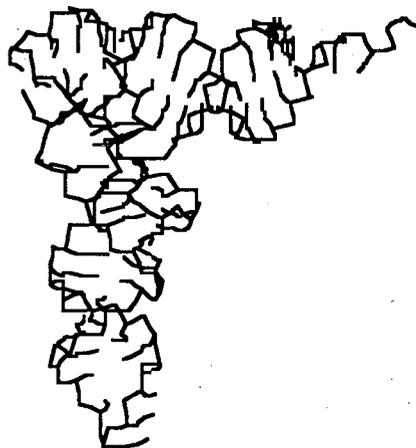
Figure 9. Six transfer RNA structure produced with the 5mer pseudoatom reduction scheme.

crystal structure is refined with respect to the bounds matrix the resultant structure shows the same distortions around the 5' phosphate and anticodon loops as is seen in the DSPACE generated structures (Fig. 10). From the 408 atoms, four helical segments, and five long range constraints, DSPACE is able to determine 4749 distances and the average atomic separation is 43.12 angstroms. When EXPANDED to an all atom representation, the xt4 structure has a volume of 60,000 cubic angstroms, a 24.28 angstroms radius of gyration, and an average phosphate to phosphate separation of 5.67 angstroms. The bounds matrix can be created and a single folded structure fully refined in four hours of microVAX computer time.

Representative structures:	bounds violation err	superposition fit err
xray crystal	2304.50	-
crystal refined	179.94	7.13
cloverleaf	22077.11	19.76
xt4	187.41	12.86
wt5	189.61	10.43
yt15	190.01	12.98
zt5	262.01	11.68
ytx2	309.08	13.48
yt4	370.48	11.13



xt4



Xray crystal
structure



cgr minimized
crystal structure

Figure 10. A comparison of the xt4 5mer structure to the crystal structure before and after conjugate gradient refinement with the crystal structure of tRNA.

2) Helical pseudoresidues with 8 constraints per helix (dyn)

The limited success achieved with the first reduction scheme and the complete failure to compile, much less execute, a version of DSPACE which would be able to handle the 5mer representation of 16S rRNA, lead to the use of the reduced helices representations. As this set of pseudoresidues cannot be made to assume realistic spacefilling qualities, it is even more important to avoid the overcompaction problems of the early runs. When minimized until the change from one step to next is less than a 0.1 decrease in the error function, the total error function approaches 10 angstroms. With the helical pseudoresidues it is possible to obtain error functions of the same order of magnitude (i.e. <100) after only 64 minimization steps. The values of the error functions for new structures minimized in this manner bracket the error value which DSPACE assigns to the crystal structure. The resultant structures also have dimensions similar to the crystal structure of tRNA. The 100 trials using these parameters and eight constraints per helix showed great promise. The phosphate to phosphate, phosphate to hydrogen bonding pseudoatom of the paired residue were given an angstroms leeway about their ideal separations and Hbond donor to Hbond acceptor distances were confined to 3.5 to 4.0 angstroms. When the structures were visually inspected it became apparent that there were still some problems with overcompaction. As a possible remedy these structures were subjected to simulated annealing which 'heated up' the molecules by adding small random distances to the bonds. The structure is then allowed to reach a new dynamic equilibrium before it is reminimized. This did not significantly improve either the compaction or correctness of the structures (Fig. 11).

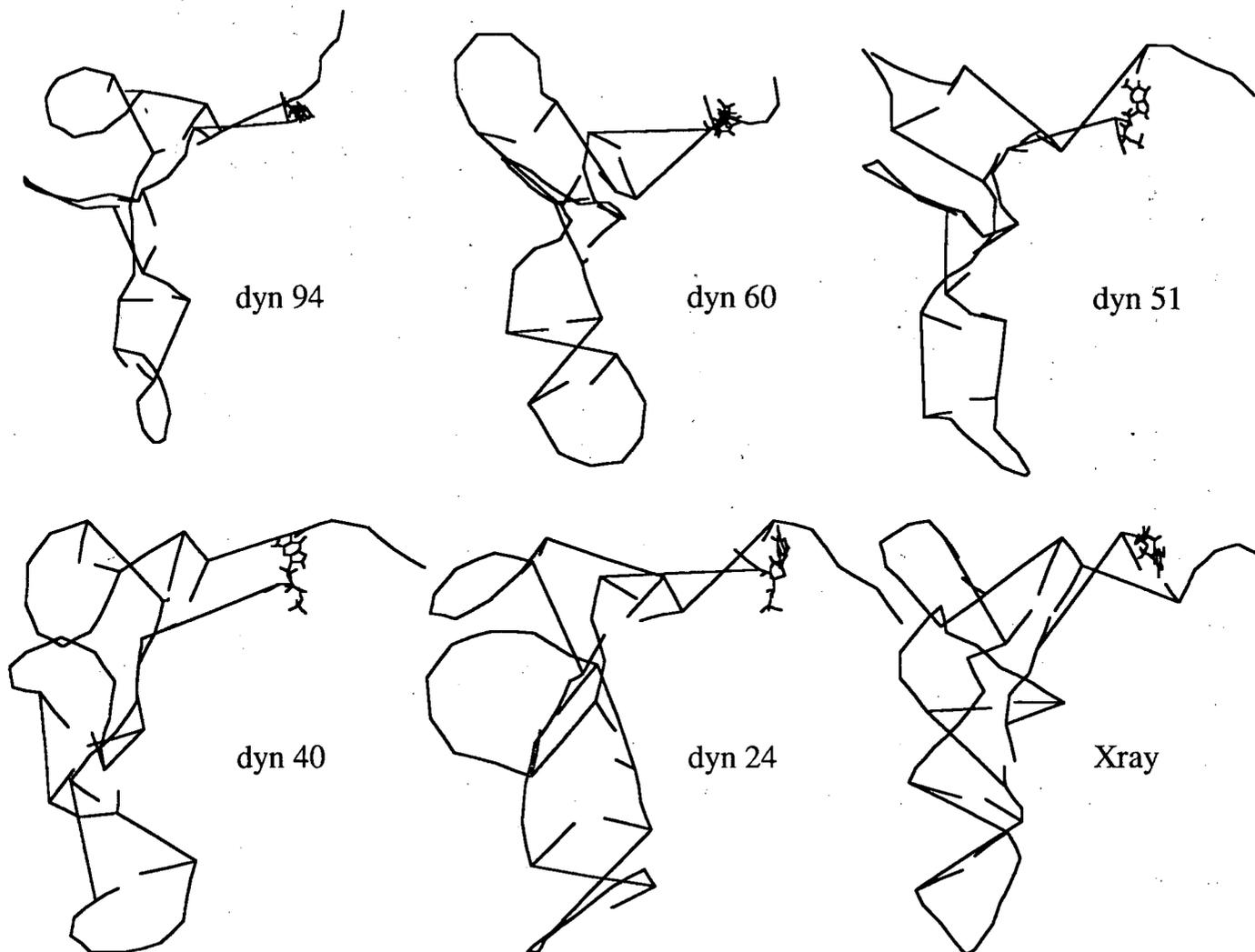


Figure 11. The five best tRNA models produced with simulated annealing compared in scale with the crystal structure of tRNA.

From the 99 atoms, four helical segments, and five long range constraints, DSPACE determines 374 distances and the average pseudoatom separation is 37.86 angstroms. In 100 trials this procedure produced 47 independent structures. All of these structures resembled the 'L' shape of the tRNA crystal structure. The cgr error functions for these structures ranged from 41.67 to 177.88 square angstroms. When superimposed on a pseudoatom version of the crystal structure, this series of structures had fit errors that varied from 4.56 to 12.30 angstroms.

Representative structures:	after cgr only		after annealing		
	cgr err.	fit err	cgr err.	bounds err.	fit err
dyn94	35.71	4.56	11.99	40.93	10.44
dyn60	41.64	8.74	22.39	64.13	8.23
dyn51	41.87	9.43	19.32	50.83	11.05
dyn40	54.79	6.02	20.28	59.01	8.82
dyn24	55.75	10.85	9.21	44.81	10.89

3) Helical pseudoresidues with 16 constraints per helix

a) Structures created with 5 long range constraints (vt)

In the next 100 trials named vt, the dynamics cycles were abandoned and an additional eight constraints per helix were added. The distances from the phosphates and hydrogen bonding groups for one strand of a helical region to the end of the opposite strand that is not directly basepaired are constrained to be within 2.5 angstroms of the ideal values for an A-form helix. The excellent results suggest that simulated annealing is too sophisticated an approach for such a simple model. All of the structures possess the characteristic bent shape although the sharp kink in the T Ψ C loop of vt50 makes this less obvious (Fig. 12).

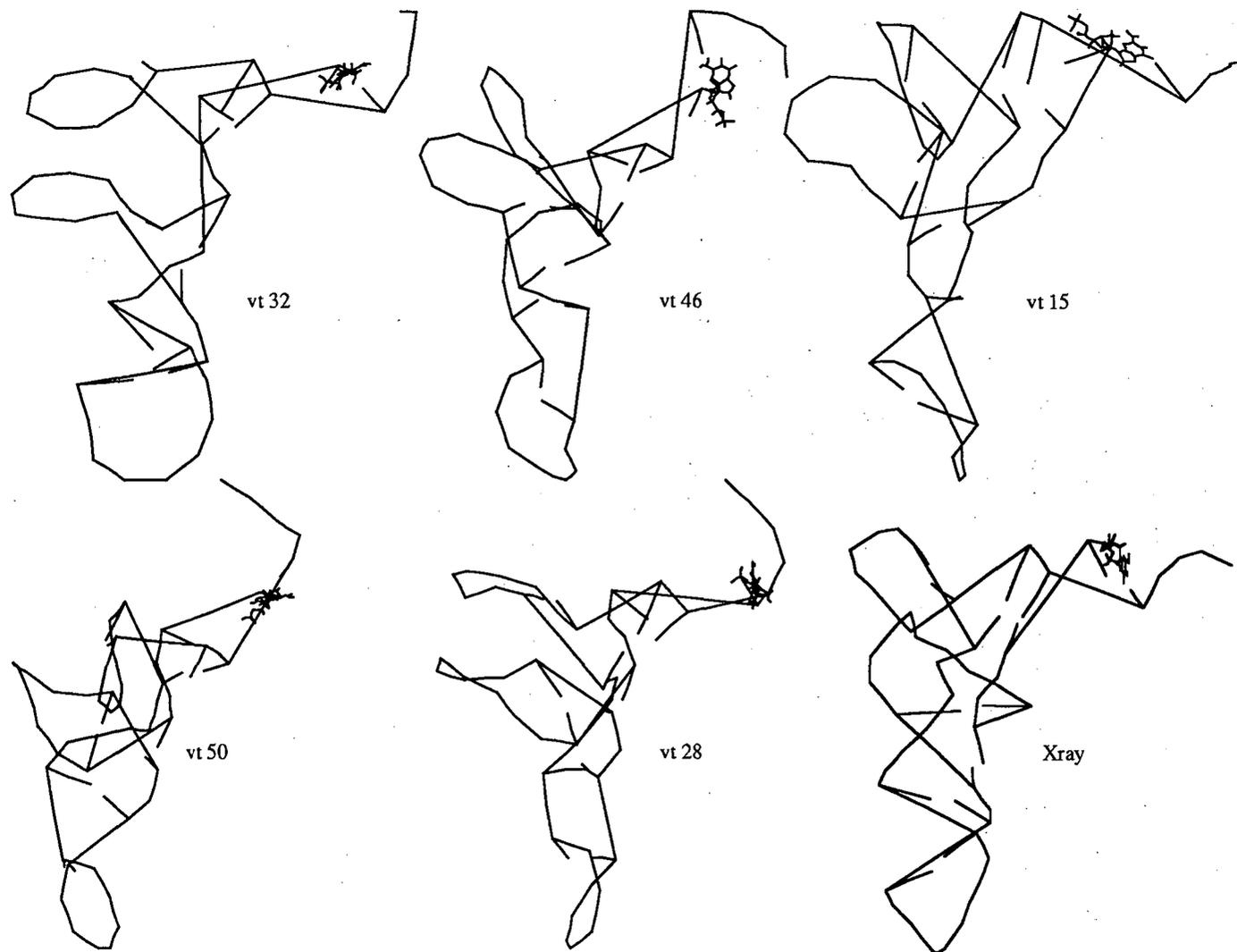


Figure 12. The five best vt structures compared in scale to the crystal structure of tRNA.

The same problem can be seen in the anticodon loop of vt15. The too close approach of the 5' and 3' strands of the amino acceptor stem is also a common problem and is directly attributable to poor spacefilling characteristics of the pseudohelices. Despite these shortcomings all the vt constructs clearly resemble the crystal structure. Among the ten best structures, vt58 is an almost perfect match, especially considering the primitive pseudoresidues (Fig. 13). From the 99 pseudoatoms, four helical subunits, and five long range crosslinks, DSPACE is able to determine 374 distances with an average separation of 36.31 angstroms. All 45 independent structures produced in 100 trials have the distinctive 'L' shape. The cgr error functions for the models ranged from 32.35 to 274.72 square angstroms and the error violations for all bounds varied from 85.15 to 248.48 angstroms. The error for superposition on the crystal structure was as low as 7.06 and as high as 12.43 angstroms. With the pseudohelical residues and limited conjugate gradient refinement it was possible create the 100 models in only 1.5 hours of microVAX computer time.

Representative structures:	cgr err.	bounds err.	fit err
vt32	32.35	85.15	9.58
vt46	45.43	96.02	10.19
vt15	46.16	107.70	10.70
vt50	56.69	120.92	11.92
vt28	59.34	113.66	10.37
vt58	75.78	132.48	7.06

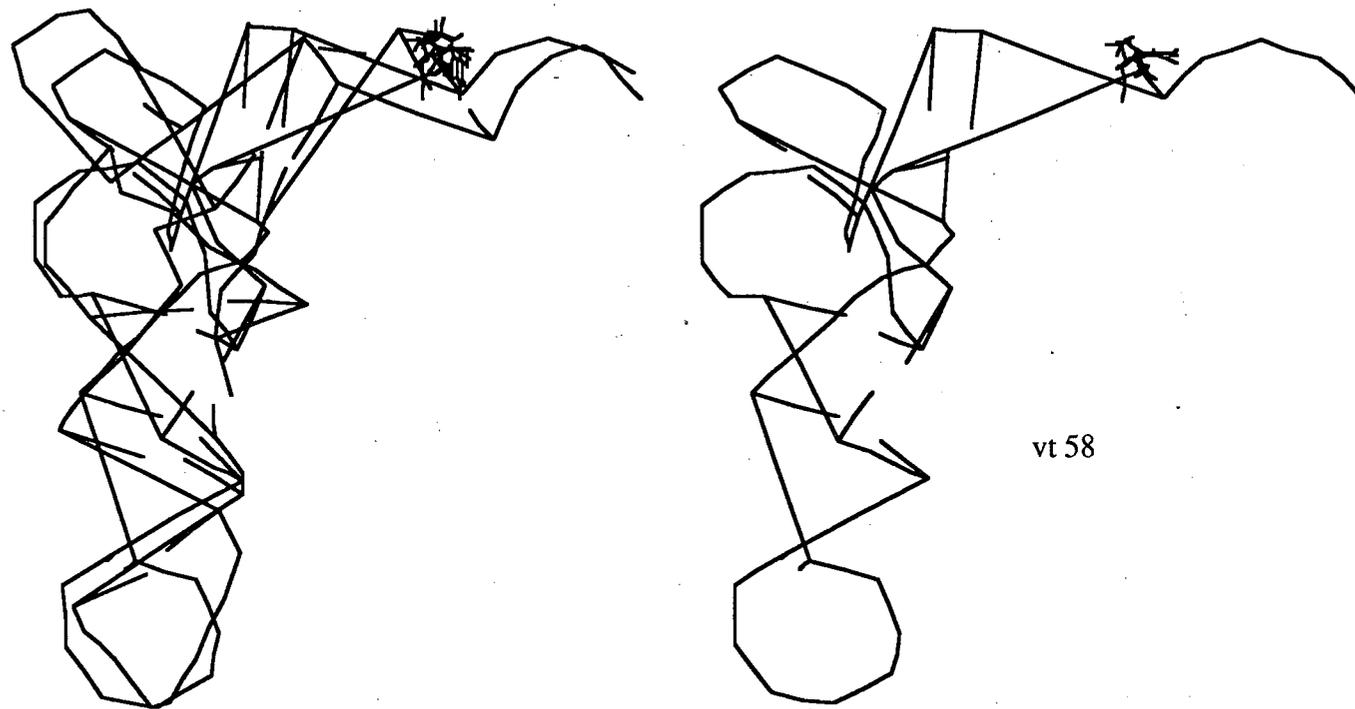


Figure 13. The structure of vt58 superimposed on the crystal structure of tRNA.

b) Structures created with 12 long range constraints (bad)

To see if the initial results were somehow dependent on the choice of crosslinks, a new bounds matrix was constructed which included all the suspected relationships listed in Levitt's paper. When all the suspected interactions listed in 1969 are included, 31 of the 48 distinct structures are of the bent variety (Fig. 14). The remaining 17 have a linear or tangled conformation and the majority of these resemble the model which Levitt constructed (Fig. 15). The fact that almost two-thirds of the distinct structures formed evince the distinctive 'L' shape and follow the proper folding conformation demonstrates that the use of distance geometry to objectively fold RNA is eliminating the subjectivity of a human modeler and that the structure of tRNA is so firmly determined by the helical relationships that it can tolerate bad data. The structure which has the lowest error is also the structure which can best be superimposed on the reduced form of the tRNA crystal structure and of the ten structures with low cgr error functions, only bad30 has poor angularity. As before DSPACE determines 374 distances but the additional seven false long range relationships increases the average pseudoatom separation to 36.85 angstroms. 66 of the structures formed had the angular 'L' shape while 34 had rod-like or tangled conformations. The cgr error range was 47.11 to 168.80 square angstroms and the violations for all bounds varied from 99.86 to 191.48 angstroms. The fit error for superposition onto the crystal structure ranged from 5.07 to 13.35 angstroms. The bad40 structure can be EXPANDED to an all atom structure with a volume of 49,100 cubic angstroms with a radius of gyration of 22.72 angstroms and an average phosphate to phosphate virtual bond length of 5.79 angstroms.

Representative structures:	cgr err.	bounds err.	fit err
bad40	47.11	106.71	5.07
bad67	48.75	105.60	10.41
bad30	49.80	99.86	8.65
bad3	53.31	105.02	11.77
bad2	58.14	110.43	12.42
bad42	62.41	113.05	10.24
bad45	76.28	128.98	8.06

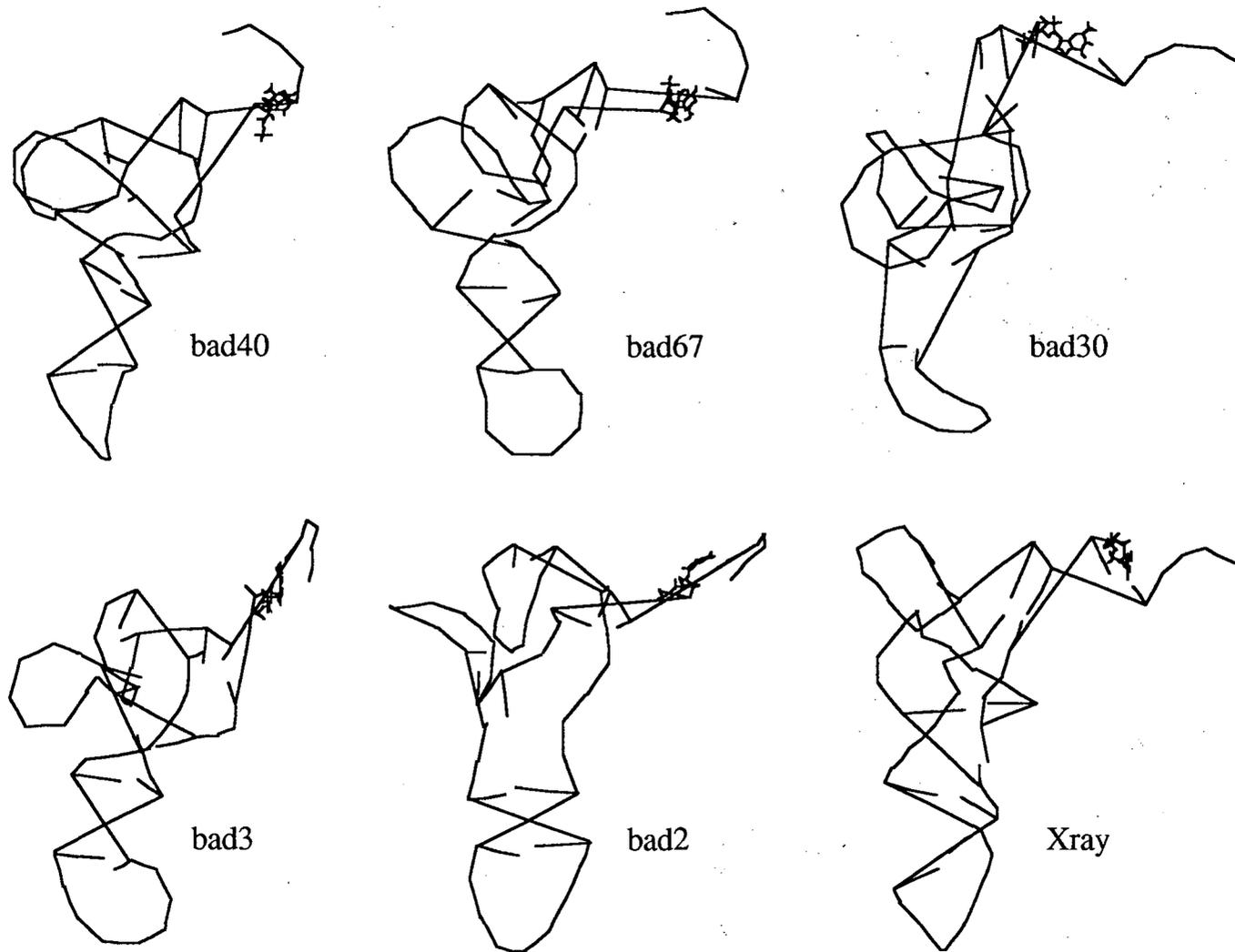


Figure 14. The best five structures created with all the suspected phylogenetic relationships.

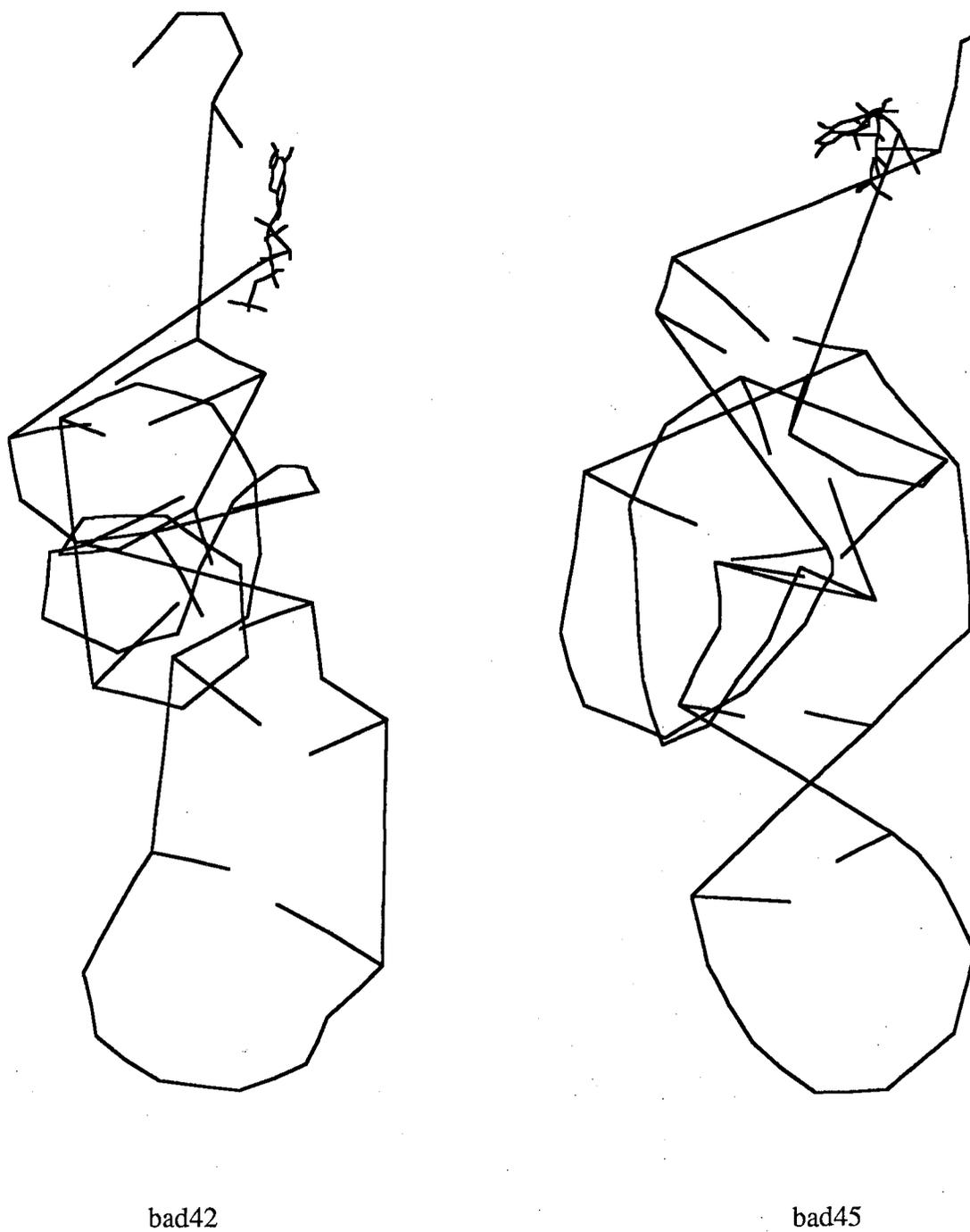


Figure 15. Two examples of the linear structures which resemble early models of tRNA.

c) Additional structures made with 5 long range constraints (ext)

Further transfer RNA runs were resumed after the first 16S rRNA work was completed. The final 300 tRNA structures were generated with a newer issue of DSPACE (version 2.1). The upgraded program required a redefinition of the library files and the change of some atom names (e.g. H5A' -> h5'e). As it was necessary to update all the residue definition files to conform to the new format, this opportunity to include the o3' oxygen of the first residue to the first pseudoresidue was utilized. In the DSPACE residue definition library, the o3' atom is considered to be part of a phosphate group. Rather than change the common residue definitions, a set of four reduced residues were created to provide a smooth transition from standard to reduced residues. This increased the total number of atoms in the reduced version of tRNA to 100. The van der Waals radii definitions were moved from a FORTRAN subroutine to a library definition file (ATOMS.DEF) that was easier to alter and does not require that the entire program be relinked for every change in these parameters.

The new series of structure constructions yielded results similar to the previous runs based on helical pseudoresidues (Fig. 16). From the 100 pseudoatoms, four helices, and five long range crosslinks, DSPACE is able to determine 377 distances. The average pseudoatom separation is 37.7 angstroms. Three different sets of 100 trial structures produced 41, 47, and 50 independent structures in each set. When the results for all the trials were collated, there was a total of only 60 independent conformations. Four of these structures are tangles which do not resemble the proper angular conformation.

Extra care was taken in the generation of the ext structure set with the conjugate gradient refinement being done in separate sets of 32 cycles. It is clear that 32 cycles of refinement are sufficient to reach the error neighborhood which DSPACE assigns to the

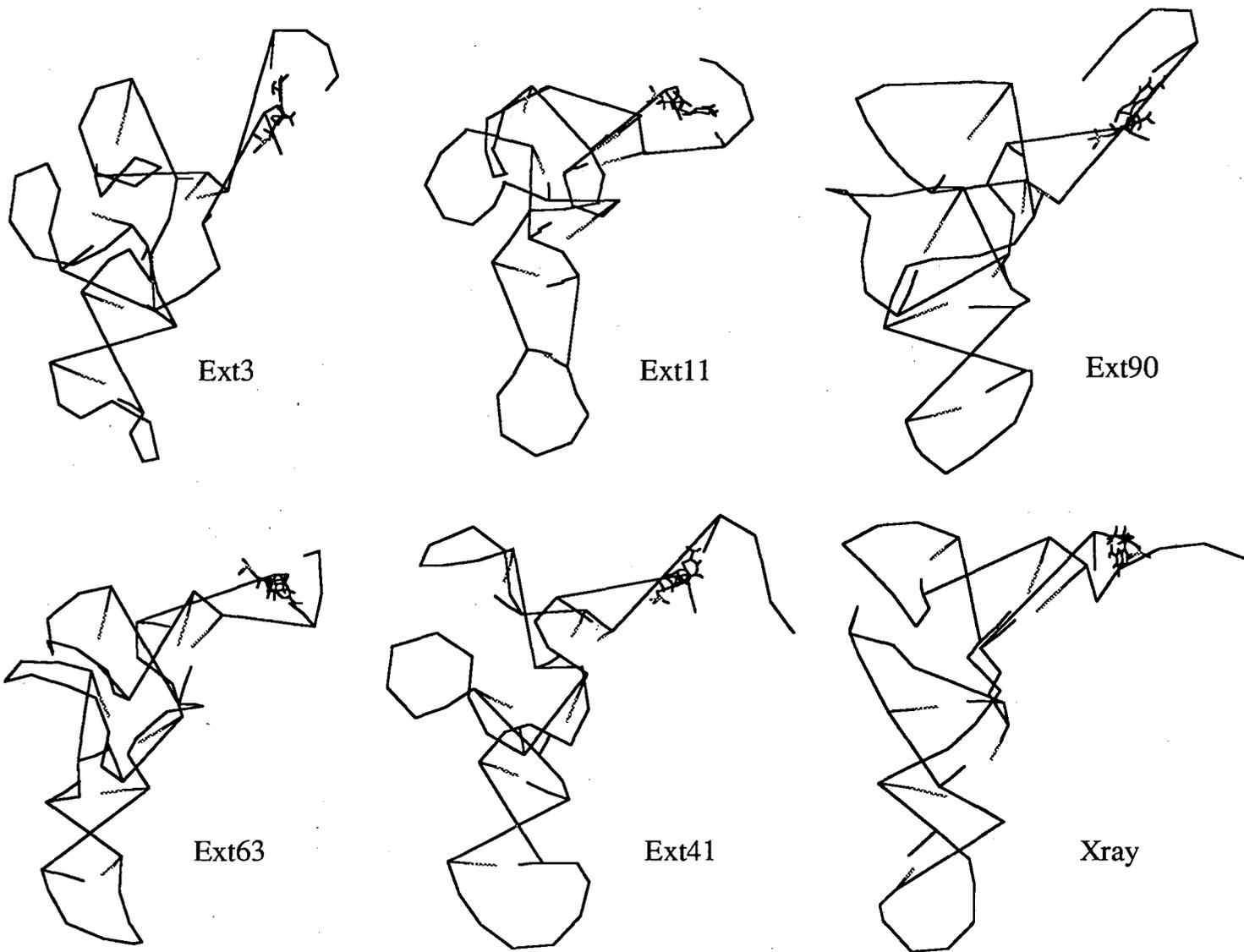


Figure 16. The five best ext models compared to the crystal structure of tRNA.

crystal structure. After the initial 32 cycles of conjugate gradient refinement the error functions for the structures ranged from 136.55 to 645.00 square angstroms while that of the crystal structure is 423.01 square angstroms. The bounds matrix violations for the ext structures varied from 212.51 to 452.82 angstroms as opposed to 345.06 angstroms for the crystal structure. At this point in the refinement process the DSPACE generated structures show similar ranges in the fit of one on another (fit = 6.76 - 10.29 angstroms, rms = 81.34 - 120.34 angstroms) as that seen for superposition on the crystal structure of tRNA (fit = 7.59 - 12.14 angstroms, rms = 85.61 - 135.97 angstroms).

Representative structures:	cgr err.	bounds err.	fit err
ext3	136.55	136.54	10.09
ext11	154.52	217.45	8.43
ext90	164.74	213.88	9.46
ext63	193.22	262.80	8.27
ext41	195.66	212.51	9.10

After 32 more cycles of minimization, the cgr error functions for all structures ranged from 49.82 to 282.36 square angstroms. The comparison of the changes which DSPACE makes to ext11 and the crystal structure given these additional refinement steps shows how further minimization may actually distort the structure away from the ideal (Fig. 17). In particular it is apparent that the loops are being forced into a rounder, less stacked conformation, although the overall superposition error is improved. The more refined version of ext11 has a cgr error function of 49.82 square angstroms, bounds violations totalling 107.75 angstroms, and a superposition fit error of 7.63 angstroms. When refined in the same manner the crystal structure cgr error function is decreased to 33.07 square angstroms, the bounds violations to 82.11 angstroms and it has a superposition error of 1.89 angstroms when compared to the original crystal structure.

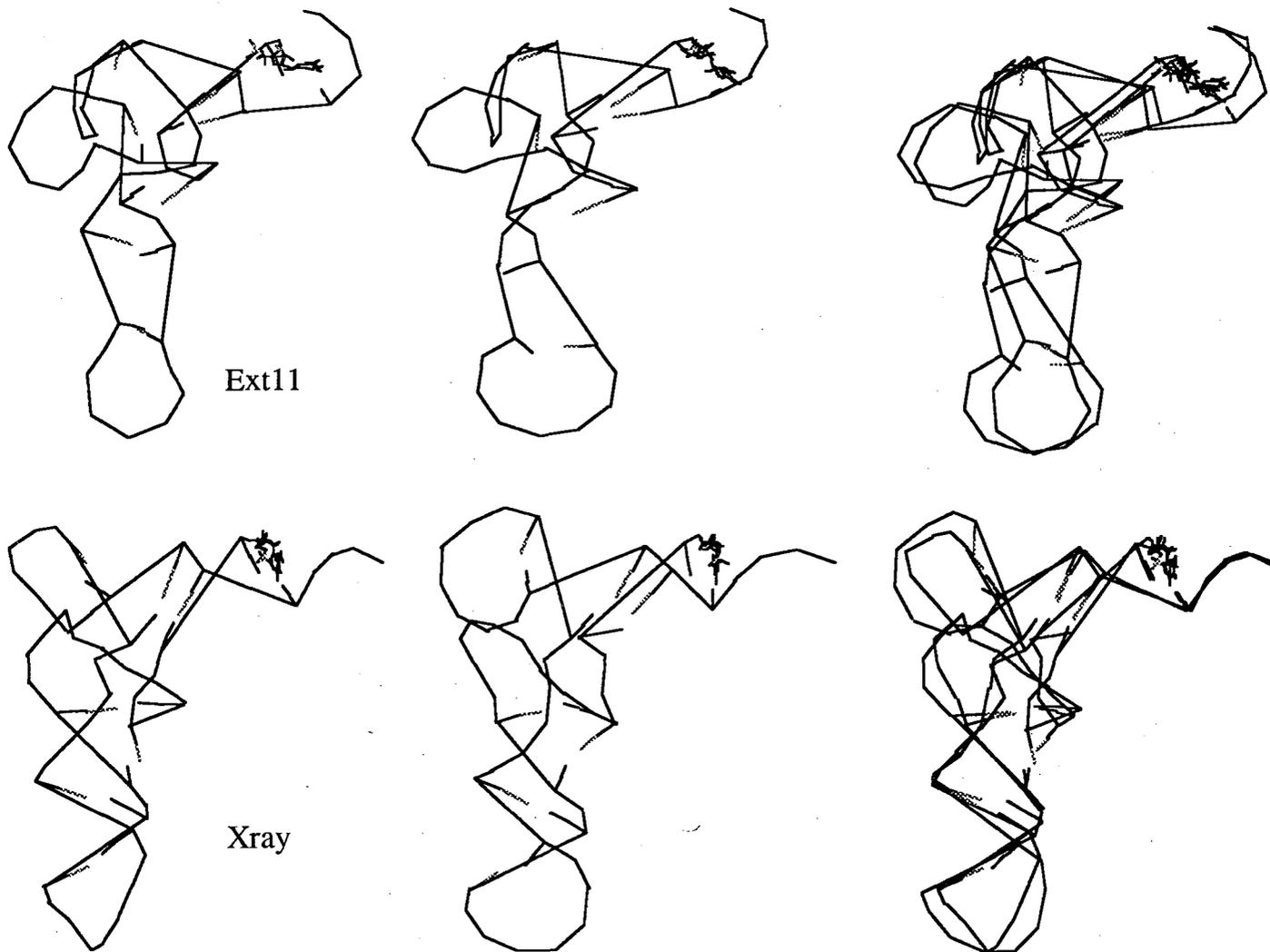
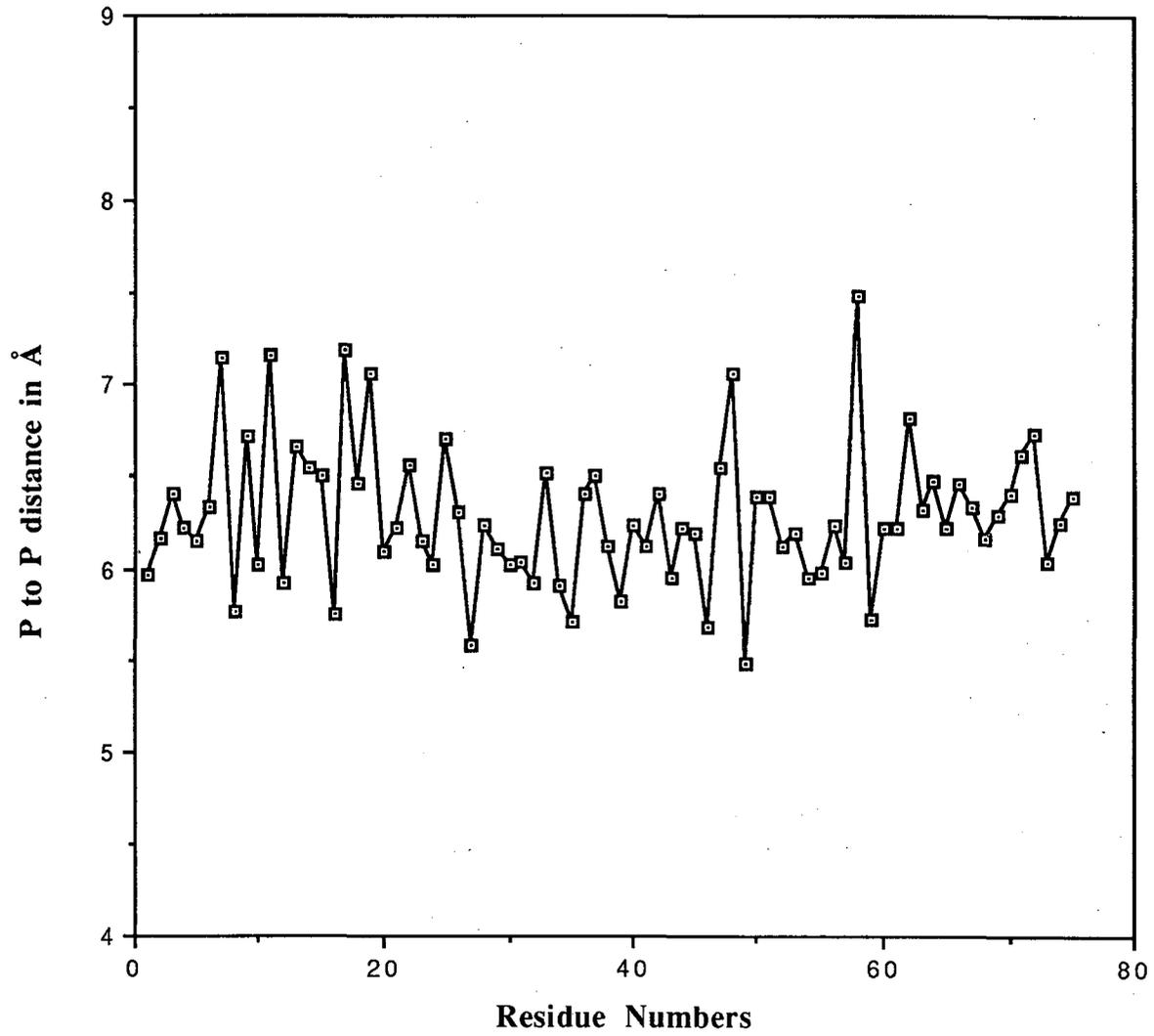


Figure 17. The effects of additional refinement on ext11 and the crystal structure of tRNA.

AMBER results

The crystal structure of yeast phenylalanine transfer RNA obtained from the Brookhaven Protein Data Bank is the standard of comparison used for all structures (Fig. 6). It should be noted that both the AMBER evaluation of the all atom structure and the DSPACE error function of the reduced crystal structure, suggest that there are serious inadequacies in the atomic parameter data bases. This was to be expected with the DSPACE pseudoatoms since there are only six atom types and the phosphate replacement atom assumes a regular, compacted phosphate to phosphate distance of 5.8 angstroms. The 75 phosphate to phosphate backbone distances present in the crystal structure are not uniformly distributed about this mean or throughout the structure. Four phosphate to phosphate lengths exceed seven angstroms and five phosphate to phosphate distances are less than 5 angstroms. Fully extended the phosphate backbone should be able to span at least eight angstroms as is seen in the sharp turn (residues U8 and A9) between the AminioAcid and the DiHydroUracil stems (Fig. 18). When minimized with AMBER the tRNA crystal structure is slightly changed in a manner which suggests the application of the Calladine rules (Calladine, 1982) to the positioning of the bases. The average phosphate to phosphate distance in the minimized version of the structure is 5.97 angstroms. AMBER detects some problems in the crystal structure with the van der Waals forces being very poor initially. But these clashes with the AMBER database are quickly resolved and the atoms which have the largest RMS deviations after the conjugate gradient minimization are the H1' protons which are inserted by the AMBER edit module. The arbitrary conversion of the 14 modified nucleotides found in tRNA into the standard A, C, G, and U residues produces some problems, but the crystal structure is easily adjusted into a conformation

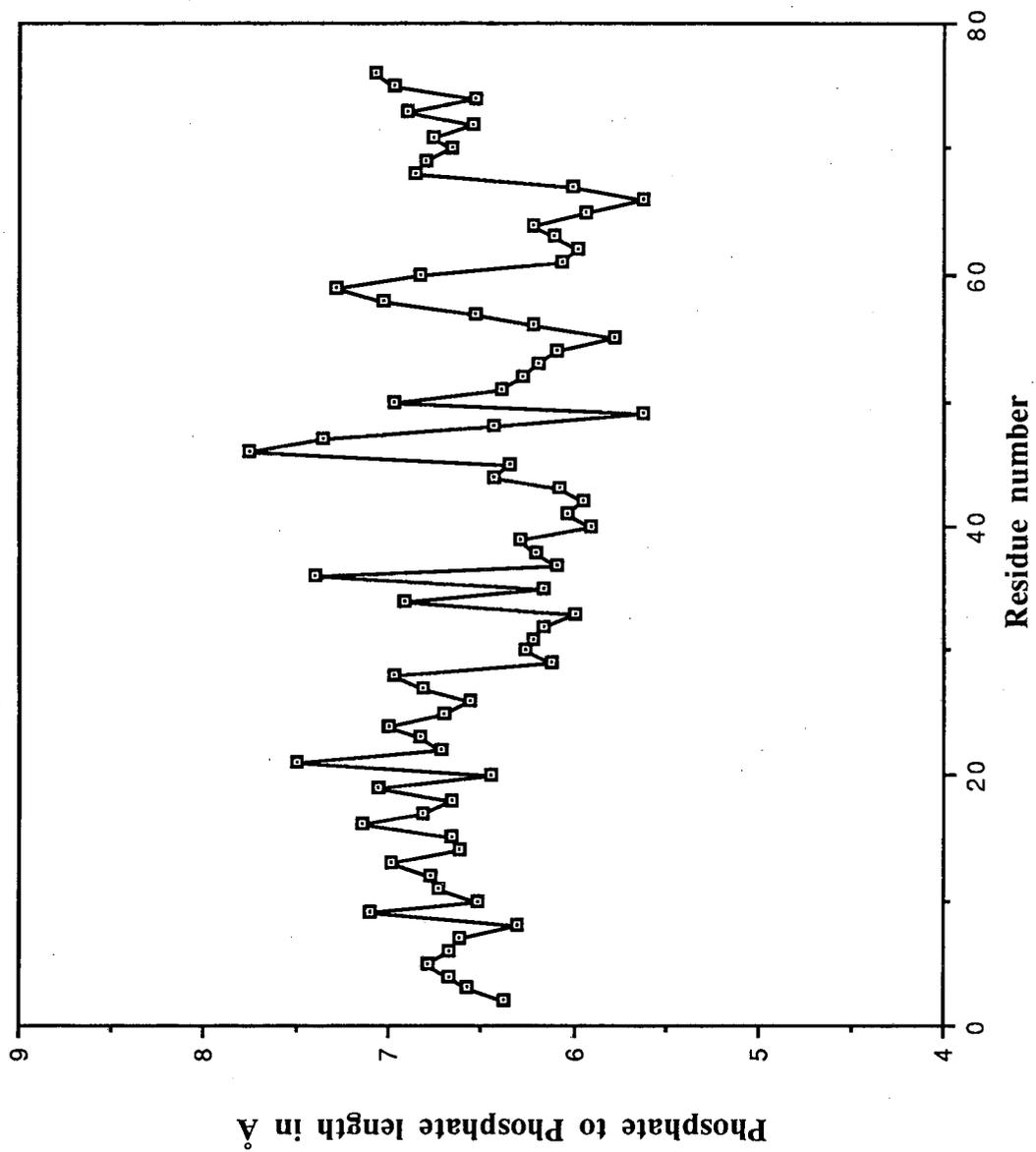
Figure 18. Yeast Phenylalanine tRNA



with a low empirical energy. This analysis of the crystal structure emphasizes that we may not take AMBER modeling at face value. Minimization of a structure means only that we have removed any serious conflicts with the input database and a structure which is totally regular, with ideal bond lengths, angles, etc., is almost certainly not a conformation which exists in solution.

This is clearly demonstrated by the model made from the cloverleaf hydrogen bonding pattern of transfer RNA. The four major stems were built separately from ideal A-form helices. The DHU stem, anticodon stem, and TΨC stem were then arbitrarily bridged with the appropriate loop sequence and then AMBER minimized separately. As was seen in the DNA hairpin modeling, the initial free energies were all poor with some being as large as ten million kilocalories. After several thousand cycles of minimization, the stems were refined to stable conformations with negative free energies of about one kcal. The small sequences which link the stems were generated with AMBER and the complete cloverleaf was constructed by manually docking the idealized fragments. Minimization of the cloverleaf was straightforward and merely regularized the backbone bonding between the pieces. The average phosphate to phosphate distance of 6.55 angstroms is indicative of the more extended sugar-phosphate backbone conformations that are generated by the AMBER structure library (Fig. 19).

Figure 19. tRNA cloverleaf



At least one member of each set of the DSPACE models was EXPANDED and taken through the AMBER minimization. As was seen in the original cloverleaf constructs, all the models had very large initial free energies. Unlike the crystal structure and the cloverleaf minimization was not allowed to proceed until complete convergence had occurred. Examination of the structures at intermediate steps indicated that the gross changes in the structures being introduced by the minimization process. While improving the overall energy balance, simple minimization was unable to improve the conformation beyond the removal of close van der Waals conflicts. The search for more appropriate AMBER parameters was hampered by the large amounts of computer time required. For example AMBER processing of the ext11 structure required more than five days of microVAX computer time. Despite these problems the results are not egregiously bad and suggest that further work or a more experienced user might produce very outstanding models.

Before minimization					
molecule	volume(Å ³)	Rg(Å ²)	E(kcal)	p/p(Å)	
crystal	51600	23.09	+7.31 e6	5.82	
cloverleaf	70000	25.57	+4.71 e2	6.62	
dyn94	76400	26.32	+2.33 e10	5.70	
vt32	52700	23.26	+6.44 e9	5.76	
ext11	41300	21.45	+4.64 e13	5.94	
After minimization					
molecule	volume(Å ³)	Rg(Å ²)	E(kcal)	p/p(Å)	# cycles
crystal	51900	23.14	-2.60	5.97	250
cloverleaf	69900	25.56	-3.36	6.55	200
dyn94	77300	26.43	+1.05	6.14	1760
vt32	54300	23.49	+9.38	6.40	1550
ext11	44100	21.91	+8.24	6.52	1550

Discussion

The adaptation of DSPACE to the folding of larger molecules required an initial expenditure of effort in deciphering the atomic and residue description formats. Once that was accomplished building the pseudoatoms and pseudoresidues was a fairly straightforward task that could be done with a text editor and the printouts of the standard geometries generated with AMBER. Converting the X-ray crystal structure from the PDB format to the XYZ format of DSPACE was done with a simple FORTRAN program (PDBDS). Tailoring the DSPACE program itself for use with large molecules was not difficult but did require the alteration of the FORTRAN source code. With the exception of the van der Waals radii which are specified in a subroutine, most of the arrays and variable sizes are sequestered in a single file (GLOBAL.CMN). The $N \times N$ distance matrix is the major consumer of memory and by making this data structure no bigger than is absolutely necessary, the size of the program is minimized. Some space can be saved by deciding to limit the number of library definitions and molecules that will be used at one time. Finally a normal protein or nucleic acid consists of several groups for each residue. With the special pseudoresidues the number of groups is approximately equal to the number of residues and the size of the array for groups can be reduced accordingly. Adapting DSPACE to the UNIX-style operating environment of the CRAY-XMP did require some debugging but no major revisions. Under the VMS operating system present on all the VAX computers, the program compiled and ran with no problems when sufficient memory was made available.

In the hope that the spacefilling problems could be solved by moving to the all-atom representation, some of the early trial structures were transferred to the next step in the modelling protocol. Transforming the XYZ files generated by DSPACE into the PDB format was accomplished with a FORTRAN program which is the converse of the PDB to XYZ program (DSPDB). Using AMBER to reinsert all the atoms and then minimizing, did not improve the overall conformation of these early DSPACE constructs. Attempts to improve the helicity of the DSPACE folding of tRNA by a narrowing of the H-bond donor

to H-bond acceptor constraints from a three angstrom range to one half angstrom range, did not produce the desired curvature in the tertiary structure. It did improve the quality of the distance matrix and was retained in later runs.

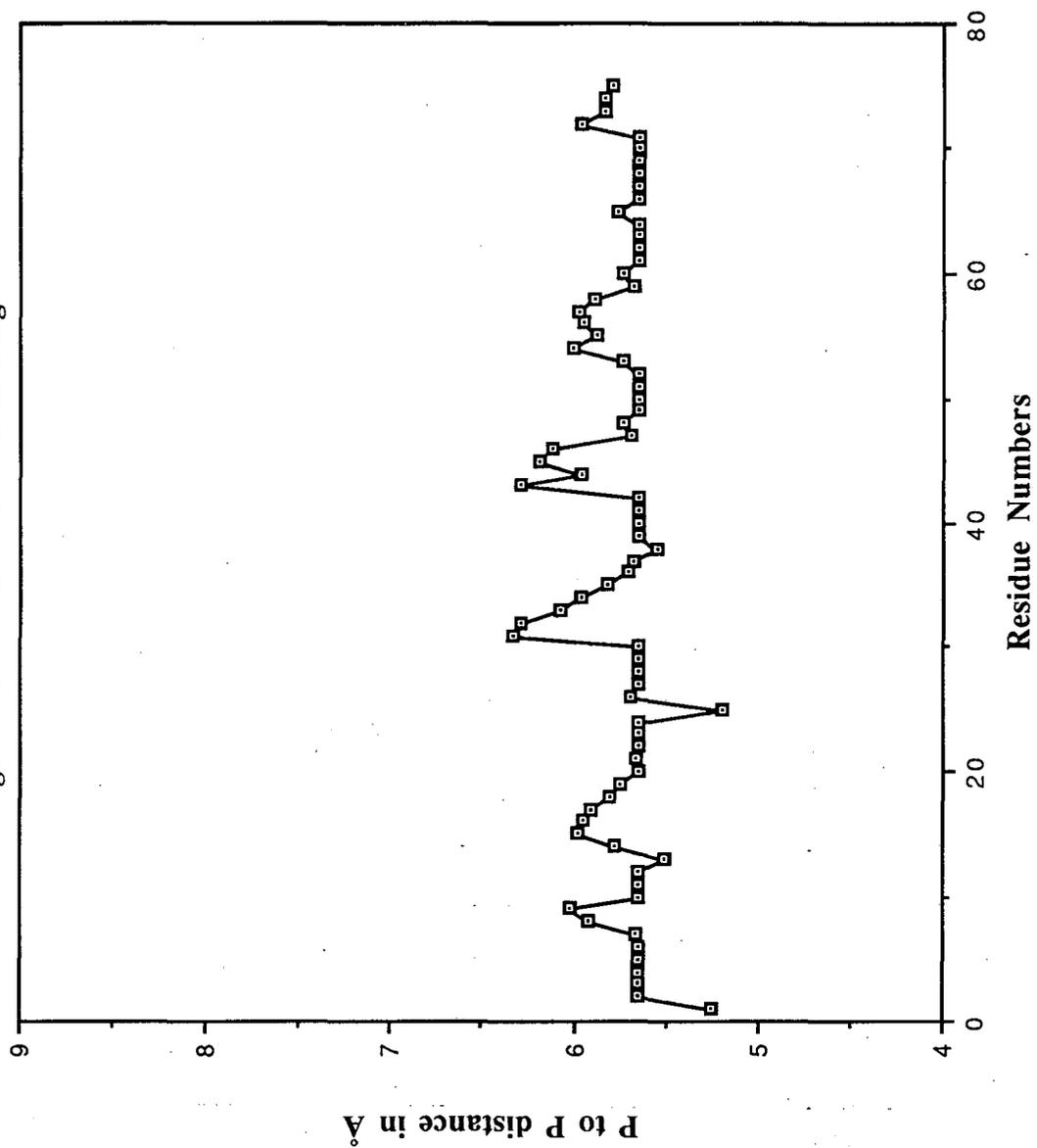
Several alternative pseudoatom representations were explored. More drastic reduction schemes in which only the phosphate or only the phosphate and hydrogen bonding group of each nucleotide were retained, greatly increased the speed at which bad results could be obtained. In this manner various means of enforcing helicity on the double-stranded regions were tried. Adding explicit phosphate to phosphate constraints for hydrogen bonded residues improved the overall structure but did not increase the helicity. Even adding distance constraints for the neighboring residues of a hydrogen bonding partner produced only a gentle winding of the strands. Far too little improvement considering the amount of time it took to create the necessary constraint files. Usually the entire structure is refined by conjugate gradient refinement against the library structures and all distance constraints. As the grouping of residues into helical subunits was insufficient to produce the desired structure through simple refinement, perhaps some preferential treatment would help. Several schemes were attempted which minimized the helical regions in differing orders or groups before the single-stranded regions were minimized. When these efforts failed, simulated annealing of the structures was attempted. By adding small, random amounts to the energies of the atoms throughout the structure, allowing the structure to dynamically adjust, and then reminimizing, it is sometimes possible to escape local minima and reach a better conformation. Finally even including extra distance constraints derived from the crystal structure was tried. All of these attempts failed and seemed to indicate that DSPACE undervalues the significance of the van der Waals radius in establishing a minimum distance constraint. Increasing the penalty function weighting of vdW contact did not significantly affect the overcompaction problem that was apparent in all the 5mer structures.

It is common practice to refine structures generated by distance geometry until they can no longer be minimized. This usually results in a small number of closely related structures, and it is argued that these conformers represent the solution structure. With such drastically reduced pseudoresidues, complete refinement leads to overcompaction and reflects the simplicity of the library residues. Refining in a stepwise manner until the most egregious discrepancies have been eliminated is more revealing and will leave the structures well placed for further minimization with AMBER. Rating the structures is more difficult under this approach. Ordering the structures based on how well they can be superimposed on the crystal structure is an attractive idea, but if the protocol is extended to molecules of unknown structure, this is not acceptable. There is some correlation between the conjugate gradient function and the bounds violation error. Sorting the structures based on either quantity yields similar ratings and can be used interchangeably to look for duplicate structures. The cgr error function has been used because it indicates which structure is most directly poised over its final conformation on the error function hypersurface. The bounds violation function may be skewed by a few large errors which would be easily corrected if refinement were to continue. In practice, the order of the best five or ten structures may vary depending on which value is chosen, but not the members of the preferred set. This stepwise approach and the decision to arbitrarily halt the refinement process far short of convergence, preserves the diversity of solutions that DSPACE finds to the folding problem while revealing the most basic characteristics that the structures share.

The DSPACE foldings prior to any AMBER minimization show that this minimal data set was sufficient to determine the basic 'L' shape. But as was expected for so simple a model, there are some problems. The early 5mer foldings, while having the proper shape, showed two major defects. The compaction which seemed to undervalue van der Waals contacts may have been a result of the minimization procedure. More significant were the deviations in handedness of the helices and the DHU and TPsiC loop stacking orientations. These 'twisting' problems result from the lack of chirality in the pseudoatoms and the low

number of secondary and tertiary constraints in comparison to the total number of pseudoatoms. When judged on the basis of those structural characteristics which correspond to the information represented by the pseudoatoms, it is clear that the predictions of the fully reduced model are more successful than the more explicit model. The overall geometry, size, and shape are consistently predicted. The correct chain path from 5' to 3' is followed, the TpsiC loops and DHU loops are correctly stacked and even the kink in the aminoacid acceptor stem is present. The increase in the number of constraints specifying a helical relationship, combined with a decrease in the total number of pseudoatoms results in much better foldings. This overall improvement is partially negated by the primitive method used to reintroduce ideal helices. It can produce abrupt shifts in backbone geometries that would require extensive structure refinement. Several of the conformers have problems with the positioning of the single-stranded loops but this is to be expected considering the simplicity of the pseudophosphate backbone. The large distances between pseudophosphates was the apparent cause of the knotting of the DHU and TpsiC loops of the ext11 structure. Some of the overcompaction problems are attributable to the extremely asymmetric nature of the pseudoatoms in the replacement helices. When approached at an angle which is perpendicular to the helical axis the phosphate pseudo atom should have a vdW radius which is similar to that of the phosphate ester (2-3 angstroms). This is also the proper contact radius for an approach parallel to the helix from the exterior. But the contact radius from the opposite direction should be forbidden with a vdW radius approaching the length of the helical segment (15-20 angstroms).

Figure 20. vt32 after EXPANDING



The problems resulting from the tight packing of the pseudoresidues by DSPACE and the subsequent superposition of a fuller atomic representation are not easily resolved. AMBER finds serious vdW conflicts in all the DSPACE generated structures. The most common problems occur in the loops of the hairpins where the short phosphate to phosphate backbone geometry of a helical region is forced on a very different conformation. This causes serious base/base clashing in these minimal loops. AMBER minimization results in the rotation of these bases out of the loop interior instead of the lengthening of the backbone geometry just as was seen in the DNA oligonucleotide modeling. This precludes the establishment of hydrophobic base stacking in the loops which could improve the empirical energy balance. Structures generated by DSPACE start off with a much narrower range of distances clustered around the ideal phosphate to phosphate value of approximately six angstroms. The bonds from the 3' end of a superimposed ideal helix to the next residue are most likely to show any significant variation because of the way in which the superposition is achieved. The transformation matrix is compiled by first translating the 5' phosphate onto the 5' pseudophosphate, then ideal helix is rotated until the phosphate to H-bonding group vector is colinear with the 5' pseudophosphate to pseudobase bond. Finally the vector connecting the 5' and 3' phosphates of the ideal helix is rotated about the 5' phosphate to base vector until it lies in the same plane as the 5' pseudophosphate to 3' pseudophosphate bond. Thus any discrepancy between the ideal helix and the pseudohelix structure produced by DSPACE will show up in the phosphate backbone and hydrogen bonding pattern at the 3' junction of a helix (Fig. 20). This is most evident in the distance from residues 72 to 73 of the AA stem. As the adenine at position 73 is a single-stranded residue with only standard bonding constraints, DSPACE will always be able to place it without violating a distance bound. The departure of this phosphate to phosphate distance from 5.9 angstroms can only be the result of superimposing the ideal helix. Even with this source of error, the use of pseudohelices improved the structures of the helical stems over that achieved with the 5mer

model. When these structures are minimized with AMBER local variations are introduced (Fig. 21). But because of the very sharp energy barrier caused by vdW conflicts, unrealistic phosphate to phosphate lengths of less than four angstroms or more than eight angstroms may appear. Even a slight overlap of atoms can produce an unfavorable energy term of a billion calories and the resultant atomic motions used by AMBER to resolve the conflict may create very poor, but less energetically expensive, bond lengths (Fig. 22). It is clear that a more sophisticated approach such as restrained molecular dynamics will be necessary to produce all atom model structures. Perhaps the addition of a different pseudoatom type for single-stranded phosphates with longer phosphate bond lengths and more linear bond angles would help.

Figure 21. vt32 after AMBER minimization

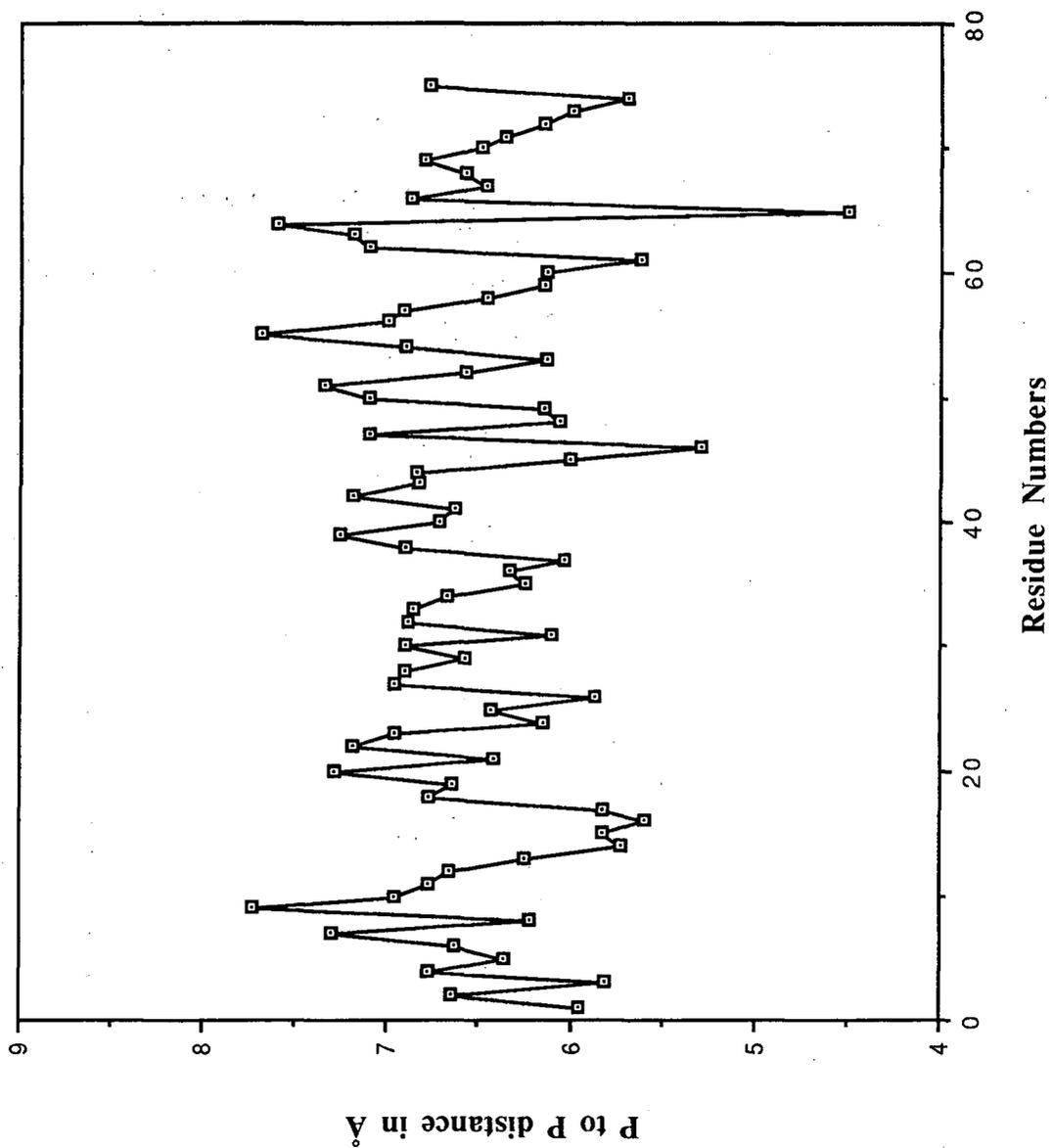
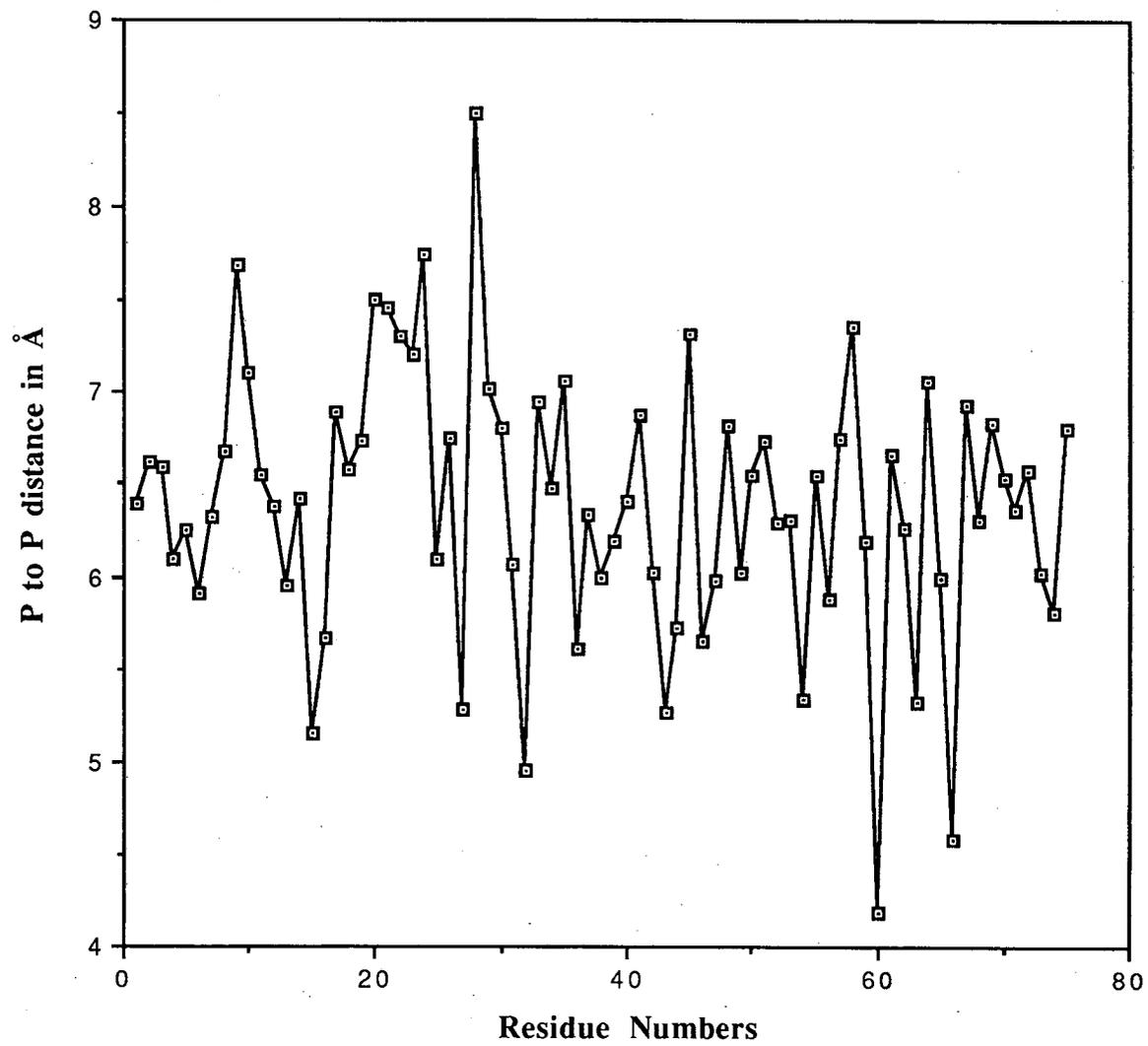


Figure 22. vt32 after extended AMBER minimization



During the 5mer exploratory runs a distance geometry matrix was constructed and structure created where by accident the directory specification of the helical and long range distance constraints was wrong. The result was an extended, single-stranded helix. At that time DSPACE cgr minimization was allowed to proceed until the error function between two steps had converged to less than 0.1. This result is mentioned because it illustrates two important points. As in structures produced by AMBER, minimizing until convergence is achieved will produce structures which reflect the nature of the reference structures. As the reduced residues are derived from the Arnott parameters for A-form helices, even single-stranded residues will be forced into a helix given sufficient minimization. Thus care must be exercised in using the DSPACE cgr routines to minimize intermediate structures. That's the bad news. The positive aspect lies in the lack of folding produced by this mistake. Lacking explicit distance constraints DSPACE will not produce a compact structure. Just as the preferred bond angles derived from the A-form helix will eventually reproduce a helix, the dihedral angles derived from the pseudoatoms will tend to produce extended conformations. As the initial structures produced by DSPACE embedding are even more extended, with some bond lengths exceeding ten angstroms (Fig. 23), the models will

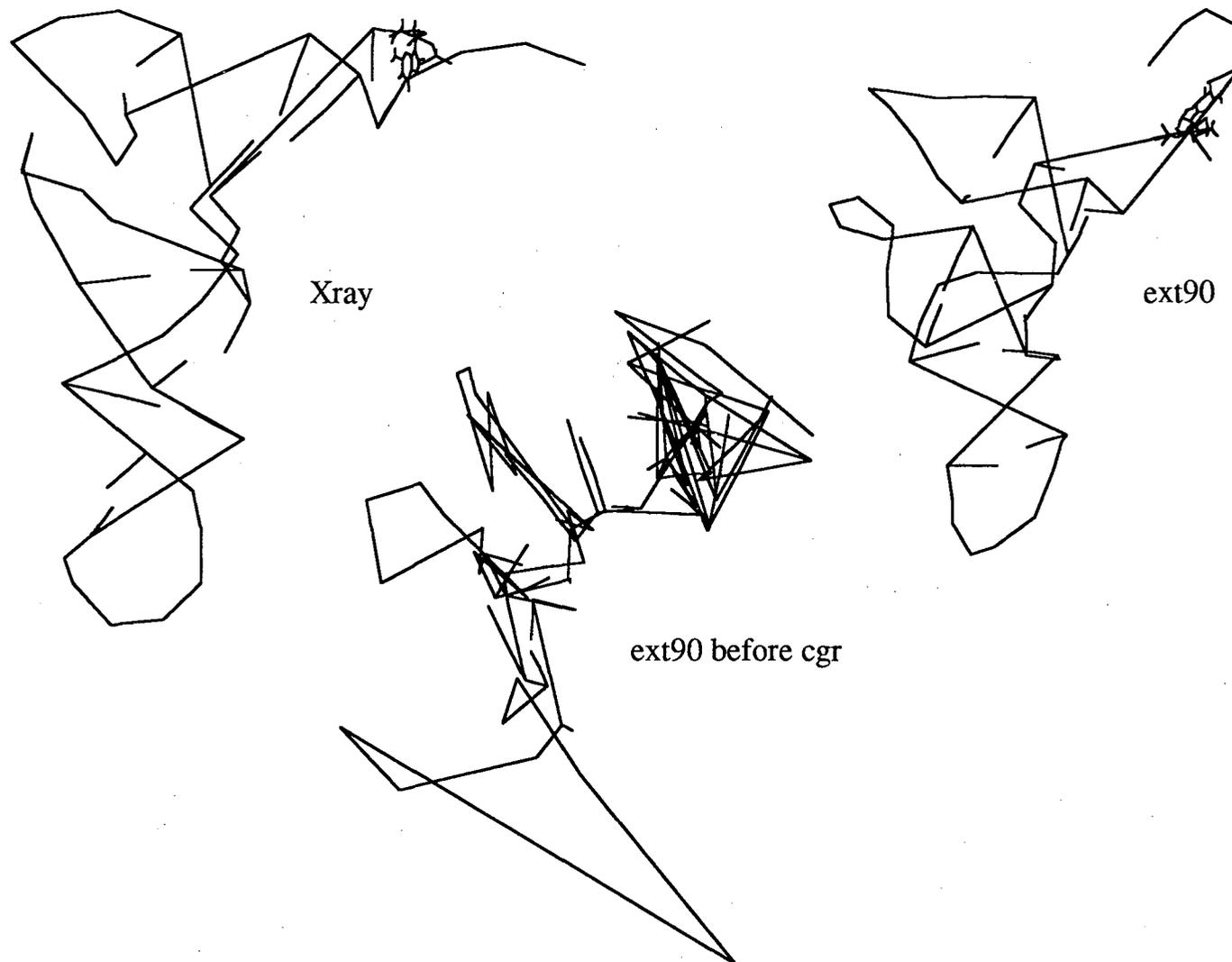


Figure 23. The unrefined and final versions of ext90 compared to the crystal structure.

approach the correct structure from an area of conformational space which is larger in volume and less ordered. Therefore any structure which is consistently produced should be characteristic of the molecule being modeled. This gradual introduction of structure may also resemble the folding of RNA in solution. It is well known that in low ionic solution, nucleic acids form extended single-stranded tangles due to the strong electronic repulsion of the negatively charged phosphate backbone. It has also been demonstrated that divalent cations or polycationic proteins are required for the proper compact folding of some nucleic acids (Schimmel & Redfield, 1980). By adding in the secondary and tertiary distance constraints in stages, we may be able to study the general folding process. As a check of this idea tRNA foldings were performed with no secondary or tertiary constraints using the helical pseudoatoms and once again an extended coil was obtained. When just the helical constraints are included, the resulting model strongly resembles the classical cloverleaf (Fig. 24).

Recently, some researchers (Metzler et al., 1989) have been highly critical of the manner in which distance geometry samples conformational space. Continued cgr minimization of the DSPACE structures will produce a very narrow range of conformations which are all highly similar but different from that found by X-ray crystallography. This

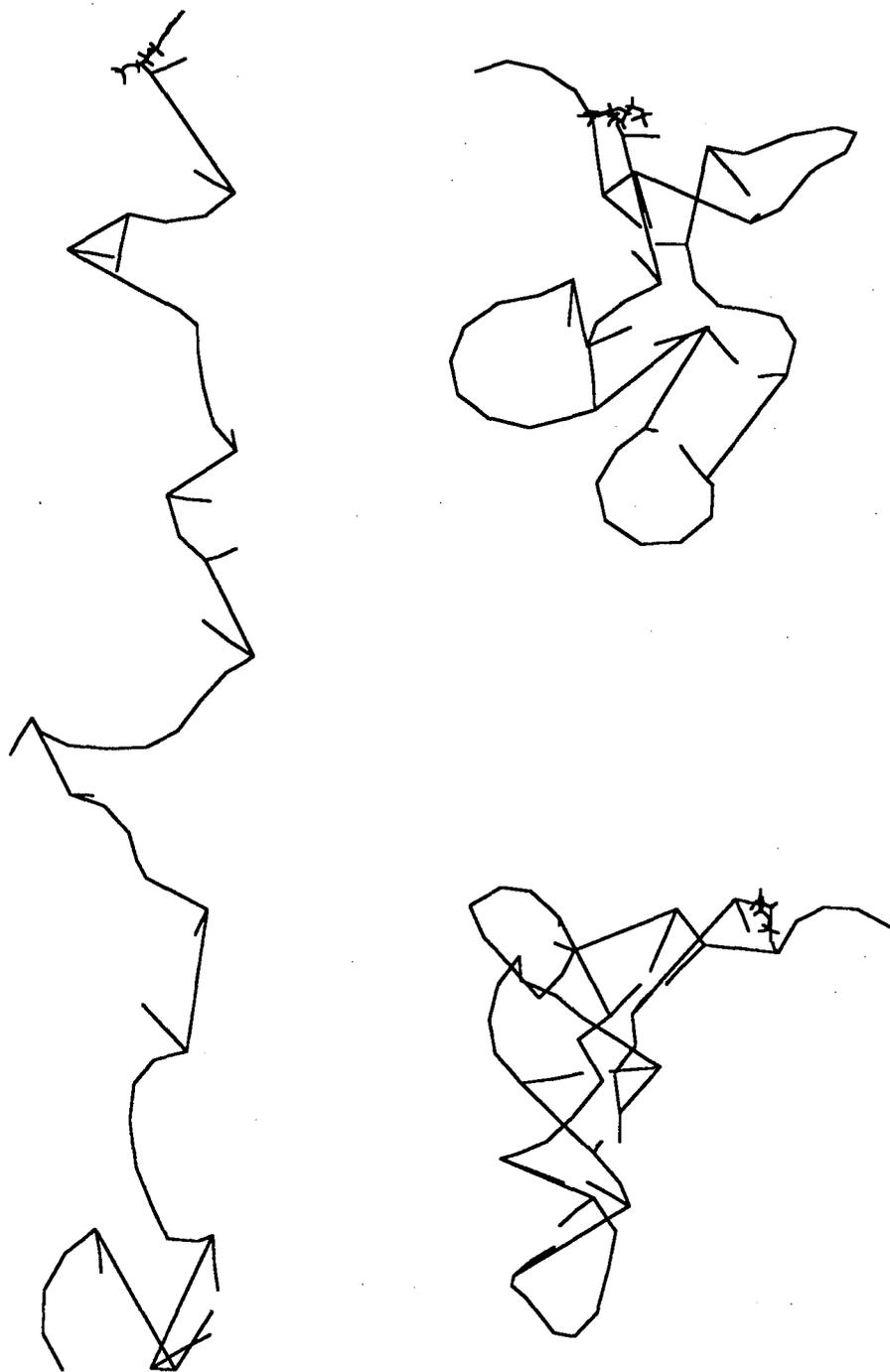


Figure 24. The structures produced when only the primary or primary and secondary structure are used to fold tRNA compared to the crystal structure.

criticism is not particularly relevant to the models created by this protocol. The results should differ from the crystal structure since a full atomic representation is not used. Even an all atom model should fail to reproduce the accepted structure since all computational methods use average empirical approximations of the atomic building blocks. Beyond these shortcomings of modeling constructs, the structure which a molecule has in a crystalline environment is not necessarily the same as that which it will take in a different environment. This is especially true when biological molecules form an active complex as shown by the recent tRNA/tRNA synthetase cocrystal (Rould et al., 1989). An analyzable crystal is necessarily well ordered and cannot reflect the myriad equivalent conformers of a molecule which coexist in solution. Despite all these unavoidable complications, this series of DSPACE runs consistently finds solutions to the tRNA folding problem which are in the proper global conformational domain.

Using physical modeling techniques and similar data sets, no researcher was able to successfully model transfer RNA. Levitt complained at the time that his attempts were hampered by the weight and size of the CPK models he employed (Levitt, 1969). With modern computer technology we can avoid the problems of gravity and space encountered with physical models. By using distance geometry to fold the molecule we eliminate operator subjectivity and can reproducibly predict the overall dimensions and shape of an RNA molecule. We can even determine the general path of the backbone conformation. But prediction of atomic resolution coordinates from first principles is not yet possible at the present level of abstraction. This study shows that the major determinants of the size and shape of tRNA are inherent to its basic primary and secondary structure. With the addition of only a few tertiary interactions we can begin to probe the rules governing the formation of three dimensional structures. To the extent that this is a general rule in RNA structure, we should be able to use the same approach to model other RNA molecules.

References

- Arnott, S., Hukins, D.W.L., Dover, S.D., Fuller, W., & Hudgson, A.R. (1973) *Journal of Molecular Biology* 81, 107-122.
- Calladine, C.R. (1982) *Journal of Molecular Biology* 161, 343-352.
- Dock-Bregeon, A.C., Dhevriere, B., Podjarny, A., Johnson, J., de Bear, J.S., Gough, G.R., Gilham, P.T., & Moras, D. (1989) *Journal of Molecular Biology* 209, 459-474.
- Garrett-Wheeler, E., Lockard, R.E., & Kumar, A. (1984) *Nucleic Acids Research* 12, 3405-3423.
- Haselman, T., Camp, D.G., & Fox, G.E. (1989) *Nucleic Acids Research* 17, 2215-2221.
- Jaeger, J.A., Turner, D.H., & Zuker, M. (1989) *Proceedings of the National Academy of Science* 86, 7706-7710.
- Kim, S.H., Quigley, G.J., Suddath, F.L., McPherson, A., Sneden, D., Kim, J.J., Weinzierl, J., & Rich, J. (1973) *Science* 179, 285-288.
- Levitt, Michael (1969) *Nature* 224, 759-763.
- Metzler, W.J., Hare, D.R., & Pardi, A. (1989) *Biochemistry* 28, 7045-7052.
- inio, J., Favre, A., & Yaniv, M. (1969) *Nature* 223, 1333-1335.
- Rould, M.A., Perona, J.J., Soell, D., & Steitz, T.A. (1989) *Science* 246, 1135-1142.
- Sampson, J.R., DiRenzo, A.B., Behlen, L.S., & Uhlenbeck, O.C. (1989) *Science* 243, 1363-1366.
- Saenger, Wolfram (1984) *Principles of Nucleic Acid Structure*. (Cantor, C. Ed.) Springer-Verlag, New York.
- Sprinzi, M., Hartmann, T., Weber, J., Blank, J., & Zeidler, R. (1989) *Nucleic Acids Research supplement* 17, 1-172.
- Srinivasan, A.R. & Olson, W.K. (1987) *Journal of Biomolecular Structure & Dynamics* 4, 895-938.

Sussman, J.L., Holbrook, S.R., Warrant, R.W., Church, G.M., & Kim, S.-H. (1978)

Journal of Molecular Biology 123, 607-630.

Tinoco, I., Jr., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers,

D.M. & Gralla, J. (1973) Nature(London) New Biology 246, 40-41.

Woo, N.H., Roe, B.A., & Rich, A. (1980) Nature 286, 346-351.

Wuethrich, K. (1989) Science 243, 45-50.

Yaniv, M., Favre, A., & Barrell, B.G. (1969) Nature 223, 1331-1333.

Zachau, H.G., Dutting, D., Feldman, H., Melchers, F. & Karau, W. (1966) Cold Spring Harbor Symposia Quantitative Biology 31, 417-424.

Chapter 5

16S RIBOSOMAL RNA MODELING

Introduction

Beyond the importance of the ribosomal RNAs in protein synthesis is the wider and more general significance of RNA structure in general. The ability to place functional groups and ligands in specific micro-environments is the basis of protein chemistry and it appears that catalytic RNAs operate in the same manner. Deciphering the catalytic functions of RNA will be important in the study of retroviruses like the human immunodeficiency virus that causes AIDS. The self-processing of RNA has even suggested to some scientists that the earliest complex lifeforms were RNA polymers and that RNA chemistry played an indispensable role in evolution. As it is the folded forms which do the work, knowing the three dimensional structures that RNAs can form will be essential for understanding how they function.

There is a large body of evidence which shows that 16S ribosomal RNA is intimately and inextricably involved in ribosomal function. An early study of the ribosome demonstrated that both the 16S and the 23S RNAs are cleaved by ribonuclease, even when they are part of the 70S *E. coli* ribosome (Santer, 1963). Santer believed that the ribosomal RNAs were accessible for a reason and suggested that they may play roles in messenger RNA binding and association of the ribosomal subunits. The importance of ribosomal rRNA is further underlined by the maintenance of 7 copies of the rRNA genes in *E. coli* (Nomura, 1987). 16S rRNA is involved in the proper positioning of the mRNA through interaction with the Shine-Delgarno sequence and appears to have a role in maintaining the proper reading frame during translation (Weiss et al., 1987). The general premise is that the small subunit ribosomal proteins are important for assembly and stability but play a peripheral role in translation. This viewpoint is supported by reconstitution experiments, including heterologous reconstitution (Higo et al., 1973). and mutation experiments which at least in the case of the small subunit, have been unable to isolate a single instance where a protein is directly and indispensably involved in translation (Nomura, 1987).

In contrast there are many instances in which a change in the ribosomal RNA has a profound effect on ribosomal function. Colicin removes the last 50 bases of 16S rRNA and abolishes all translation (Bowman et al., 1971). Mild treatment with kethoxal methylates six guanosines of 16S and thereby prevents the binding of aminoacylated tRNAs (Noller & Chaires, 1972). The 3' end of 16S can also be photochemically crosslinked to the transfer RNA which carries the growing peptide chain (Prince et al., 1982). Kasugamycin blocks the initiation of translation and a resistant mutant lacks the enzyme responsible for posttranscriptional methylation of A1518 and A1519 (Helser et al., 1972). It has also been shown the resistance of ribosomes to kanamycin plus gentamycin or kanamycin plus apramycin is achieved in mutant *E. coli* by methylation of G1405 or A1408 (Beauclerk & Cundliffe, 1987). Changes in the sequence of the nominally single-stranded loop near residue 530 affect translational fidelity and can produce resistance to streptomycin (Melancon et al., 1988). Deletion experiments have demonstrated that G1401 is absolutely necessary for proper ribosomal functioning (Denman et al., 1989).

The primary sequence of *E. coli* 16S ribosomal RNA was determined over the course of many years by laborious RNA sequencing techniques (Brosius et al., 1978). The hydrogen bonding pattern was established by a combination of equally difficult means and this basepairing map was confirmed by single-strand and double-strand specific modification experiments (Noller & Woese, 1981). Helices in ribosomal RNA's are resistant to reagents like kethoxal and RNase S which attack single-stranded RNA's but can be cleaved with the double-strand specific enzyme, Cobra Venom nuclease. By transforming RNA into DNA with reverse transcriptase or by locating the ribosomal RNA genes, DNA sequencing technology can be used to rapidly expand the catalog of rRNAs. 106 small subunit ribosomal RNA sequences from different species are now known and can be aligned with the secondary structure map. Although the size of the ribosomal RNAs may vary by more than a thousand nucleotides in different species, base additions or deletions are confined to nine regions of the sequence, one of which is unique to eucaryotes

(Dams et al., 1988). In addition to the common hydrogen bonding map, there are as many as 360 absolutely conserved bases in the primary structure of small subunit RNAs (Gutell et al., 1985). These sequence homologies are the basis for the evolutionary trees used by molecular geneticists to interrelate all lifeforms and were the primary argument which underlay the reclassification of life into eucaryotes, eubacteria, and archebacteria (Woese & Fox, 1977). These conserved bases must have vital roles in determining the tertiary structure of 16S, or in subunit association, or perhaps even a direct role in translation (Dahlberg, 1989). As an absolutely essential element in the transition of information from nucleic acids to protein, the ribosome must be very stable with respect to random evolutionary change. If we had the three dimensional structure of the small subunit RNA in hand, we could use it as the basis for interspecies comparison and the construction of more significant phylogenetic trees.

16S rRNA is large enough to be directly imaged with an electron microscope and some researchers report a shape similar to that of the 30S particle (Vasiliev et al., 1986). Statistical reconstruction of the 30S subunit structure provides a confining upperbound on the size and conformation of the RNA (Verschoor et al., 1984). In solutions of low ionic strength, 16S RNA has an estimated radius of gyration of 160-190 angstroms (Folkhard et al., 1975). When placed in a buffered salt solution appropriate for reconstitution of the ribosome, 16S RNA has a radius of gyration of 84-86 angstroms. A significant compaction of the RNA occurs on the addition of a subset of the small subunit proteins reducing the radius of gyration to 70 angstroms. A decrease in the hypochromicity of 16S RNA during this stepwise reconstitution suggests that there is a decrease in RNA secondary structure despite this compaction. The X-ray scattering profiles indicate that the RNA maintains a similar overall shape under all conditions (Serdyuk, 1983). Inside the assembled 30S particle, 16S rRNA has a radius of gyration of 66 angstroms (Ramakrishnan, 1986) and neutron scattering measurements indicate that the radius of gyration of 16S is 60 to 65 angstroms in the assembled subunit (Capel et al., 1988). Attempts to directly probe 16S

RNA for tertiary structure have been made with a variety of chemical crosslinkers (Noller et al., 1987). Immune electron microscopy and RNA/protein crosslinking have been used to look for clues to placement of the RNA within the overall shape of the 30S subunit. And methods have been developed to rapidly probe for tertiary interactions through footprinting and chemical protection of 16S RNA. Combining these clues into a coherent whole will begin to resolve the structure of the ribosome.

The importance of the structure of the ribosome and the increasing amount of data about 16S rRNA structure have lead several research groups to construct models of the system. Both physical and computer models have been constructed (Expert-Bezancon & Wollenzien, 1985; Brimacombe et al., 1988; Stern et al., 1988). The models vary in important aspects and it is difficult to compare them. The size of the problem and the subtle nature of human judgements concerning the weight to be given differing data sets, compound the differences and the contentiousness with which the models are discussed. Due to the stature of the authorities involved, some people may conclude that the structure of the small subunit is essentially resolved and that only the details need to be agreed upon. A close reading of the model descriptions reveals that there are serious conflicts among the various data sets. It may be that the ribosome is a highly mobile complex that can assume a number of related but distinct conformations which cannot be described with a single static model. Computer modeling is better suited to dealing with this potential complication, but to the extent that it is dependent on subjective human manipulation, it may merely mask the problems.

Using the helical substructures developed in modeling tRNA it is now possible to objectively produce folded 16S rRNA models. These well-defined computer models also make it possible to develop a quantitative method for comparing models made by other means. Finally by developing a common format for displaying the models, it will be possible for researchers who have not devoted themselves to ribosomal modeling to comprehend and make use of the information which the models embody.

Materials and Methods

Materials - Software

The distance geometry program, DSPACE (versions 1.3 and 2.1), written by Dennis Hare and Robert Morrison, was obtained from Hare Research, Inc. The molecular modeling package of programs, AMBER version 3.0, was obtained from Peter Kollman of UCSF. The graphical interface, GRAMPS written by T.J. O'Donnell, was used for black and white vector modeling. The raster modeling programs written by Michael Connolly were obtained from the Scripps Research Institute in San Diego. Various format conversion programs and the program, EXPAND, which replaces pseudohelical constructs with ideal A-form helices as it converts DSPACE files to PDB files, were written in FORTRAN.

Materials - Hardware

The calculations were performed on a specially configured modified microVAX running the VMS operating system. 10 MBytes of main memory were supplemented with 55 MBytes of virtual memory sequestered from the 350 MBytes of magnetic disk storage. Interactive graphics were done on an Evans & Sutherland MultiPicture System connected to the microVAX for the early models. Less interactive but colorized vector displays drawn by DSPACE were done on a Macintosh II emulating a Tektronix 4105 display. The raster graphics were displayed on an E&S PS340 connected to a VAX 11/785.

Materials - Data

Primary Structure

The primary sequence of 16S ribosomal RNA from *E. coli* consists of approximately 50,000 atoms in 1542 nucleotides (Brosius et al., 1978). Regular RNA bases have been used as none of the ten modified nucleotides which are normally found in 16S rRNA are required for proper ribosomal functioning in ribosome assembly and protein translation assays (Krzyzosiack et al., 1987).

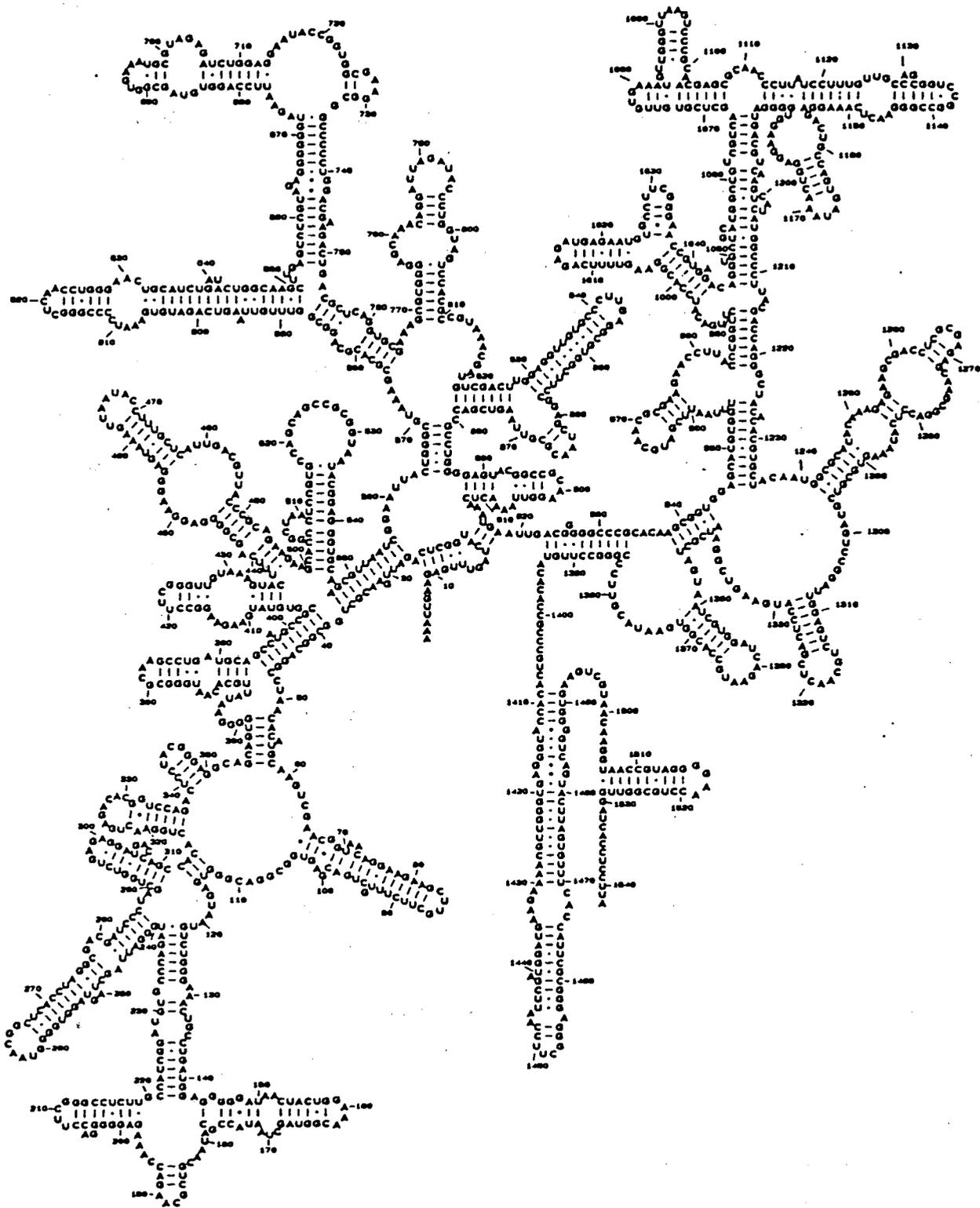


Figure 1. The secondary structure map of 16S ribosomal RNA.

Secondary Structure

Of the 1542 residues found in 16S rRNA 782 are basepaired. As drawn by Noller, the secondary structure of 16S rRNA forms three distinct domains (Gutell et al., 1985). Residues 1-560 form the 5' domain, residues 561-920 form the middle domain, and residues 921-1542 form the 3' domain (Fig. 1). Residues 1400-1542 are sometimes listed as a subregion of the third domain or as a fourth domain. The junctions between these domains are fairly close together in the flat representation centered about the pseudoknot (bases 17-19:916-918). The structure of 16S is too large and complex to attempt to further divide it into subunits which can be uniquely named on the basis of local characteristics as was possible with tRNA. Even numbering the helices is not simple because the secondary map is highly branched and some helices are made from as many as seven different sections of the primary structure. Several of the helices also vary greatly in size, being many basepairs longer than E.coli 16S rRNA in eucaryotes and disappearing altogether in other bacteria. As the molecule will be reduced to a collection of nonbasepairing pseudophosphates and pseudohelices representing one strand of a basepaired helix, the pseudohelices were created and named as they are encountered along the primary structure. The first helical region encountered is named s1h1 for segment1, helix1, the second region is s1h2. The third helical region is divided into s1h3, s1h4, and s1h5 by the bulge guanosines G31 and G38. A traversal of 16S from 5' to 3' yields 51 helical segments in the 5' domain, 30 helical segments in the middle domain, and 50 helical segments in the 3' domain (Fig. 2). Segments of less than three uninterrupted bases which are used to extend helical runs are not converted into pseudohelices. They are retained as pseudophosphates as their double-strandedness is marginal and representing them as helices would greatly increase the number pseudoatoms.

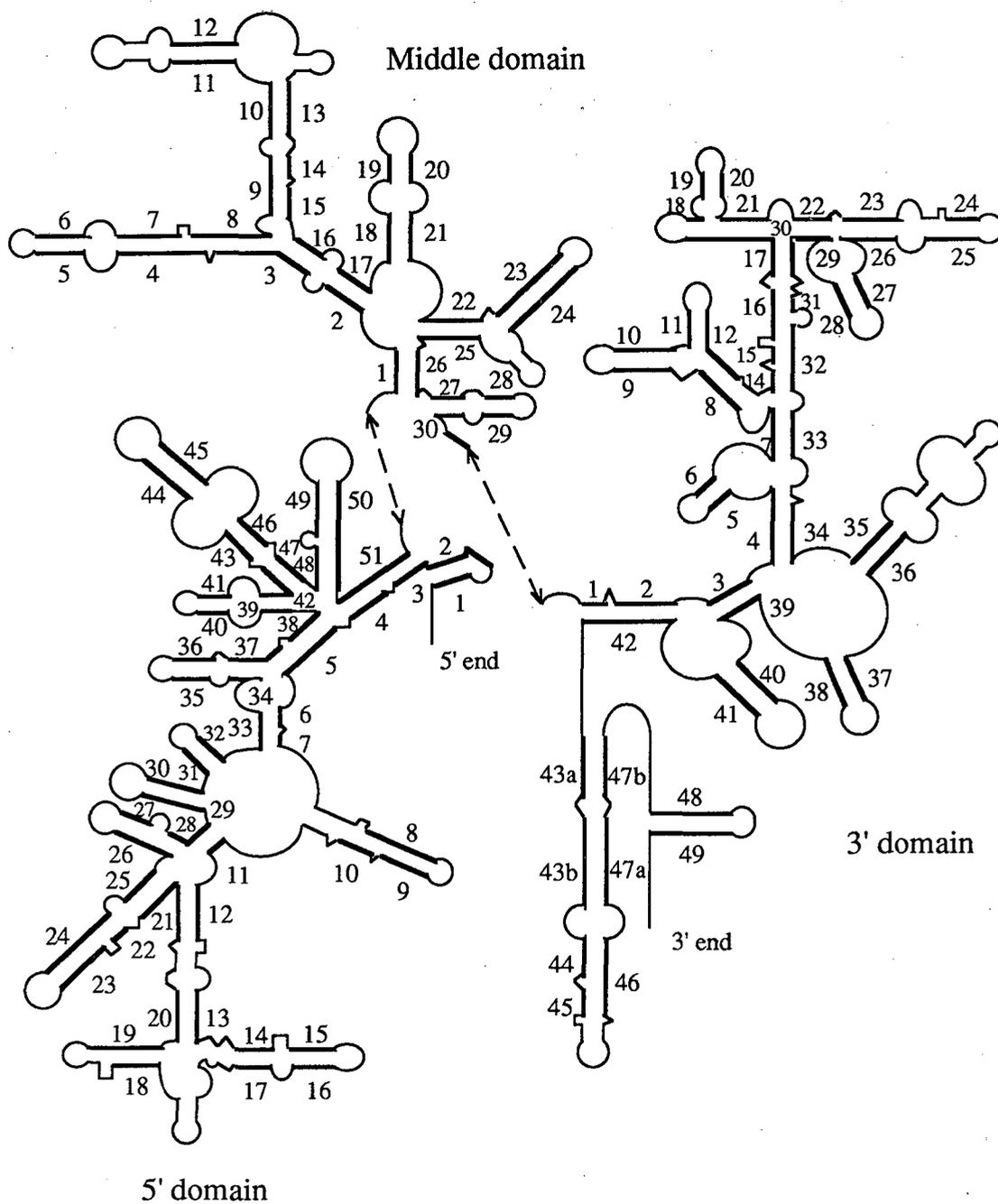


Figure 2. The secondary structure map of 16S RNA divided into domains and the helical segments numbered to reflect their pseudohelical counterparts.

The phylogenetic analysis has been extended in a manner similar to that used to search for tertiary interactions, to include some segments of only two or three basepairs in length. Also included are some helices which contain unusual G:A basepairs presumably of the sort seen in DNA oligomers (Kan et al., 1983). These helices were not included initially because a three basepair stack should be only transitorily associated as its melting temperature will be at or below physiological ones.

The first 50 bases of 16S rRNA are converted to reduced residues in the following manner:

Base 1 - The 5' adenosine is not converted to a pseudoresidue in order to facilitate reimposition of the all atom format for the whole molecule. (residue 1)

Bases 2-8 - These single-stranded bases are reduced to pseudophosphates. (residues 2-8)

Bases 9-13 - These bases form the 5' strand of the first helix. Only the pseudophosphates and hydrogen-bonding groups of G9 and U13 are retained. The reference helix S1H1.PDB will be superimposed on the helical skeleton by EXPAND when the DSPACE XYZ file is transformed into a PDB file. (residue 9 named s1h1)

Bases 14-16 - Single-stranded bases reduced to pseudophosphates. (residues 10-12)

Bases 17-19 - The 5' half of the pseudoknot is treated as a weak helix. (residue 13 named w10a)

Base 20 - This single-stranded uridine is reduced to xura. (residue 14)

Bases 21-25 - The 3' half of the first helix. (residue 15 named s1h2)

Base 26 - Another single-stranded adenosine. (residue 16)

Bases 27-30 - Double-stranded bases. (residue 17 named s1h3)

Base 31 - A bulge guanosine. (residue 18)

Bases 32-37 - Double-stranded bases. (residue 19 named s1h4)

Base 38 - Another bulged guanosine. (residue 20)

Bases 39-47 - A smooth helical stack interrupted by a bulge on the opposite strand. As usual the phosphates and Hbonders at each end are retained. The pseudophosphates and Hbonders for bases 44 and 45 are also included in the helical pseudoresidue definition as referents for the ends of the facing helical strands. Reference helices S1H5.PDB and S1H5.PDC will be used by EXPAND to fill in the helix. (residue 21 named s1h5)

Bases 48-50 - Single-stranded residues. (residues 22-24)

There are two irregularities in the naming of the 3' domain. An extra helical segment was accidentally included in the 1030 to 1080 region and rather than rename all the subsequent helices s3h13 was simply deleted. The long helix formed by the basepairing of 1409-1430 and 1470-1491 includes three G:A associations, two of them next to each other and adjacent to a G:U basepair. To reflect the fact that this region has marginal stability and is certainly not a regular A-form helix, a junction was introduced into the middle of this region and the 5' and 3' halves of each helical segment (s3h43 and s3h47) were designated as a and b respectively.

Tertiary Structure

The much greater size of 16S as compared to transfer RNA and the presence of proteins in the 30S subunit complicates both the determination of crosslinks and their interpretation. Data obtained from protein-free solutions of 16S rRNA can be safely assumed to be derived from a homogeneous population of closely related conformers. Great care must be used in considering data derived from 30S subunits. Unless a significant excess of each protein species is used, it may be possible to form uneven populations of complete and incomplete subunits. In the presence of excess protein,

nonspecific binding to the ribosomal RNA can produce a particle with detectably different physical properties (Ramakrishnan, 1986). Additionally it is possible to form both an active and an inactive 30S particle from the same starting materials by adjusting the mixing conditions (Ericson & Wollenzien, 1989). Furthermore a large part of the data on the small subunit is qualitative in nature, locating a protein or RNA sequence in the top or bottom of the structure. Only slightly more precise are the studies in which electron micrographs of antibodies bound to an antigenic structure are used to create a surface map of the small subunit. Because it is very difficult to derive meaningful quantitative data from most protein work and since this is primarily an RNA modeling protocol, only quantitative RNA/RNA tertiary data will be used to construct models of 16S rRNA and the protein data will be used to evaluate the results in the final chapter.

Phylogenetic Relationships

The rapid expansion of the number of known small subunit RNA sequences and the improvement in the computer algorithms used to correlate these sequences have lead to the prediction that some residues will be involved in highly conserved tertiary basepairing like that seen in tRNA. The phylogenetic interactions may not be canonical Watson/Crick basepairs but the spatial separation of the phosphates should be similar. Consequently a lower bound of 17.5 angstroms and an upper bound of 20.5 angstroms have been used for the distance between residues 9:507, 570:866, 673:717, 921:1396, 1398:1406, 1399:1405, 1399:1504, 1401:1501, and 1405:1496 (Gutell et al., 1985) (Fig. 3).

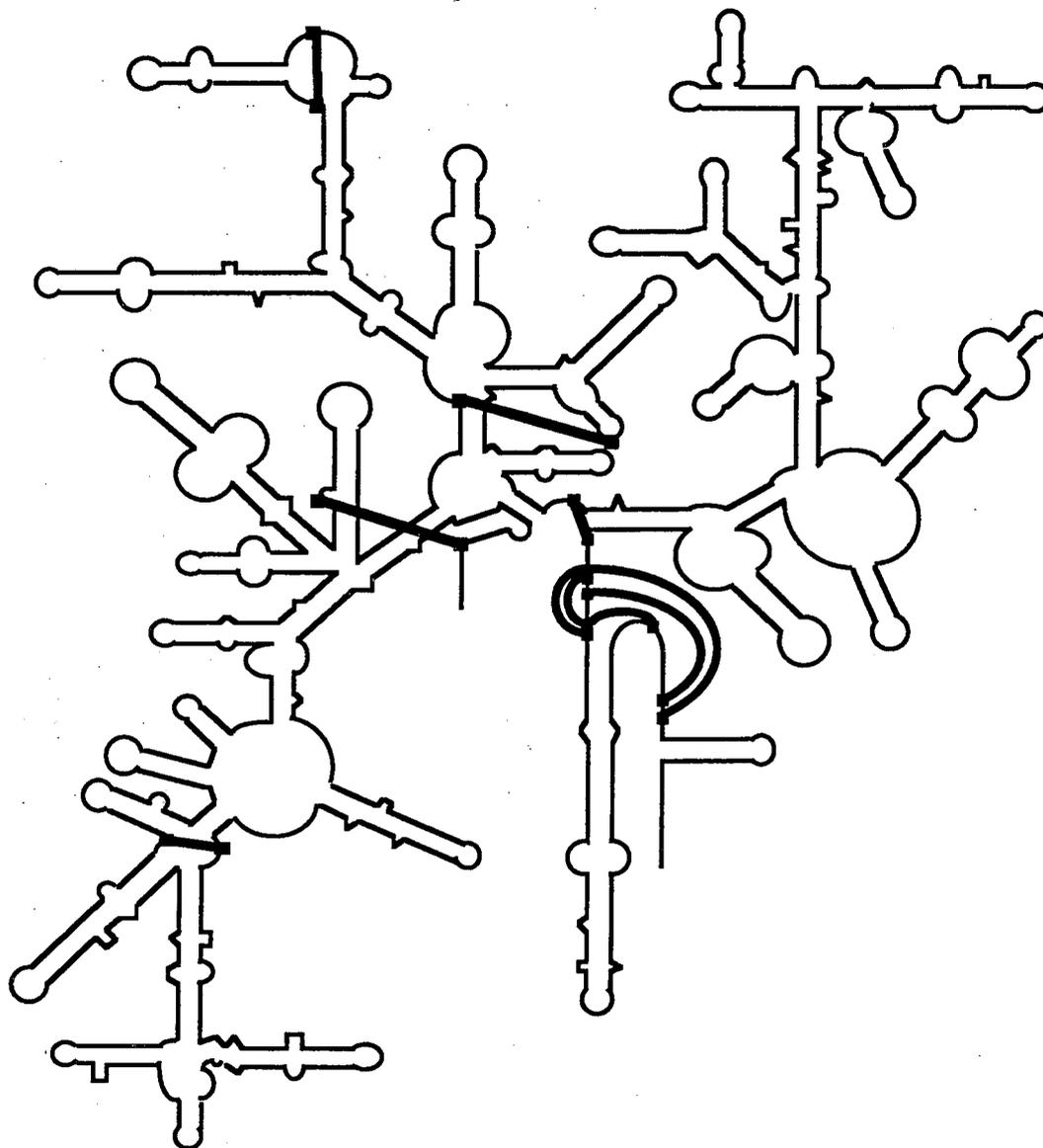


Figure 3. The tertiary phylogenetic relationships of 16S RNA.

As mentioned in the conversion of the secondary structure into pseudohelices, several marginal or unusual interactions were originally omitted. Helices of less than four basepairs require the same eight pseudoatoms as larger helices and only six pseudoatoms if they are single-stranded. After the developmental work on the regular secondary structure and the other classes of tertiary interactions, it became clear that there was sufficient room in the DSPACE arrays for a few additional pseudoatoms. Therefore the three basepair helices were added as a special class of 'weak' helices. Included in the weak helical definitions are the pseudoknot, formed by bases 17-19:916-918, and the helices formed by 184-186:191-193, 688-690:697-699, 1253-1255:1282-1284, and 1263-1265:1270-1272. Some additional weak interactions such as in the region 65-70:98-105 were constrained with pseudophosphate to pseudophosphate bounds but no hydrogen bonding groups were added to the structure.

Psoralen Crosslinks

The bifunctional intercalator, psoralen, in conjunction with UV radiation, is a well characterized crosslinker of helically stacked pyrimidines. Eight psoralen crosslinks which are indicative of tertiary structure have been isolated (Thompson & Hearst, 1983; Hui & Cantor, 1985). The upper and lower bounds for the crosslinks U14 X U921, U619 X U1420, U921 X U1532, U955 X U1506, U1116 X U1183, U1240 X U1298, and U1308 X U1330, were derived from the coordinates of a standard A-form helix. It was possible to identify within one or two residues which bases are crosslinked for these seven but that is not possible for the rare crosslink between the 5' and 3' domains (G354 X U1330). Other research confirms the existence of this interaction (Spitnik-Elson et al., 1985) so it is retained. Bases from the middle of the crosslinked fragments were chosen and a van der Waals contact lower bound was used. An upperbound of 25.5 angstroms was used to reflect the uncertainty as to exactly which bases are crosslinked (Fig. 4).

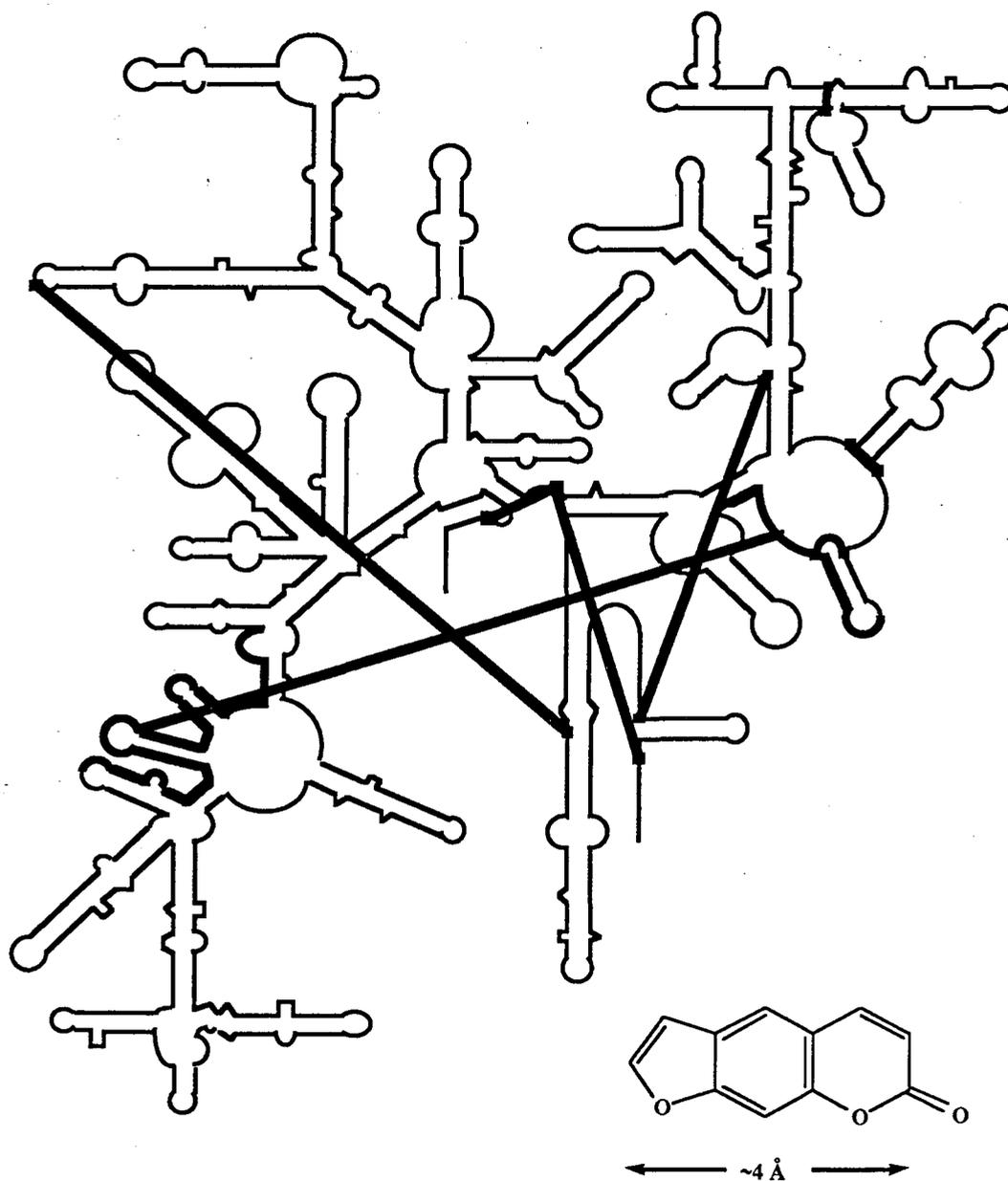


Figure 4. The psoralen crosslinks mapped on the secondary structure of 16S RNA.

Ultraviolet Crosslinks

Under ultraviolet irradiation cyclobutane bridges may be formed between bases which are in direct contact. The geometry of these crosslinks requires that the linked bases not be helically related. DSPACE constraints were used which would force van der Waals contact of the bases but not necessarily of the sugar-phosphate backbone of the crosslinked nucleotides. The crosslink G497 X A546 seems to indicate that the s1h47 and s1h50 helical strands are stacked but the 3' piece of the crosslink cannot be narrowed down to fewer than four nucleotides. As four nucleotides could form a very extended structure, a larger upper bound has been used for this crosslink. The other UV crosslinks, G31 X C48, U118 X A288, U367 X A397, U1091 X C1181, and U1348 X A1377 can be narrowed down to one or two nucleotides. Two of the UV links listed by Stiege (C582 X A759 and U1240 X U1298) are not included in the UV constraint files as they are superseded by more stringent psoralen (U1240 X U1298) and secondary constraints (helix s2h3:s2h16) (Stiege et al., 1986). After 16S RNA modeling was well advanced, an additional ultraviolet crosslink, A246 X A892, was determined (Stiege et al., 1988). This circumstance demonstrated the superiority of computer modeling as it was easy to add this crosslink to the input parameter set and produce new models (Fig. 5).

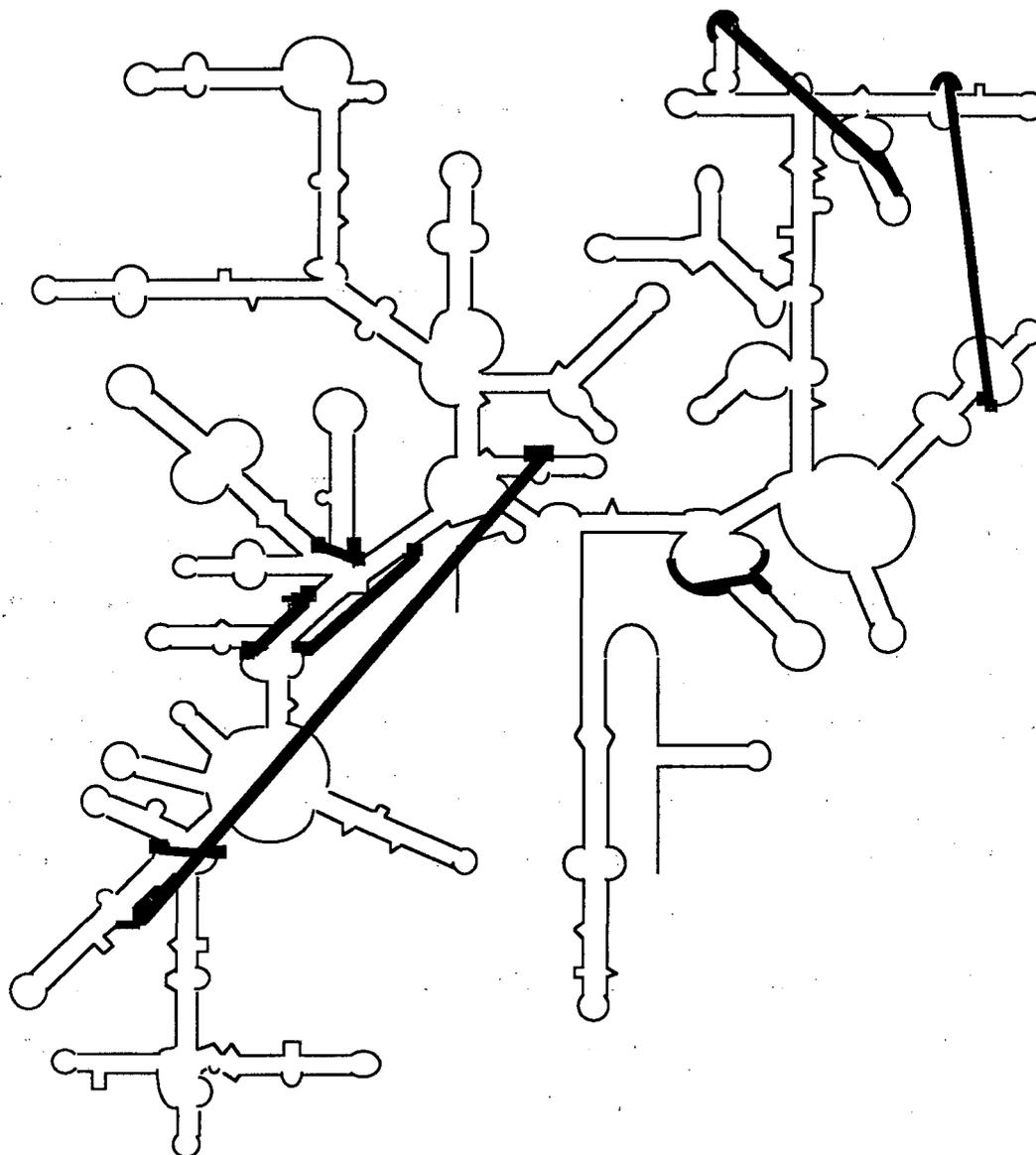


Figure 5. The ultraviolet crosslinks mapped onto the secondary structure of 16S RNA.

In addition to the phylogenetic, psoralen, and ultraviolet types of crosslinks used in folding tRNA, there are two new crosslinkers used to probe for 16S RNA tertiary structure.

GbzCynAc Crosslinks

N-acetyl-N'-(p-glyoxylylbenzoyl)cystamine forms a closed ring linkage to the hydrogen-bonding atoms of guanosine bases. Therefore bases that react with this reagent are single-stranded. The disulfide bond of this molecule can then be reduced, leaving an active sulfide group some seventeen angstroms from the covalently attached guanosine. Subsequent oxidations can produce a new disulfide linkage between adducted guanosines which are less than thirty-four angstroms apart. The minimum distance for a linker with this many degrees of freedom is essentially the van der Waals contact radius of the crosslinked bases. The maximum separation is calculated from an all-trans conformation of GbzCynAc. The seventeen GbzCynAc crosslinks that have been identified fall into two classes. Because GbzCynAc is so long and flexible, the ten crosslinks, G177 X G530, G265 X G450, G570 X G778, G926 X G1405, G966 X G1297, G1015 X G1405, G1084 X G1094, G1094 X G1276, G1138 X G1381, and G1316 X G1381, which occur within a domain of the secondary structure, merely confirm the correctness of the hydrogen bonding map. The remaining seven crosslinks, 265 X 703, 299 X 791, 703 X 926, 703 X 1381, 760 X 1127, 844 X 1258, and 869 X 1297, are between different domains of the basepairing map and will restrict the conformational space of 16S RNA by requiring that the crosslinked bases are on the same side of the 30S subunit and exposed to the solvent (Wollenzien et al., 1985) (Fig. 6).

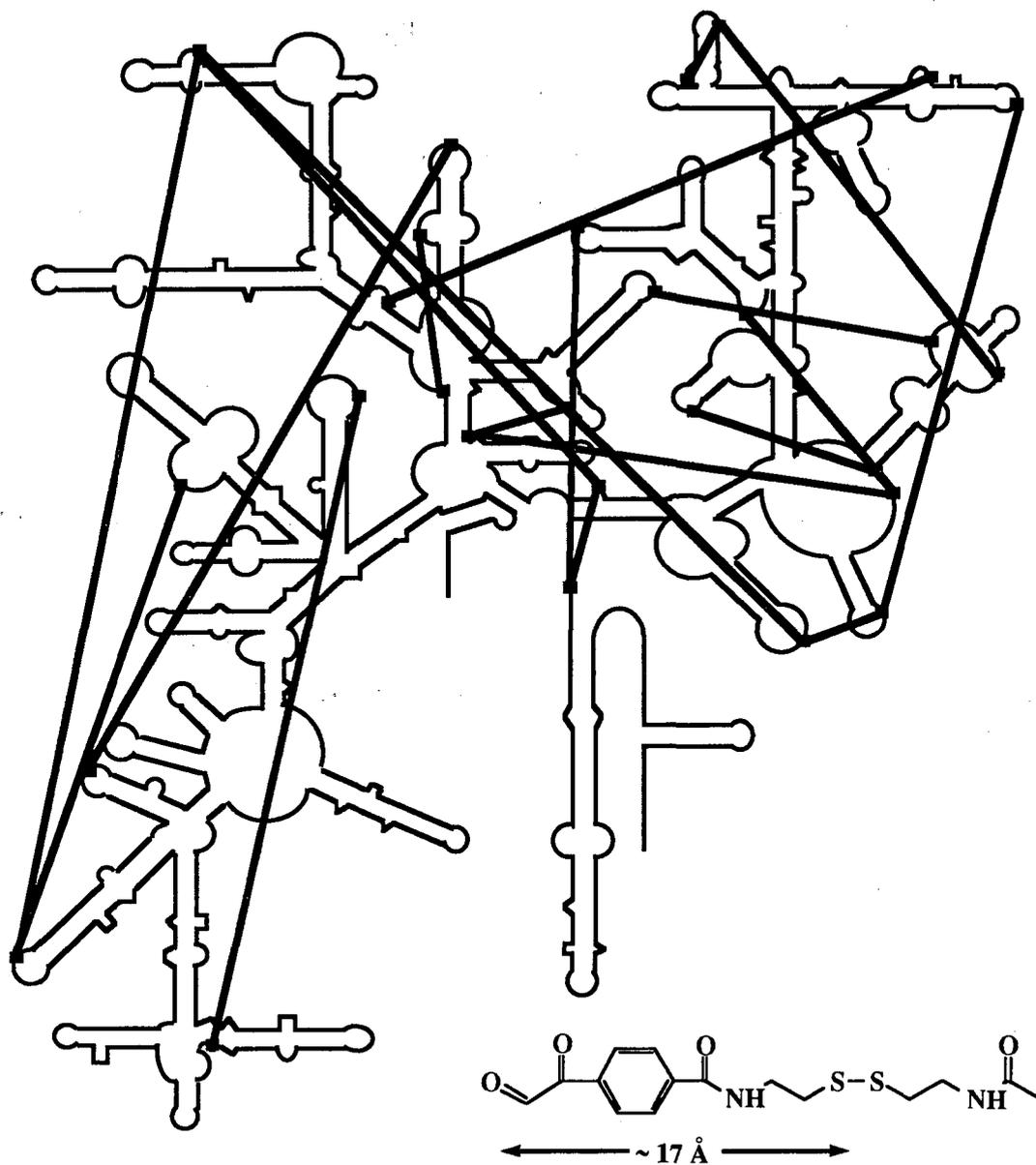


Figure 6. The GbzCynAc crosslinks are shown superimposed on the secondary structure map of 16S RNA.

Nitrogen Mustard Crosslinks

Nitrogen mustard is a much smaller distance probe as the maximum extension between the reactive chlorines is only four angstroms. The chlorine atom is a strong electrophile which preferentially reacts with the N7 of guanosines, although interactions with the N7 of adenosines is also known. Five mustard crosslinks which are indicative of tertiary interactions, G31 X A306, G46 X A306, G46 X A383, G693 X A794, and A695 X G799, have been identified (Atmadja et al., 1986). It is possible that the crosslinked bases are stacked in a helix, but it is not required, therefore loose constraints which allowed van der Waals contact or a coplanar alignment of the bases were used. As G47 is a member of the pseudohelix s1h5, it does not exist as a separately identifiable entity in the reduced 16S RNA representation. Therefore distance constraints with an expanded upper bound were specified to C48 which terminates the pseudohelix (Fig. 7).

In summary, the reduced structure of 802 pseudoresidues will be folded on the basis of 801 primary constraints, 132 helical segments, 10 weak helical segments, and 47 tertiary interactions.

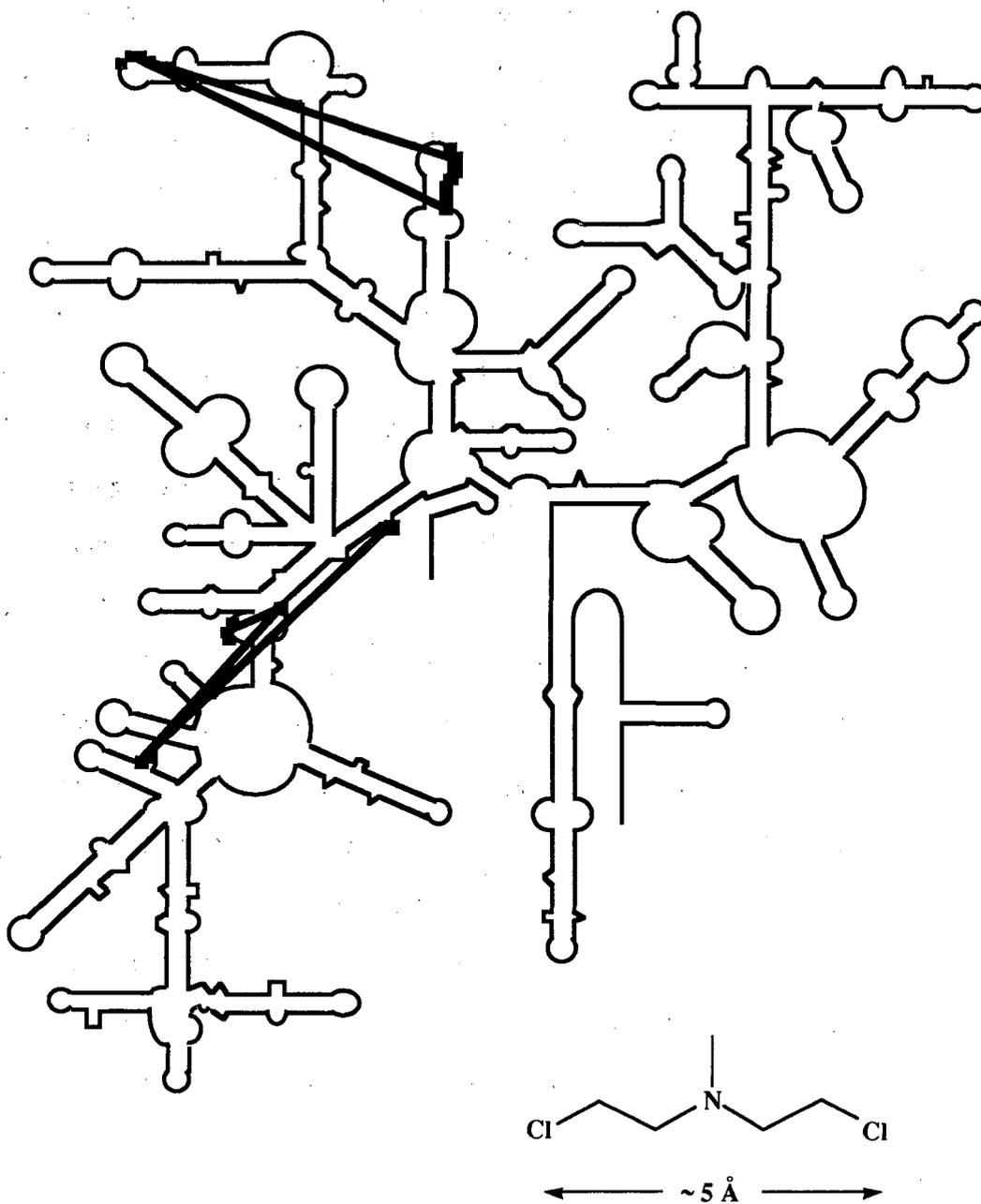


Figure 7. The nitrogen mustard crosslinks of 16S RNA mapped onto the secondary structure.

Methods

Physical Modeling

The physical model, byhand, was made by using a single piece of steel wire as the phosphate backbone and then threading this wire through solderless electrical connectors. The wider female connectors were used to represent the purines and the male connectors were used for the pyrimidine nucleosides. Basepairs were formed by mating the appropriate 'purine' and 'pyrimidine' connectors. Relying primarily on the psoralen crosslink data which he produced (Thompson & Hearst, 1983) and an early secondary structure map of 16S (Noller & Woese, 1981), John Thompson folded this crude representation of 16S RNA into a structure which was suspended in an open framework of approximately 2 feet square. The model was rather stiff and so no attempt was made to helicize the basepaired regions. Using a laser to light up each 'base' in turn, a set of three dimensional coordinates were determined and entered into the computer. An arbitrary conversion factor was used to convert the data set into angstroms such that the average phosphate to phosphate distance was 6.5 angstroms. While this compromise value will properly reflect the single-stranded and loop regions it will overexpand the helical regions. Due to errors in data collection and the crude nature of the model it was necessary to adjust by simple linear interpolation several phosphate coordinates which produced phosphate separations of more than 8.5 angstroms. Once these obvious errors had been corrected, it was possible to do real time manipulations of the digitized model on the black and white MPS display (Fig. 8). Attempts to produce a more spacefilling version of the model by

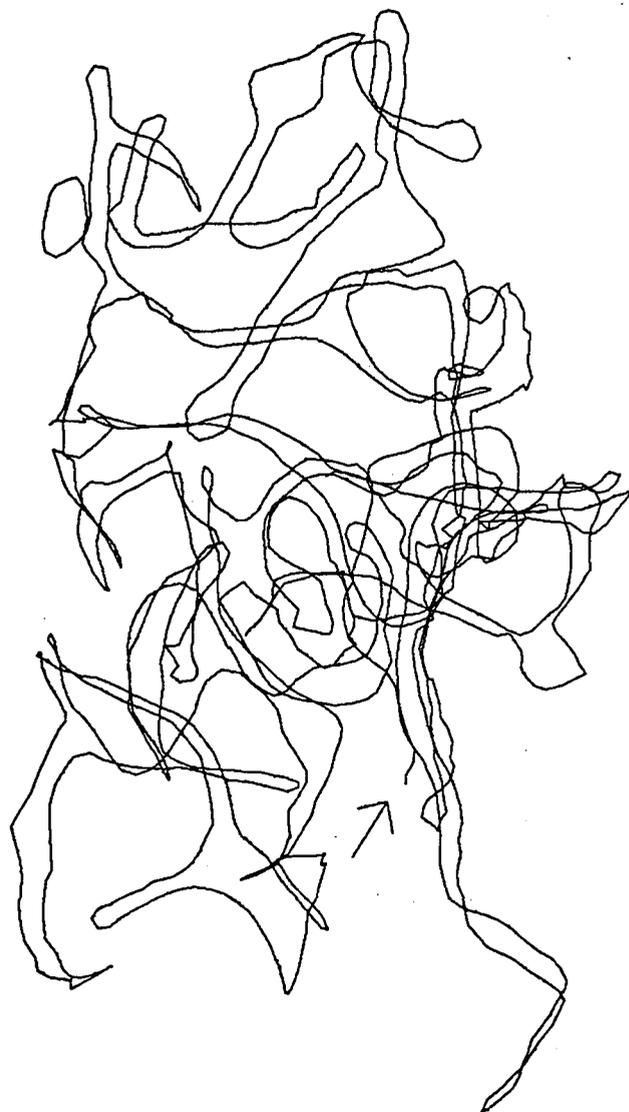


Figure 8. Line drawing of the model of 16S RNA that was constructed byhand.

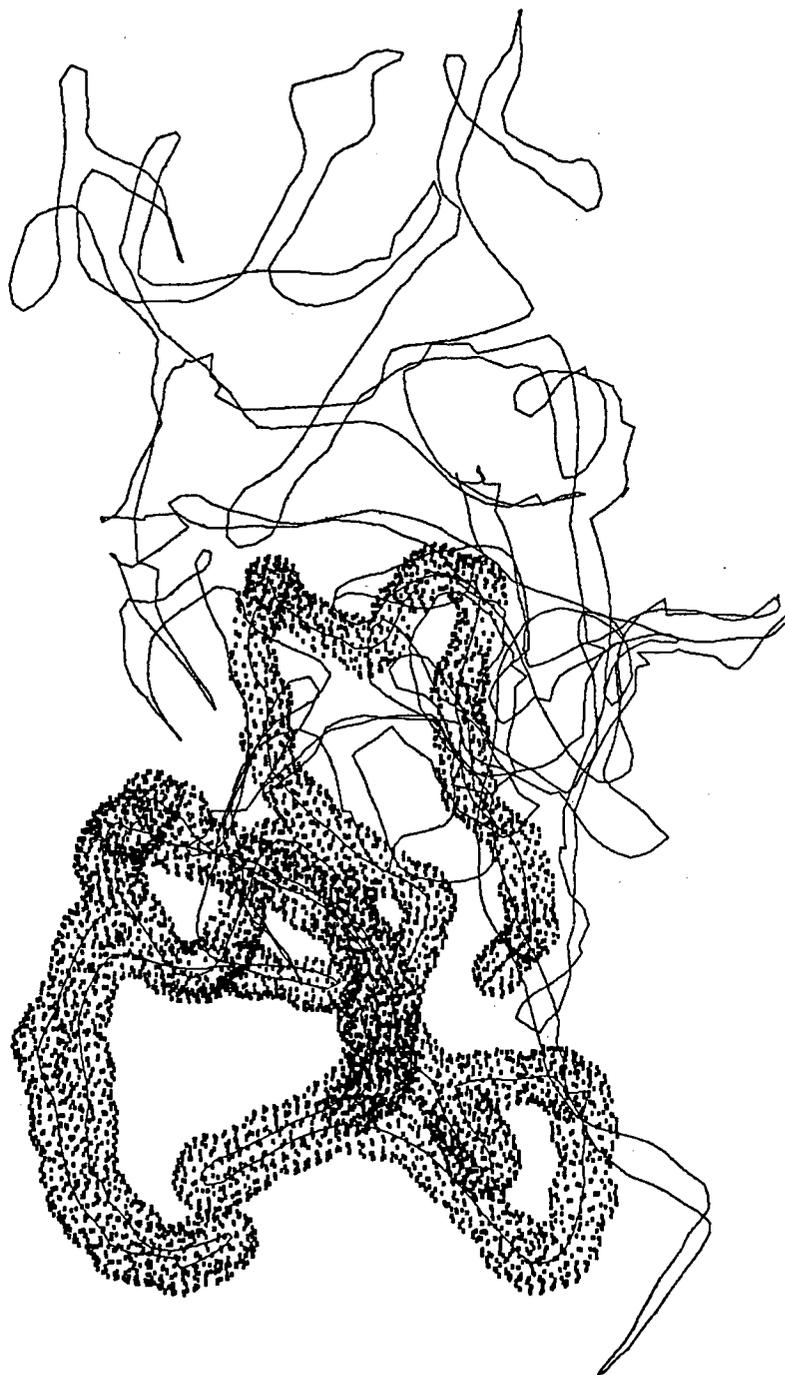


Figure 9. Line drawing of the byhand model with a dot surface at a radius of one angstroms covering the first 250 residues.

adding a halo of dots about the model backbone at a radius of one angstroms very quickly exhausted the memory capacity of the MPS. The contact surface for no more than 250 residues could be displayed at any one time (Fig. 9). No further work on the structure derived from the byhand model was pursued until the computer modeling protocol had been revamped and computer generated models had been completed. Then for purposes of comparison, AMBER was used to insert the full atom representation into the digitized version of the byhand model for each base as it occurs in the primary structure. This full structure was then reduced to the pseudoatom representation and the DSPACE minimizer was used to bring the arbitrarily uniform structure into agreement with the helical constructs. The byhand model was then aligned and compared to the models produced by the DSPACE folding protocol. Finally the byhand model was converted back into the PDB format for reinsertion of the full atom representation and production of the raster graphics displays.

Computer Modeling

The procedure for computer modeling a molecule involves a series of independent protocols which perform specific, limited data manipulations. The first task is to collect and encode the primary, secondary, and tertiary data. Distance geometry is used to objectively transform these structural clues into a three dimensional model. The pseudoatom constructs of the distance geometry model are then replaced with their real atomic representatives. Energy minimization of the all atom structure is used to resolve serious physical conflicts. Finally the structure is converted to a graphical display which facilitates analysis by the researcher.

Data Preparation

Creating the 132 pseudohelices required for 16S was a large task which could not easily be automated. A general residue definition pattern file was created which was adjusted with a text editor to reflect the particular sequence for each helical strand. Since only the pseudoatoms at the end of a helix are retained, some helical residues are identical

and required only a change in the residue name. Even with such shortcuts it took three full days to create all the necessary files. An unexpected benefit of this onerous task is that it forces a careful examination of every nuance of the secondary structure map.

The helical constraint distances are calculated from a reference set of double-stranded ARNA helices created with the regular version of NUCGEN. The initial structures were created with only eight constraints per helix. Regular helices of more than four basepairs were constrained by an additional eight interactions designed to enforce helicity in the later structures. The weak helices were not given these additional constraints since they may form the phylogenetically indicated hydrogen base pairs without folding into a regular A-form helix.

Distance Geometry

The adaptation of DSPACE to handle the 16S ribosomal RNA sequence required some tailoring of the program arrays so that the bounds matrix and the residue list could be expanded. Fortunately these arrays are collected in a single file which is added to the individual FORTRAN program source files during the compilation process. The final configuration of the program could fold and evaluate a molecule with a maximum of 1400 atoms. The program could display and superimpose at most two such molecules at a time. A significant increase in speed of operation was achieved by increasing the amount of real memory that the program could use to four MBytes. Under these conditions it took two full days of microVAX computer time to create and smooth the distance bounds matrix for the pseudohelical representation of 16S rRNA. Once the bounds matrix had been created, it was possible to produce a new 16S conformer in eight hours of computer time. Thus the more than 200 structures created during the course of this research represent approximately three months of computer time.

Atomic Reconstruction

The program, EXPAND, which calls up the appropriate ideal helix and superimposes it on the pseudohelix used by DSPACE, requires either 4, 8, or 12

pseudoatoms per helical residue. The reference helical files were named 'residue_name'.PDB, *.PDC, and *.PDD for succeeding helical segments. The present maximum of three segments per helical strand could be increased by altering the EXPAND program. The NUCGEN module of AMBER was modified to take sequence data interactively from the terminal and produce single-stranded ARNA helices by default. To facilitate the superposition process, the phosphate and hydrogen bonding atom from the first residue (i.e. the N1 atom of purines or the C4 atom of pyrimidines) and the phosphate from the last residue, are copied and inserted as the first three entries of the file.

Molecular Mechanics

The AMBER residue definition library was altered to include the phosphate ester in each ribonucleotide by default. This was both necessary and desirable. By including the phosphate, the total number of residues which AMBER must consider is halved. The AMBER definition file for the LINK module is greatly simplified by eliminating the redundant POM residues between each nucleotide. For simple reinsertation of an all atom representation, AMBER requires that each residue have at least one atom in the input PDB file. In the fully reduced pseudohelical form nearly one-half of the molecules exist as only pseudophosphates and by changing the AMBER residue definition this will be sufficient.

The changes required to allow AMBER to process the ~50,000 atoms of the full representation were more difficult. It was possible to expand the preparatory programs LINK, EDIT, PARM, and the analysis module ANAL, in a fairly straightforward manner by locating the common arrays which held the list of atoms, bonds, bond angles and dihedral angles. The PARM module required some additional adjustment since a temporary work array which did not require expansion had been equivalenced to one of the altered arrays. Adapting the MINimizer program from AMBER proved to be the major stumbling block. The MIN module actually does the structural manipulations necessary to minimize the free energy of the structure and it is the most demanding of the computer resources. Consequently this program dynamically adjusts its arrays and memory usage as needed.

The major new array which the other modules do not have is the list of nonbonding interactions within a specified radius for each atom. Even when all other arrays were precisely tailored to the size of 16S rRNA it was not possible to run the MIN module with the 12.0 angstrom nonbonding radius which is listed in the example data sets distributed with AMBER. 2.5 angstroms proved to be the practical limit for MIN on the microVAX and at this setting a single update of the nonbonded list for 16S rRNA took an entire day of computer time. With such a small radius, favorable nonbonded interactions are not properly anticipated in the analysis of the conformational free energy surface and extremely unfavorable van der Waals conflicts may be allowed to develop as well. These problems effectively restrict the usage of AMBER to the reinsertion of the all atom representation, a simple Newton-Rapheson minimization of any gross structural conflicts, and an analysis of the resulting structure. Barring a massive increase in the computing power dedicated to this process, it will be necessary to reduce the number of atoms to be minimized by doing sections of the structure or by introducing pseudoatom constructs similar to the 5mer models used in the early stages of tRNA modeling.

Raster Displays

After AMBER the resultant PDB formatted files are converted to the ATM format required by the Scripps Institute modeling package. A virtual bond file which sequentially links the phosphates of a structure was created with the MKBND utility. These two files are used as input to the PGI program which produces a GEO file that describes the molecule as a series of three dimensional polygons. Given the particular orientation, framing, and coloration desired, the MCS program produces smooth tube renderings which can be viewed on the raster display of the PS340. For the backbone representation, only the phosphate atoms were used and the diameter of the tubes was chosen as two angstroms as this approximates the van der Waals radius of phosphate.

The GEO file contains the solid polygons in an unusual order and it took a few trials to figure out that the program creates the long single backbone tube by first creating a

5' hemispherical cap and then a small 5' cylinder for each phosphate. These cylinders do not extend to the next cylinder but leave a gap. After all the phosphate cylinders have been created the capping 3' hemisphere is added. The program then returns to the 5' end of the sequence and generates the cylinders necessary to bridge the gaps left in the first pass. Once this pattern was discovered, it became possible to introduce more varied effects with region specific coloring.

The first color displays were simple red, green, and blue domain partitions. More precise color patterning was introduced by shrinking the original coloration to helices well separated in the primary structure. The first helical region is red, the final helices are blue, and the well characterized helix formed by bases 588 to 651 is green. Color was then added to helices which seem to form independent subdomains and are known to be affected by the binding of particular ribosomal proteins. The color of a backbone section is specified by an MCS command for each particular section. As the 1542 residues of 16S rRNA produces 3083 polygons, the program, MCSCOL, was written to automate the production of the necessary color specification files. The colors yellow, purple, and cyan, were used because they are easy to define as equal mixtures of two primary colors and they offer striking contrast. Thus they should be reproducible and will facilitate the interpretation of the folded models.

The addition of orange to the base set of six colors still left most of the 16S backbone undifferentiated. The idiosyncrasy of the manner in which the polygons are produced became a positive feature when either the 5' or 3' set of polygons for helix are colored white, producing a striped pattern and effective doubling the number of colors. An all black helix would not be visible but a black and white striped helix is. A final colored helix was added when a trial showed that rose could be distinguished from red on both the screen and photographs of the screen.

The large cylinders meant to represent helices were created by picking the hydrogen bonding atom from the 5' and 3' purines at the end of a helix and using them as

the helical axis. A more sophisticated method of determining the true helical axis is not justified in light of the simplicity of the modeling constructs. The diameter of the helical cylinders is 20 angstroms, the diameter of an A-form helix.

The default treatment of these constructs produced capped helices which resembled a medicine capsule. Since the radius of the capping hemispheres was the same as that of the cylinder, they extended several nucleotides beyond the nominally helical region. It was necessary to edit the GEO files and set the radius of the hemispheres to 0.1 angstroms to avoid this problem. This had the added benefit of making it possible to look inside the large helices and see the backbone tube. As isolated polygons, these large cylinders consist of only one piece and to distinguish solid and striped helices, the outside of the large cylinder for a striped region is white and the interior wall is colored.

The raster displays were photographed directly from the screen of the PS340 with a 35mm camera. Using the gray background, good results can be obtained using 100 speed film, a one second exposure time, a wide open aperture, and the PS340 display set to maximum brightness.

Results

Partial Folds

The initial attempts to model 16S ribosomal RNA were done without any long range tertiary constraints. The primary purpose of these early runs was to check the primary sequence and helical substructures for errors. The elongation of the sequence as more constructs became available was also used to establish a practical lower limit on the resources needed for DSPACE to run. When the entire sequence was included, the initial 16S structure forms an extended, flat 'Y' of approximately 500 X 475 X 25 angstroms. This conformation reflects the three major secondary structure domains (Fig 10). The 5' domain forms the bottom stem, the middle domain is on the left, the 3' domain is on the right, and the regions beyond 1400 form a separate subdomain which extends some 120 angstroms perpendicularly into the page. While reflecting the three domain nature of 16S, the conformation produced by DSPACE without long range constraints is still over-extended with the longest dimension being approximately 450 angstroms, the radius of gyration being 152 angstroms, and the volume is 14.68 million cubic angstroms. In electron micrographs the small subunit of the ribosome is approximately 220 X 220 X 55 angstroms (Hill et al., 1969), with a longest dimension is on the order of 250 angstroms (Lake, 1985). The very large dimensions of the initial computer generated structure demonstrate that the DSPACE program has not introduced any unwarranted compaction. This structure may correspond to the 'Y' structures reported in electron micrographs of 16S RNA prepared from very low ionic concentration solutions (Vasiliev et al., 1978). In such solutions the electrostatic repulsions of the negatively charged phosphate backbone can be expected to dominate the three dimensional structure and should produce extended conformations.

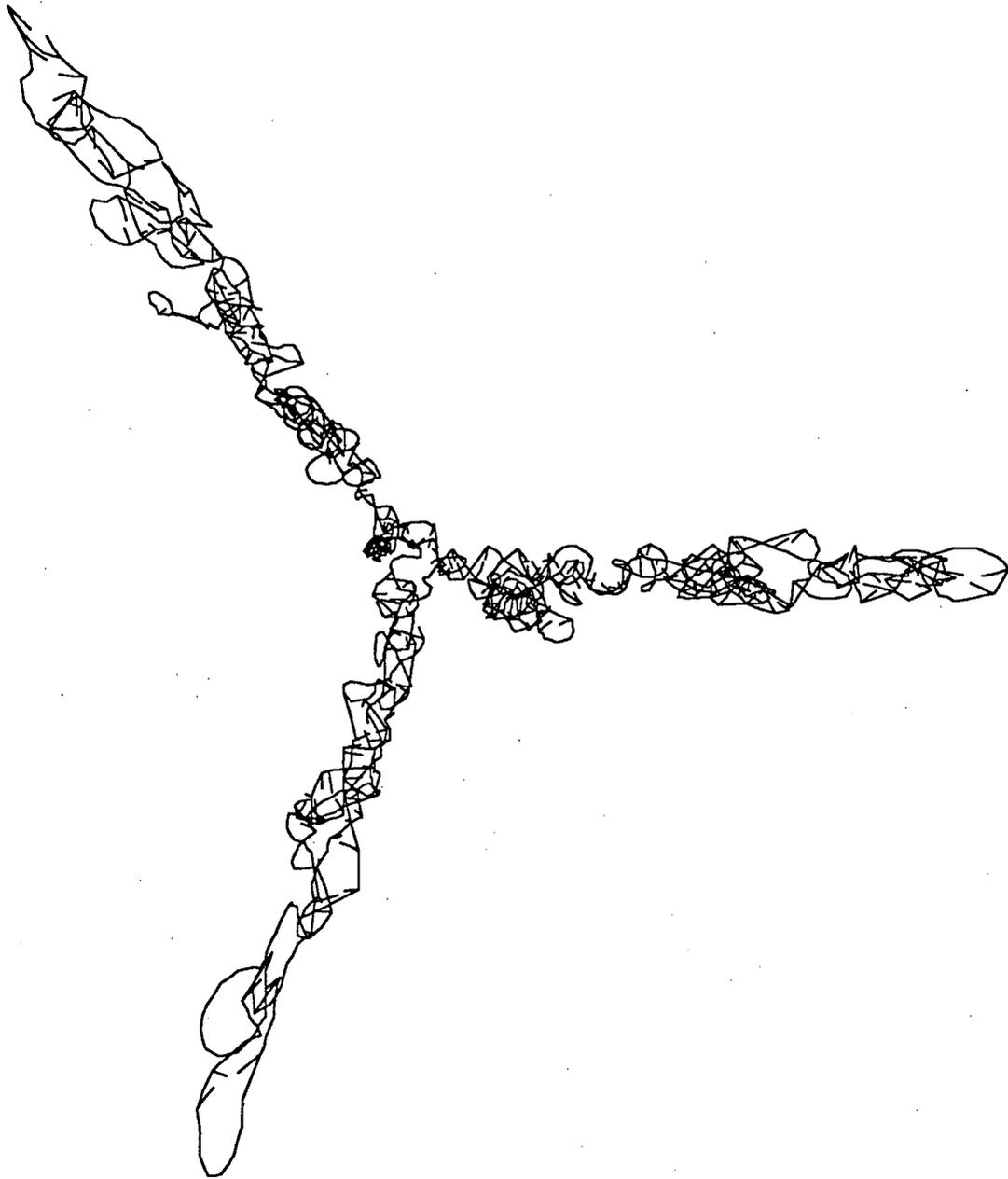


Figure 10. 16S folded on the basis of secondary structure relationships.

The intermediate models of 16S rRNA were arrived at by creating a set of constraint files, one for each data type. Then a new distance matrix and 16S conformer were produced as each data set was added to previous group of secondary and tertiary constraints. In this fashion mistakes in structure definition, simple data entry errors, or errors in bounds specification could be detected and easily isolated. The series of structures demonstrated that even a partial set of the crosslinks are sufficient to produce a significant compaction of the molecule. The tertiary phylogenetic relationships were the first set of constraints to be employed. The extended arms of the 'Y' produced by the secondary structure map are folded back toward the center of the molecule without distorting the general shape. The dimensions of the structure are reduced to 375 X 300 X 25 angstroms with the penultimate helices (residues 1409-1489) still extended along the z-axis (Fig. 11). The addition of the psoralen and UV data sets produced conformers in which the 5' and middle domains are folded back on themselves and well out of the plane which the structures based on phylogenetic data occupied. The overall dimensions have been reduced to 230 X 200 X 200 angstroms but the longest diagonal is still 300 angstroms (Fig. 12). Adding the GbzCynAc crosslink data produces a significant folding of the 3' domain back towards the center of the molecule. The overall dimensions of 230 X 225 X 90 angstroms with a maximum extent of 255 angstroms is very similar to the estimates of the size of the small subunit (Fig. 13). When the nitrogen mustard crosslink data is finally added the structure shrinks to 250 X 225 X 70 angstroms (Fig. 14) and it is not difficult to discern the resemblance to the outline of the 30S subunit (Lake, 1985). The inevitable errors were corrected, several data set permutations were tried, and the number of steps of refinement were varied from 60 to 160, but the size of the model and its resemblance of a broccoli stalk proved immune to any minor changes in parameters. This set of models are very similar, with the 5' domain forming a dense stalk, the middle domain forming parallel helices at the top end of the molecule, and the 3' domain being grouped about the middle of the structure except for the penultimate helices which extend radially from the surface.



Figure 11. 16S RNA folded with only the tertiary phylogenetic relationships.



Figure 12. Tertiary phylogenetic + psoralen + UV relationships

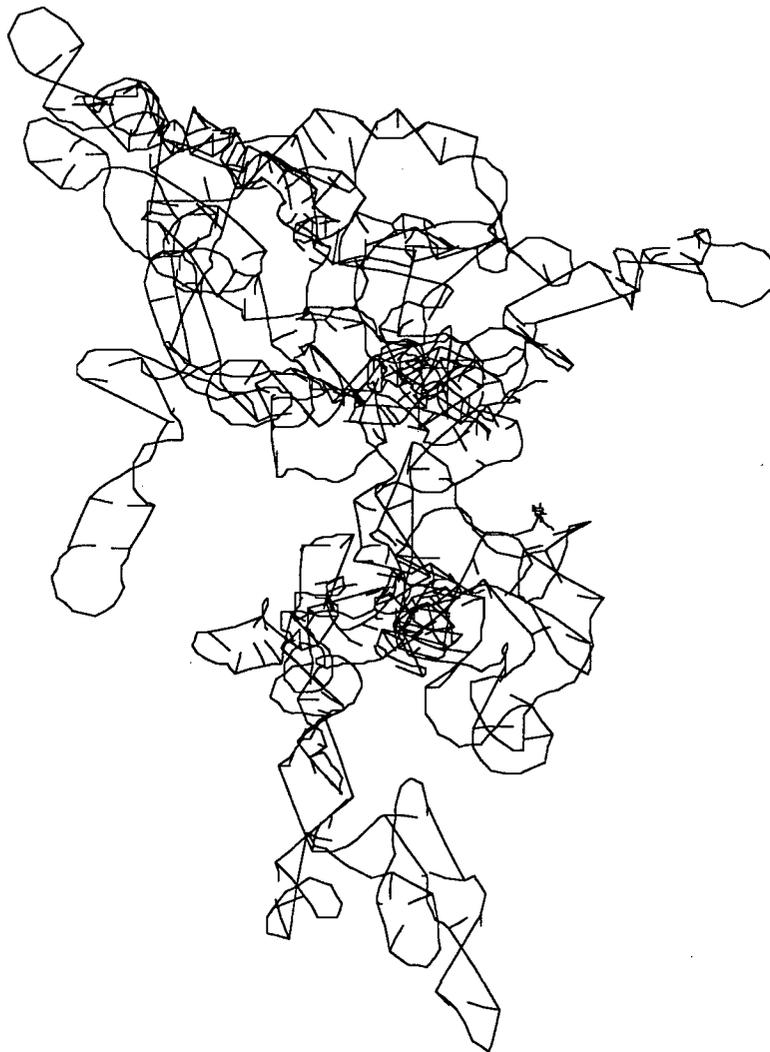


Figure 13. 16S RNA folded with GbzCynAc crosslinks added to the constraints set.



Figure 14. 16S RNA conformation produced with nitrogen mustard crosslinks added to the data set.

During this shakedown phase of the modeling process an additional UV crosslink was published (Stiege et al., 1988). Models produced without this new constraint, had separations between the crosslinked bases that were as much as 30 angstroms too large. But the models did place these bases in the same region. This compares with the separation of 40 angstroms found in the byhand model of 16S RNA. Unlike the byhand model, which would require difficult adjustment or even an extensive reconstruction to include this new data, it took only minutes to add this constraint to the UV data set. When the new crosslink is included in the input parameter set, distance geometry easily produces a model which satisfies the new constraint and still displays the structural characteristics of the earlier conformers. At 200 X 150 X 130 angstroms this structure is highly suggestive of the shape of the 30S subunit and approaches the appropriate dimensions (Fig. 15).

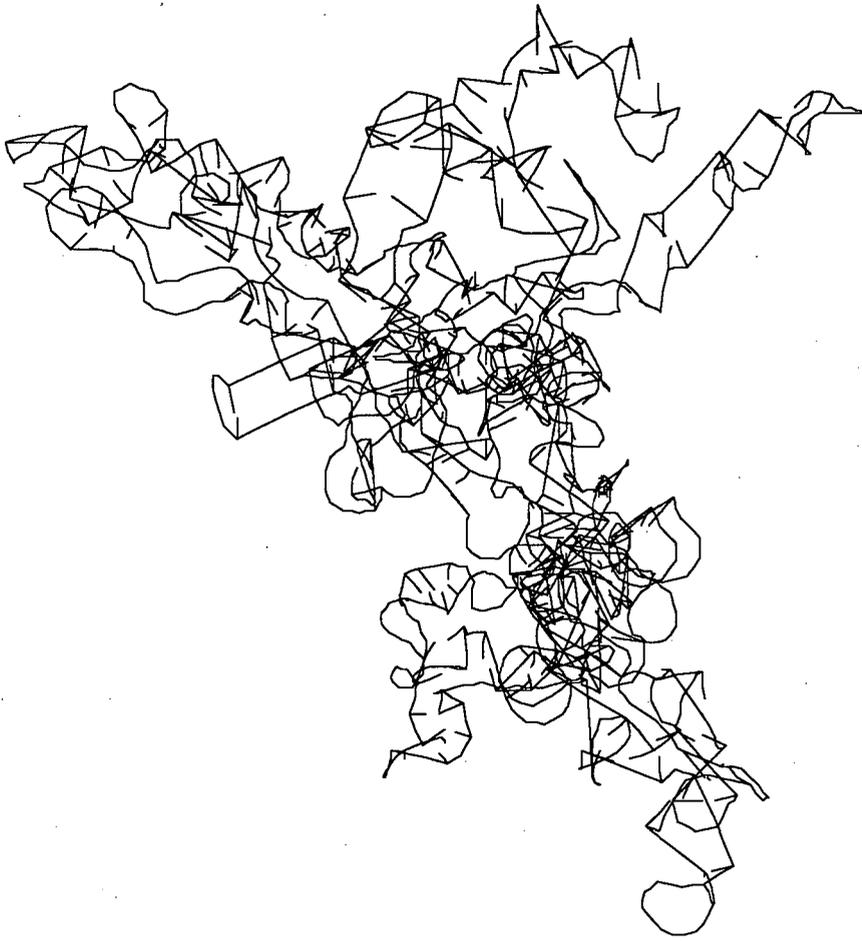


Figure 15. 16S RNA structure including the new UV crosslink.

Folds with 8 helical constraints (nt)

When the all long range crosslinks and relationships are included, DSPACE produces a compacted molecule which resembles the 30S particle, has a longest dimension of ~250 angstroms like the 30S particle, and has a radius of gyration (76 angstroms) which is intermediate between that of partially denatured 16S RNA (85 angstroms) and fully folded 16S RNA (66 angstroms) (Serdyuk et al., 1983). It is very gratifying that these physical correspondences were a natural result of the modeling process and not initial modeling assumptions. From the primary, secondary, and tertiary data, DSPACE is able to exactly determine 4329 of the $(1369)^2$ interatomic distances with an average difference between the upper and lower bound for all distances of 187 angstroms. 59 of the 100 structures created with eight constraints per helix were unique. The structures produced by distance geometry were subjected to an arbitrary 128 steps of conjugate gradient refinement and the gradient for all structures ranged from 790 to 1414 square angstroms while the total violation of all boundary conditions varied from 1599 to 2652 angstroms. The thirty-second structure produced had the lowest cgr error function and total bounds violation. This conformer has dimensions of 225 X 220 X 90 angstroms and when analyzed with AMBER its radius of gyration is 76.2 angstroms and its volume is 1.85 million cubic angstroms (Fig. 16). This is comparable to the volume of 1.77 million cubic angstroms suggested by preliminary synchrotron X-ray crystal studies of the 30S subunit (Yonath et al., 1989).



Figure 16. The nt32 model of 16S RNA made with 8 constraints per helical subunit and all the tertiary structure information.

When sorted on the basis of superposition on the structure with the lowest error function, it becomes apparent that the structures produced by distance geometry are split 50/50 into two classes. Of the 59 independent structures 29 are closely related to nt32 with the sum of the length of the vectors separating corresponding pseudoatoms after superposition ranging from 10.5 to 12.0 angstroms. The RMS deviations of this family of structures from nt32 vary from 450 to 510 angstroms. The other class shows a similar relatedness to nt13, the member of this group with the lowest error functions. When nt13 is superimposed on nt32 the fit error is 55.7 angstroms and the RMS deviation is 2473 angstroms. Squaring the fit error, multiplying the result by the number of pseudoatoms (1369), and then taking the square root yields a value of 2061 angstroms. This indicates that the large RMS deviation is the result of large superposition differences throughout the structure and is not the result of a relatively few enormous superposition problems. Although it was possible to orient the molecules based on a few prominent features (5' end, 3' end, penultimate helix), any attempt at a more detailed analysis was simply not possible with black and white displays. Using the limited color capabilities of the Tektronix 4105 vector display made it possible to pick out a few particular domains but the lack of depth was especially limiting in attempting to determine if one of the structural classes was superior to the other. It was clear though that the double stranded segments were not being properly wound into helical conformations. This is especially obvious in the helix which extends down and to the left in the nt32 structure. Therefore an additional eight constraints per helix were added to the secondary structure parameter files. These constraints specify the upper and lower bounds for the pseudoresidue at the opposite end of the opposite strand of a pseudohelix.

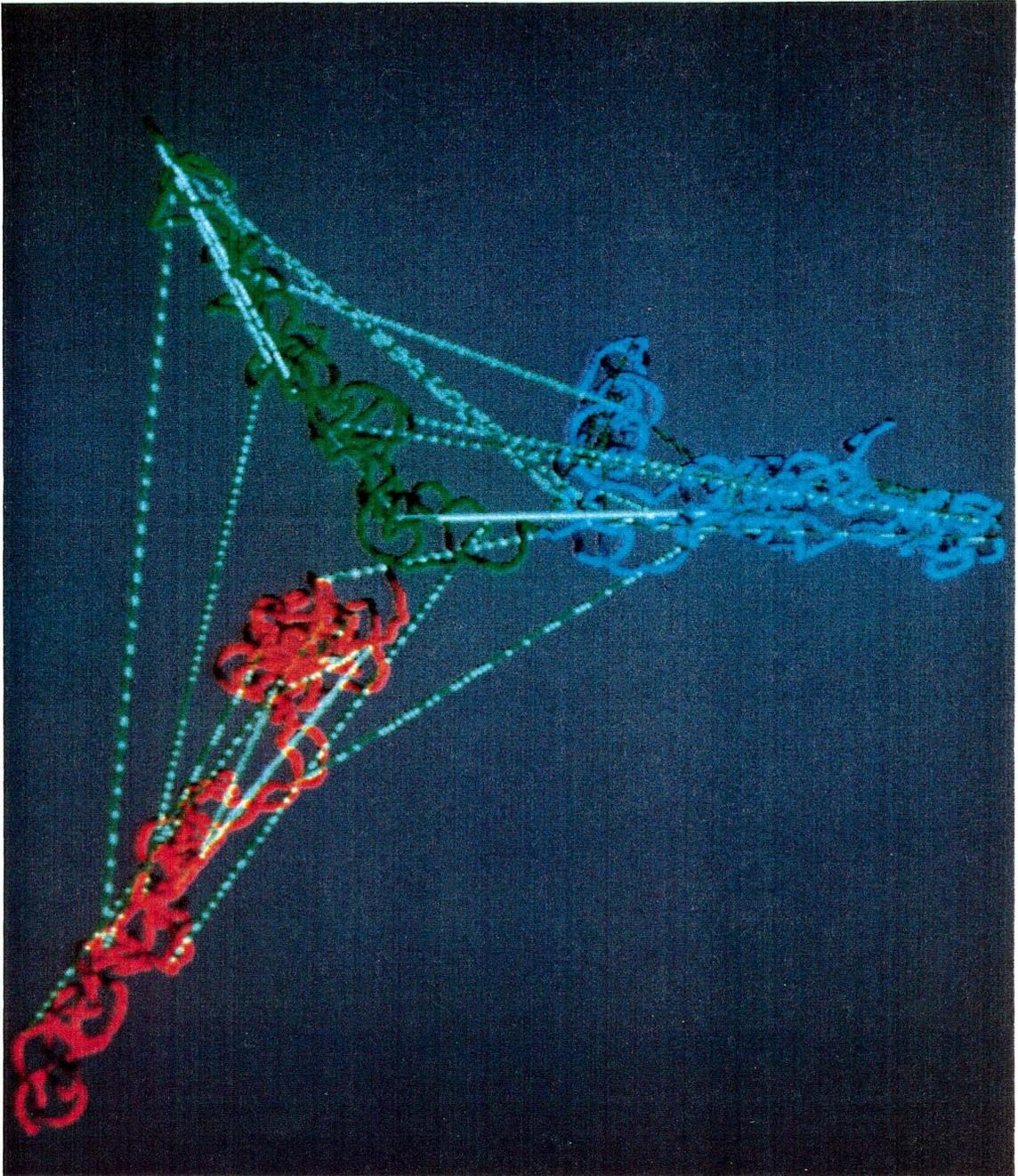
Folds with 16 constraints per helix (last)

A total of 100 structures were created with 16 constraints per helix and all the long range interactions. 58 of these structures are unique. The cgr error functions for these structures range from 807 to 1315 square angstroms while the total bounds violations vary

from 1705 to 2180 angstroms. The structure last70 has the lowest values for both these criteria. When superimposed on this conformer, the structures generated by DSPACE are again divided equally into two classes. 31 of the independent structures resemble last70 while the remaining 27 are closely related to the next best structure, last74. The cgr error function for the structures ranged from 807 to 1315 square angstroms while the overall bounds violations varied from 1705 to 2180 angstroms. The family of structures which resembled last70 could be superimposed on it with a fit error ranging from 10.7 to 12.4 angstroms and an RMS deviation of 459 to 521 angstroms. The worst superposition fit on last70 was scored by last74 with a fit error of 56.6 angstroms and an RMS deviation of 2494 angstroms. The family of structures which resembled last74 showed a similar range of superposition errors when compared to last74. When the nt set of structures, was compared to the present one, the family which resembled nt32 showed a strong correspondence to the last70 family although none of those structures scored as well as any member of the last70 set. nt32 proved to have a superposition fit error of 13.72 angstroms and an RMS deviation of 590 angstroms when compared to last70. When converted to an all atom representation and analyzed by AMBER, last70 has a volume of 1.85 million cubic angstroms and a radius of gyration of 76.16 angstroms. Its overall dimensions are 245 X 190 X 140 angstroms. Last74 has similar characteristics with a volume of 1.87 million cubic angstroms, a radius of gyration of 76.45 angstroms and overall dimensions of 240 X 200 X 180 angstroms. These compare favorably with the dimensions of the small subunit mentioned previously and the ten best structures all display similar characteristics.

On the following page:

Figure 17. The three color raster picture of 16S RNA folded without the long range interactions. The long range relationships to be added are shown as light green lines.



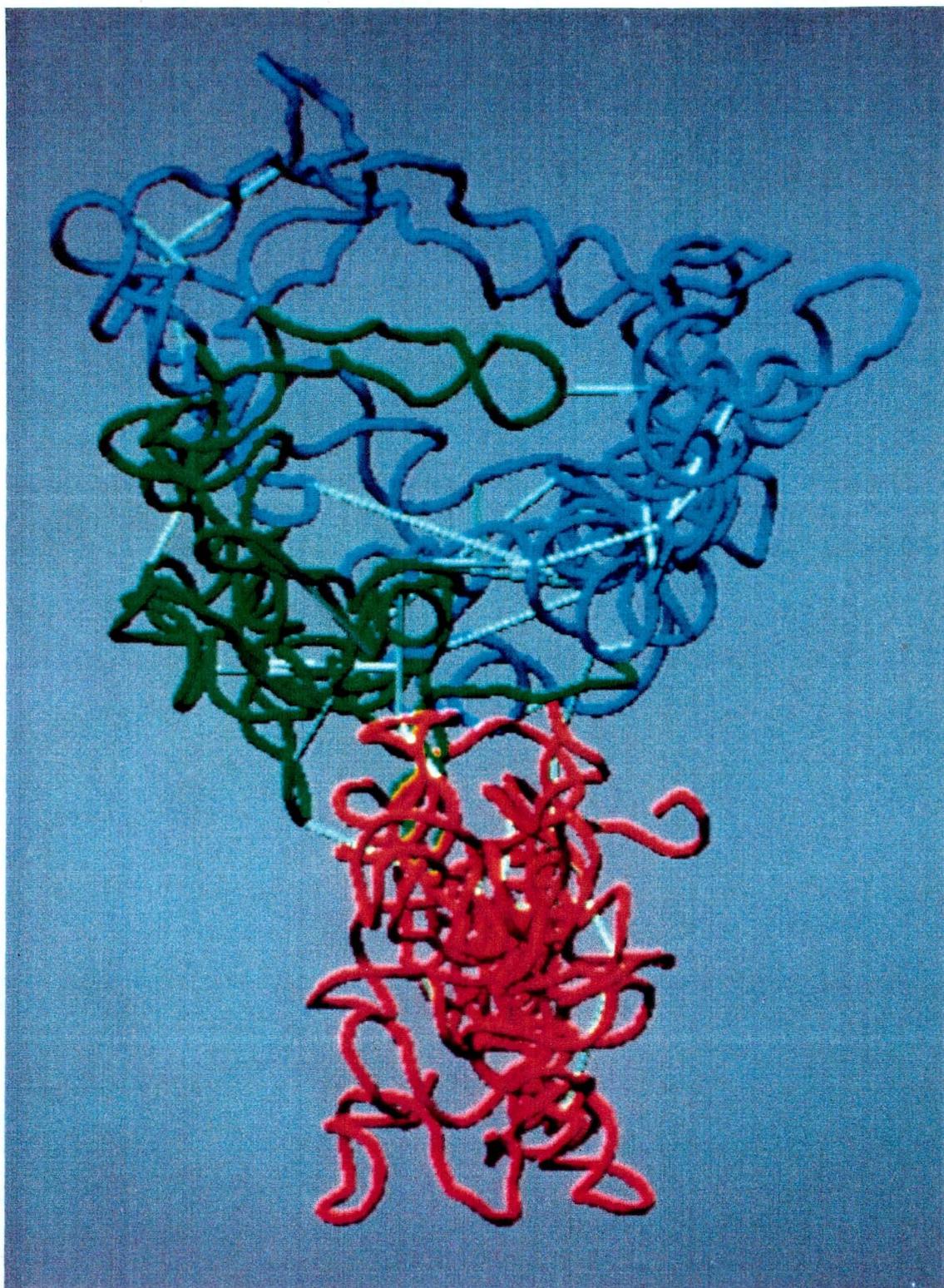
BBC 901 – 810

Figure 17

A tricolor raster display of the flat secondary structure conformer clearly shows three independent domains with the 5' domain in red, the middle domain in green, and the 3' domain in blue (Fig. 17). The light green vectors represent the tertiary relationships used to fold the structure and some of the constraints that link the different domains are very extended. A similar display of last74 shows that the three domains have been folded into distinct regions of the structure. The 5' domain forms the base, while the middle and 3' domains face each other in the upper part of the folded molecule (Fig. 18). The tertiary constraint vectors have been easily reduced to a length which lies between the upper and lower bounds for a particular crosslinker. In satisfying the constraints imposed on the folded structure, DSPACE most often has trouble with the irregular helices, especially the helical run formed by S3H14-17 and S3H30-32. This is a reassuring result, as it is difficult to predict what sort of structure this region, with its multiple bulged bases and loops, will form. Further analysis of the 16S RNA structures produced by distance geometry is very difficult with the techniques employed so far. As the models show good agreement with the physical measurements and satisfy the input constraints, it is time to create a more detailed representation of the structure.

On the following page:

Figure 18. The three color display of the computer generated structure of 16S RNA. The light green tubes show the long range interactions that were used to fold the molecule.



BBC 901 - 812

Figure 18

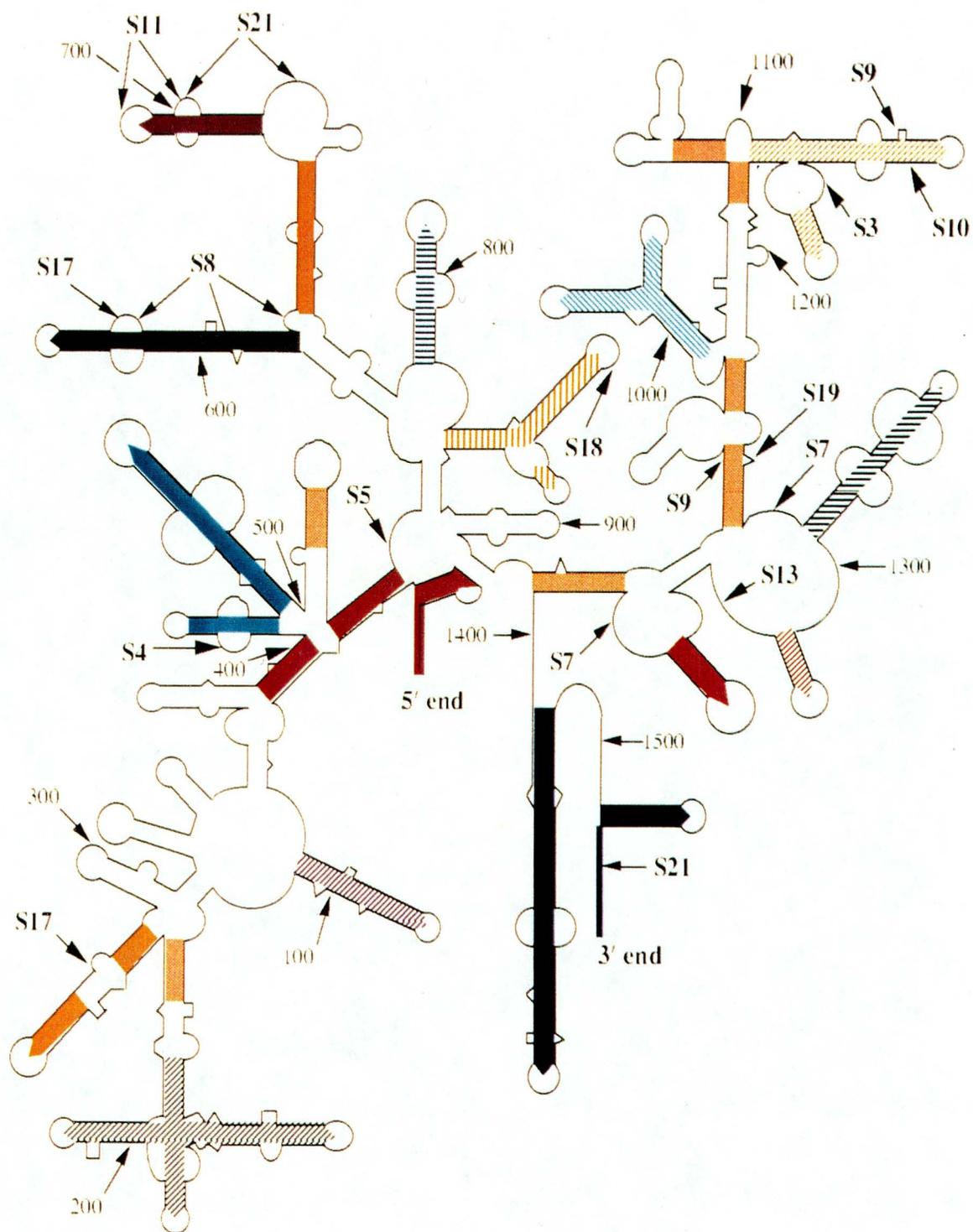
The colored schematic of the secondary structure map of 16S RNA demonstrates the system that was developed to differentiate the regions of the molecule (Fig. 19). The shaded regions indicate the color of the backbone and cylindrical constructs that will be used. Regions that were not given a unique shade were colored beige.

HELIX COLOR MAPPING OF 16S RNA

<u>COLOR</u>	<u>REGION</u>	<u>BASES</u>
Red	5' end, S5, S12	1-47, 394-403, 547-556
Purple-striped	variable region V1	65-104
Green-striped	S20, variable V2	136-227
Yellow	S17, S20	247-277
Cyan	S4, variable V3	406-497
Green	S8, S17, S16	588-651
Orange	S15	655-672, 734-751
Purple	S11, S21, S6, S18	677-713
Blue-striped	S6	769-810
Yellow-striped	S18, S6, variable V5	821-879
Beige	S9, S19	946-955, 1225-1235
Cyan-striped	S3, S2, variable V7	997-1044
Orange-striped	S3, S9, S10, S2	1113-1187
Black-striped	S7, S9, variable V8	1241-1296
Red-striped	S13, S7, S19	1308-1329
Rose	S7, S9	1350-1372
Blue	S21, variable V9, 3' end	1409-1542

On the following page:

Figure 19. The secondary structure map of 16S RNA color patterns.



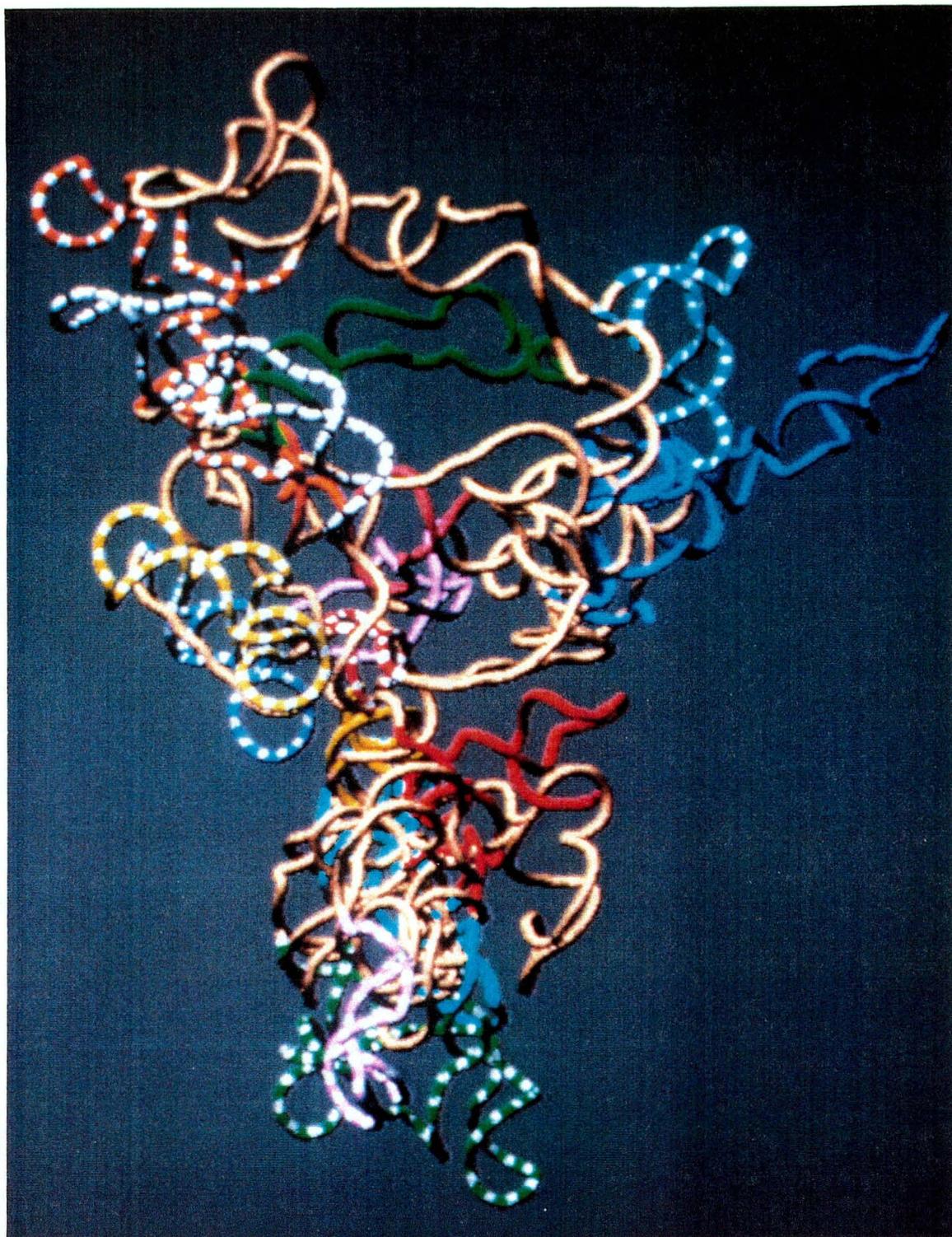
BBC 901 - 728

Figure 19

Although the switch to color raster displays entailed a new set of problems and limitations as mentioned in the Methods section, even a cursory glance at the color figures revealed that it was an eminently practical means for more detailed study of the models. With the full color version of last70 it is easier to follow the path of the backbone and it is simple to discern which region is closer to the viewer (Fig. 20). As the structure with the lowest error functions, a particular orientation of last70 was used as the guide for generating the displays for all other conformers. The view chosen was tentatively identified as the solvent-facing front of the molecule (as opposed to the side facing the 50S subunit) because of the general outline, the position of the red 5' end on the right, and the yellow-striped helix on the left. In this orientation the purple-striped helix is in front and with the green- striped helices forms the base of the molecule. The other 5' domain helices (red, yellow, and cyan) complete the bottom stem. The orange, purple, blue-striped, and yellow-striped helices of the middle domain are clustered on the left waist of the structure with the green helix protruding from the back. The 3' helices form the head of the molecule with the orange-striped and black-striped helices on the left and the blue-striped and blue helices on the right. The rose helix is in a central bridging position between these two groups.

On the following page:

Figure 20. The colored raster display of the solvent face of the last70 model of 16S RNA.



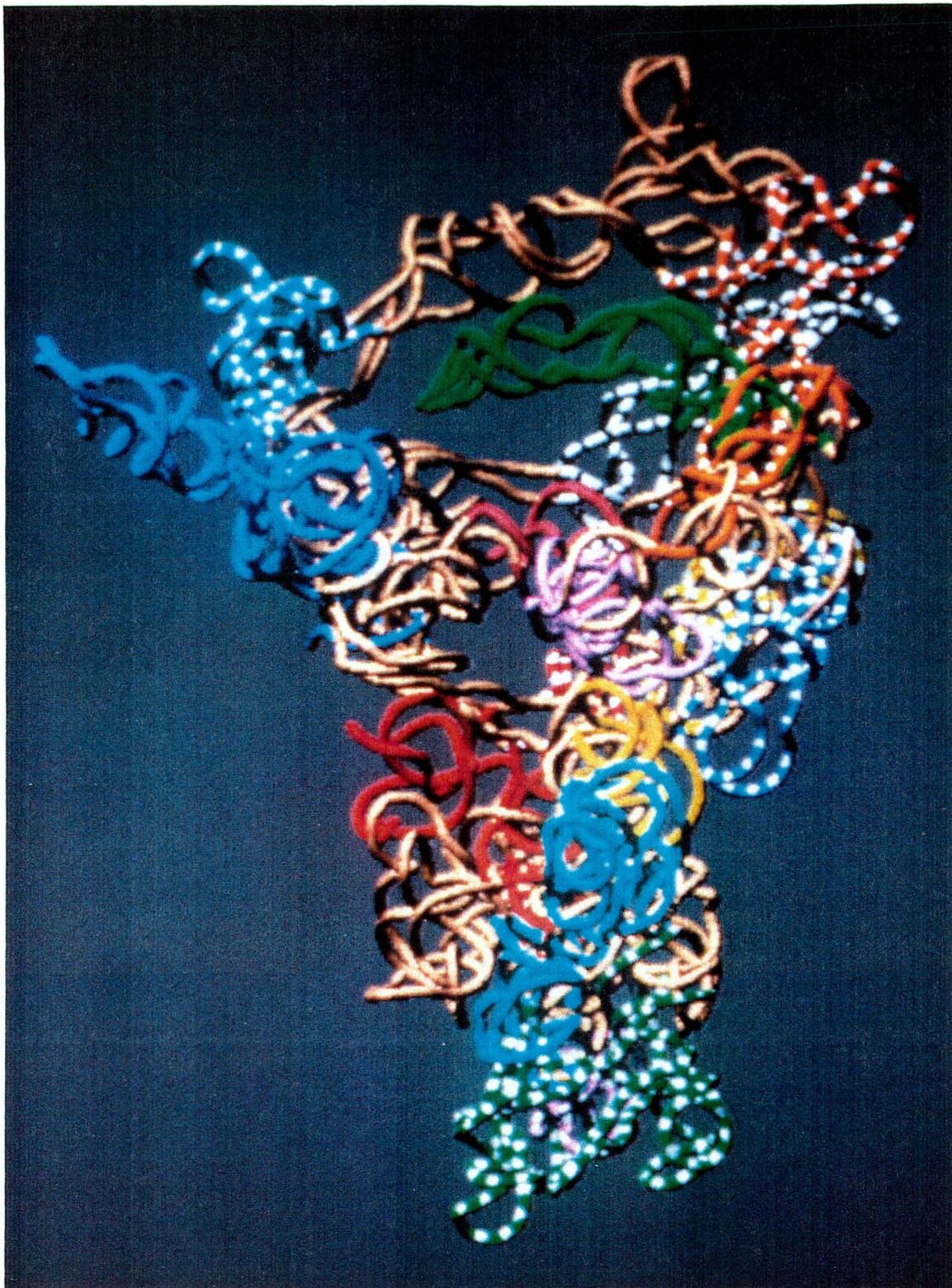
BBC 901 - 828

Figure 20

Rotating the structure 180 degrees about the y-axis reveals the side of the molecule which should face the 50S subunit. Because the raster renderings of the structure retain so much depth information, this new view does not reveal much that is new. It does reinforce the three dimensional character of the structure and it emphasizes how the penultimate blue helix juts out from the surface of the molecule. When this representation of last70 is superimposed upon and simultaneously displayed with the last15 structure, it becomes clear that a family of structures represents a single global conformer and that the superposition errors stem from variations in the precise path of the loops and helices (Fig. 21). Last15 was chosen because although it was the fifth best member of the last70 family, and 12 best overall when judged on the basis of cgr error function, it was 24th out of the 31 independent structures when judged on the basis of superposition onto last70. All of the colored regions appear in the same general areas with many of the helices having practically identical conformations. For example, the green and beige helices at the top of the molecule are not significantly different in light of the simplicity of the pseudohelical modeling constructs. More significant is the diversity shown in the red and orange colored regions which indicate that the model retains substantial degrees of freedom in these regions. Since the red 5' end of the model was one of the major determinants in orienting the molecule, the designation of solvent and 50S faces must be used with even more caution.

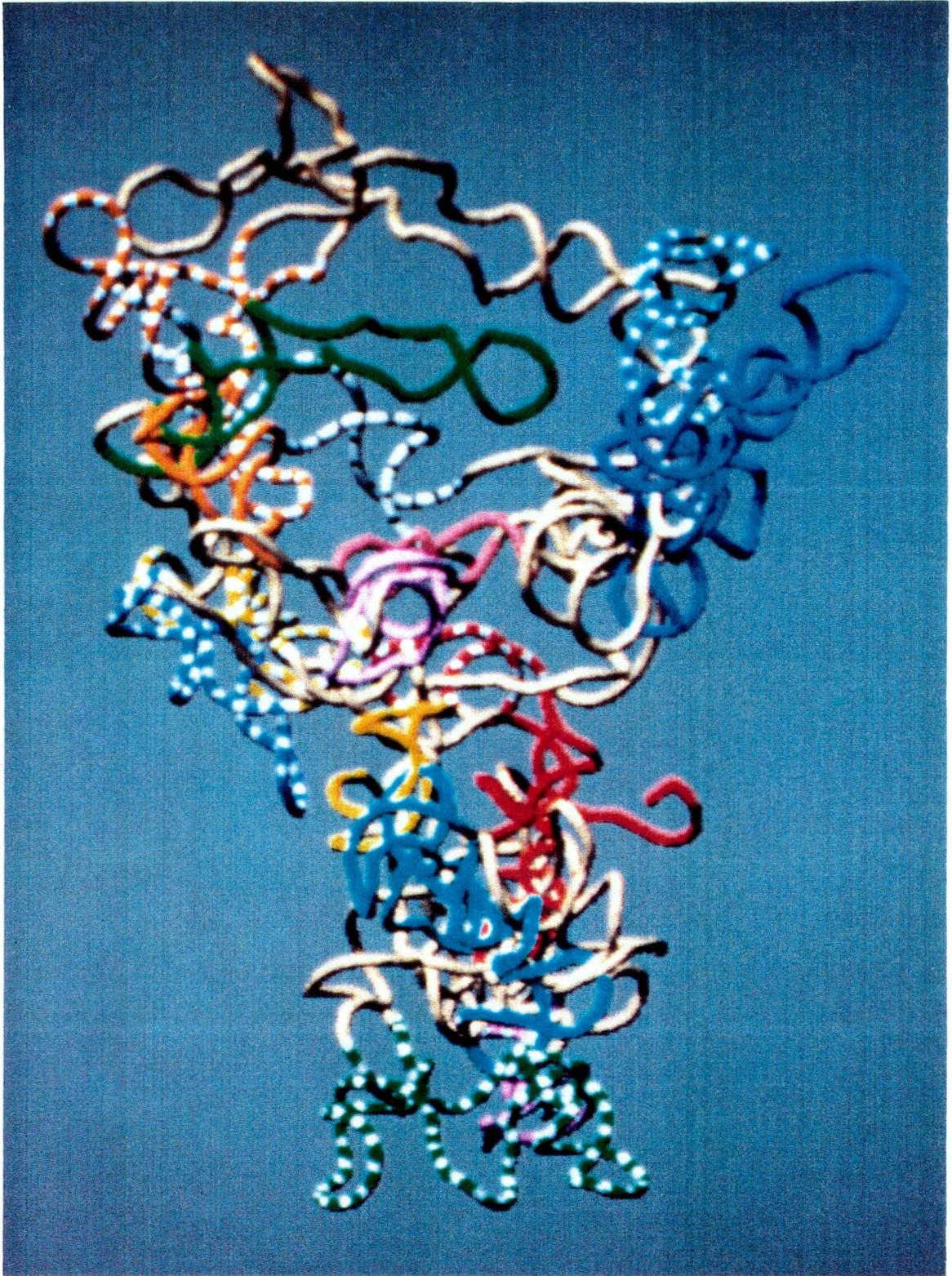
On the following page:

Figure 21. 50S subunit interface orientations of last70 and last15 are superimposed.



BBC 901 - 826

Figure 21



BBC 901 - 814

Figure 22

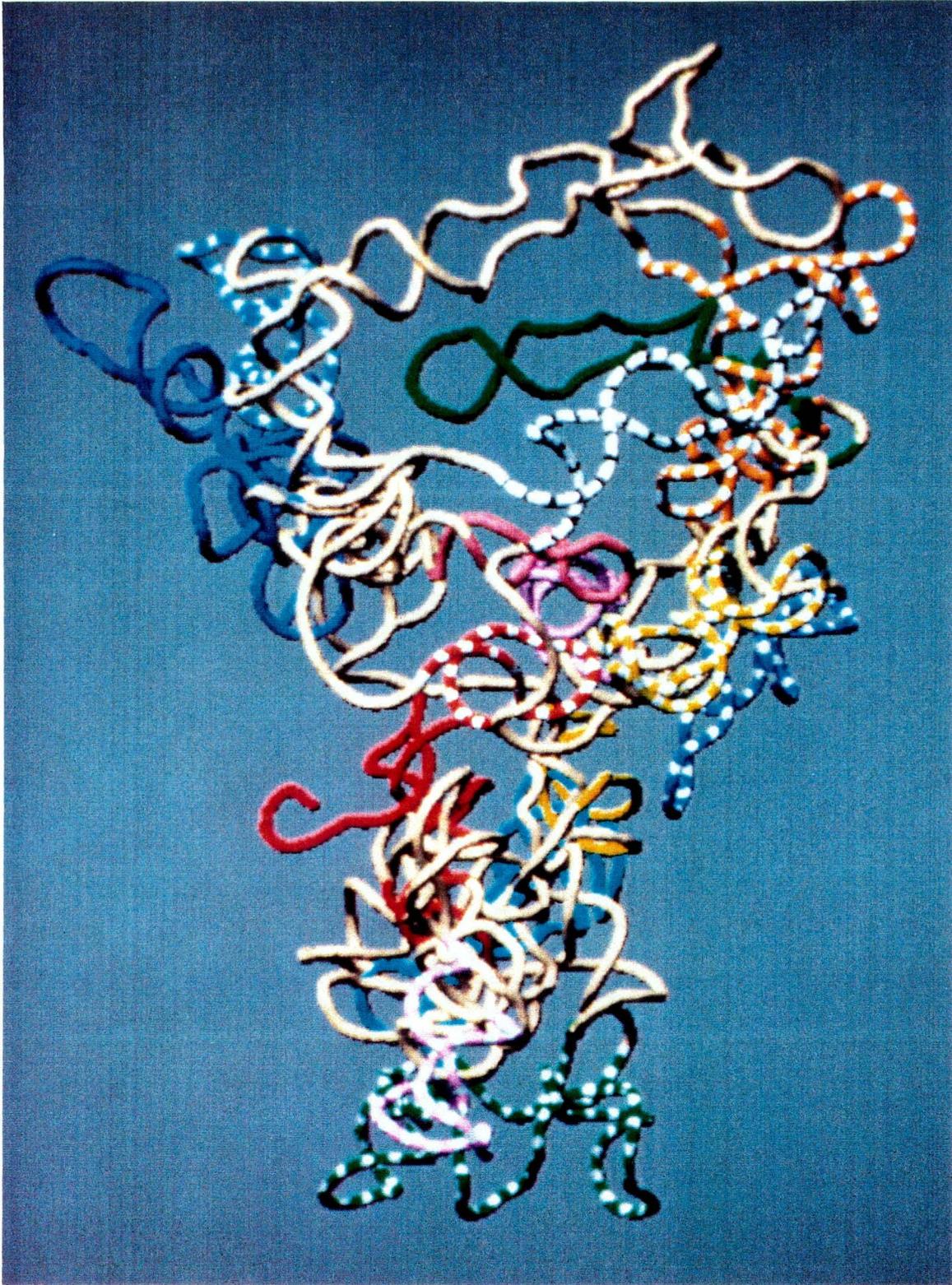
On the preceding page:

Figure 22. The solvent face of the last74 model of 16S RNA.

The raster drawing of the nominal solvent face of the last74 structure shows a great resemblance to the solvent face of last70 in terms of general outline and two dimensional placement of the colored helices (up, down, left, right) (Fig. 22). But a moments inspection reveals that several prominent helices have been dramatically shifted from the back to the front of the display. In particular the green-striped helices now occlude the purple-striped helix at the base of the molecule and the protruding green and blue helices extend out of the page toward the viewer. When last74 is rotated 180 degrees about the y-axis (Fig. 23) and compared with the original display of last70 (Fig. 20), it is clear that these two structures are global enantiomers. This explains the two classes of structures seen in the nt and last series of structures. As the pseudoatoms and pseudohelices used to model 16S rRNA are not chiral, no decision can be reached as to which of these enantiomers is correct without reference to other data sources. Judged on the basis of an internal reference frame, there are no major differences between the two enantiomers and therefore distance geometry is producing a single global folding of 16S ribosomal RNA.

On the following page:

Figure 23. The 50S interface of the last74 model of 16S RNA.



BBC 901 -820

Figure 23

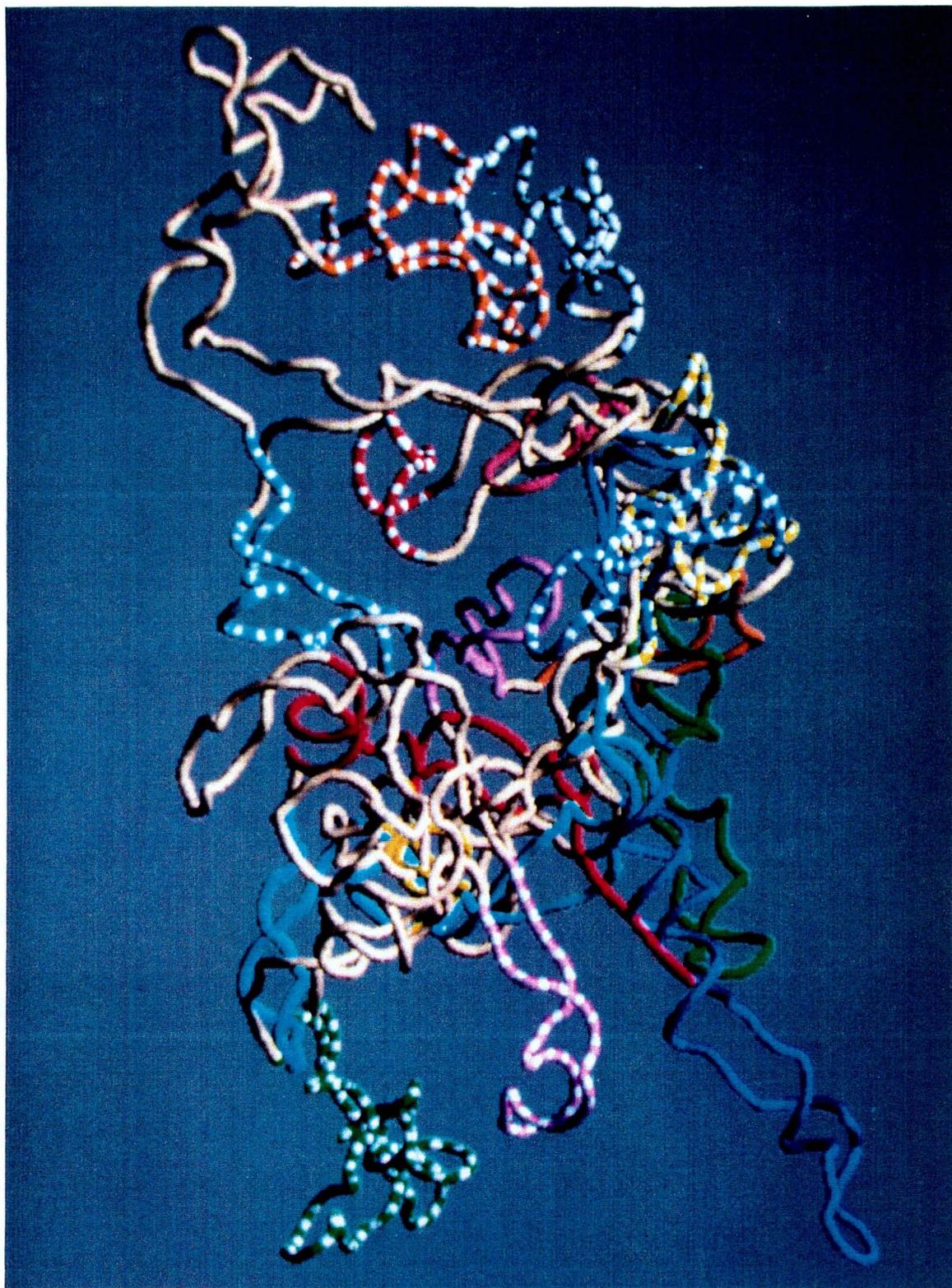
Physical Model (byhand)

The handmade model was developed from a much smaller data set many years ago. The process of converting the coordinates of the physical model into angstroms produced a structure with a volume of 4.85 million cubic angstroms and an initial radius of gyration of 105.02 angstroms as analyzed by AMBER. When compared to the values derived for 16S inside the small subunit, it is clear that the byhand model is not sufficiently compacted. This is clearly the result of the average phosphate to phosphate distance of 6.5 angstroms that was used in the transition from the centimeters of the byhand model to the angstroms of the atomic model. As seen in the modeling of transfer RNA, the structures created by distance geometry can be quite extended and refinement is required to regularize the individual bond lengths and overall structure. For this reason it was decided that leaving the byhand model in an extended conformation, rather than uniformly reducing the model size by some scalar factor, would provide a better comparison to the computer models as the byhand model is converted to helical pseudoresidues and refined.

After the conversion into the pseudohelical representation and refinement, the byhand model has a volume of 2.02 million cubic angstroms and a radius of gyration of 78.37 angstroms. Its dimensions of 230 X 230 X 90 angstroms with a largest dimension of 310 angstroms indicate that the byhand model is still a little too extended and could easily tolerate the further refinement which would bring it into close agreement with both the real molecule and the distance geometry generated structures. A color raster display of the byhand model was created (Fig. 24) which corresponded to the orientation of the

On the following page:

Figure 24. The raster display of the byhand model of 16S RNA.



BBC 901 - 836

Figure 24

original black and white line drawing (Fig. 9). The red, purple-striped, green-striped, yellow, and cyan helices of the 5' domain are clustered in the lower lefthand quadrant of the model. The linear extension of the red 5' end into the right side of the structure was an arbitrary modeling decision and should not be misinterpreted as a reflection of any physical data. It does indicate just how far a flexible single-stranded section of RNA might reach. The green, orange, purple, blue-striped, and yellow-striped helices of the middle domain form a cluster on the right side of the structure. The cyan-striped, orange-striped, black-striped, red-striped, and rose helices of the 3' domain are much more loosely associated, in contrast to the other two domains, and form the top half of the structure. The blue section of the molecule which contains the bases beyond 1400, is folded separately from the rest of the 3' domain into the lower righthand side of the model.

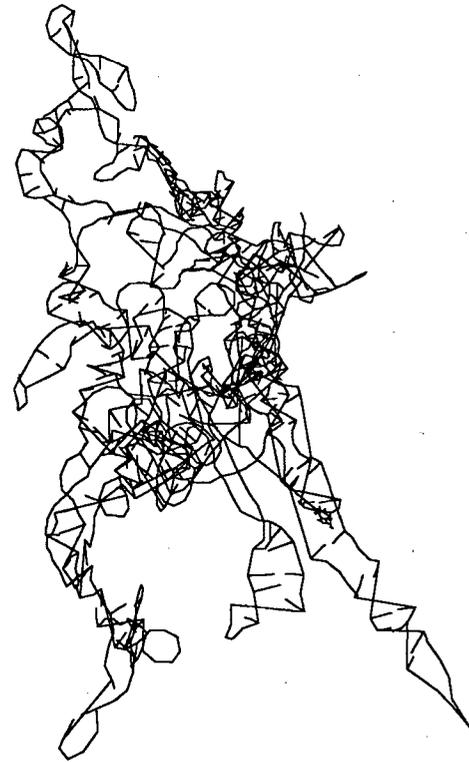
That the overall shape of the byhand structure is not as suggestive of the shape of the 30S subunit is primarily due to the purple-striped, green-striped, and blue helices which form the base of the structure. The exigencies of the physical modeling process and the dearth of data available at the time, resulted in a structure that was open and extended. This shortcoming could be easily adjusted to reflect the stacking tendencies of RNA in solution if the single-stranded regions that connect the lower helices to the body of the structure were placed in curved instead of a straight conformations. The byhand model does satisfy all the long range relationships used to produce the DSPACE models except for the psoralen crosslink between U956 and U1506. This crosslink is violated by only two angstroms and further refinement would easily bring these two portions of the 16S molecule into agreement with experiment. This could be accomplished by bringing the beige strand traversing the molecule just above the red-striped region down towards the section of the blue strand which is closest to the middle of the model.

Extensive DSPACE refinement of the byhand model was not pursued as it might have overridden the original modeling decisions which were made before the complete protein mapping and most of the nonpsoralen crosslinks were determined. Therefore the

conjugate gradient refinement was halted after 128 minimization cycles just as in the refinement of the distance geometry foldings. As the byhand model did not start out with the gross distortions of the DSPACE models this was sufficient to reach the proper error function neighborhood as judged by the magnitude of the bounds violations. In fact the cgr error function of 1084 square angstroms and the bounds violations total of 2768 angstroms for the byhand model approximated those at the high end of the structures created by distance geometry. But even with this limited refinement, substantial changes in the structure are apparent when it is compared to the original conformation of the byhand model (Fig. 25). The superposition fit error of 34.86 angstroms and an RMS deviation of 1593 angstroms between the two indicates that the byhand model had some serious shortcomings when analyzed on the basis of helical constructs and distance constraints. The prickly appearance of the initial pseudohelical byhand model is the result of the manner in which AMBER has inserted the bases. The nucleotides of the AMBER databases have dihedral angle preferences which favor a gently winding single-stranded conformation. When superimposed on the flat basepaired regions of the byhand model, the bases end up pointing out into solution instead of towards their hydrogen bonding partner. The secondary structure constraints contained in the bounds matrix correct this defect. The overall compression of the model and the helical twisting of the double-stranded regions were the main reasons for refining the structure and these have been achieved. The folding of the 3' domain down into the body of the structure was an unwanted effect of the tertiary constraints used to produce the computer models. Therefore to some extent, the byhand model must be considered a hybrid of a model constructed by physical means from minimal data crossed with a computerized model based on more recent data.



Digitized byhand model



Byhand model after refinement

Figure 25. Effects of refinement on the byhand model.

Discussion

Modeling Difficulties

The sheer size of 16S ribosomal RNA even after reduction to pseudoresidues was the source of most of the problems in this modeling protocol. It was the exhaustion of the available computer resources which forced the more straightforward method of converting the physical model to a computer model to be abandoned. The problems introduced by the transitions between various levels of representational abstraction have been mentioned previously, but the difficulty of dealing with the full structure of 16S RNA also led to significant alterations in the structure refinement and final display steps of the modeling protocol.

Refinement Protocol Adjustments

Only partial minimization with the DSPACE conjugate gradient refinement was done for two reasons. Most obvious is the over compaction that was seen in transfer RNA modeling. Especially dangerous is the possibility that the cgr minimization of the reduced structure might result in knotting. As the helices have spacefilling pseudoatoms only at their ends, it would be possible to move other sections of the molecule through the space occupied only by the bonds of the longer helices. For example, the hairpin formed by 829-857 has a helical stem of 12 base pairs. The ends of this helix are separated by approximately 40 angstroms, wide enough to allow an all-atom representation of an A-form RNA helix to lie in the middle of the pseudoatom structure with room to spare. Even partial minimization of a pseudoatom structure might produce such problems. At the present state of knot theory it is not possible to evaluate a structure for knots if it has more than ten to twenty crossings. Manually tracing the chain path with interactive graphics is time consuming and requires some expertise. It could be used to evaluate one or two structures but is impractical as a means of evaluating a large number of structures. Relying on AMBER to indicate when a structure is knotted is also a very slow process with the

present hardware and as the tRNA results show, AMBER may resolve conflicts to the detriment of the structure as a whole.

As AMBER reinserts the atoms of a residue based on its library of standard structures, it may introduce conflicts by accidentally superimposing bases. This would certainly explain the enormous vdW problems that AMBER detects. Particularly in the very tight 5' domain of 16S rRNA, placing bases at standard values may produce impossible structures. The manner in which the EXPAND program superimposes a standard helix could also be creating serious conflicts. Overlaps can produce free energies in excess of ten to the 38th power calories. This exceeds the floating point precision of the corresponding variables in the MIN modules of AMBER and caused the program to crash, often at the cost of days of computation. It was not practical to adjust the source program to carry larger values. Instead MIN was amended to detect such data overruns, report the problem to the user and set the particular interaction to the largest value that could be carried forward. Unfortunately AMBER proved ill-suited to handle such a large structure under these conditions. A plot of the phosphate to phosphate distances along the backbone shows the artificial uniformity which must result from the use of pseudoatoms and ideal replacement helices. The one significant deviation from the average is due to a sharp change in direction at the junction to the independently folding segment 588-651 (Fig. 26). After ten days of AMBER minimization this very poor bond has been shortened by an angstrom, but in attempting to resolve van der Waals or other conflicts, AMBER has begun to introduce some equally unappealing local conformations (Fig. 27). It is striking that the program distorts the bond lengths away from the ideal value of 5.75 angstroms in a pairwise fashion. If a phosphate to phosphate separation is increased to an unrealistic value of greater than eight angstroms, the neighboring phosphate to phosphate distance will be decreased to an equally unrealistic three angstroms. Consequently the average phosphate to phosphate is unchanged even though the quality of the structure may be degraded.

Figure 26. last74 before AMBER

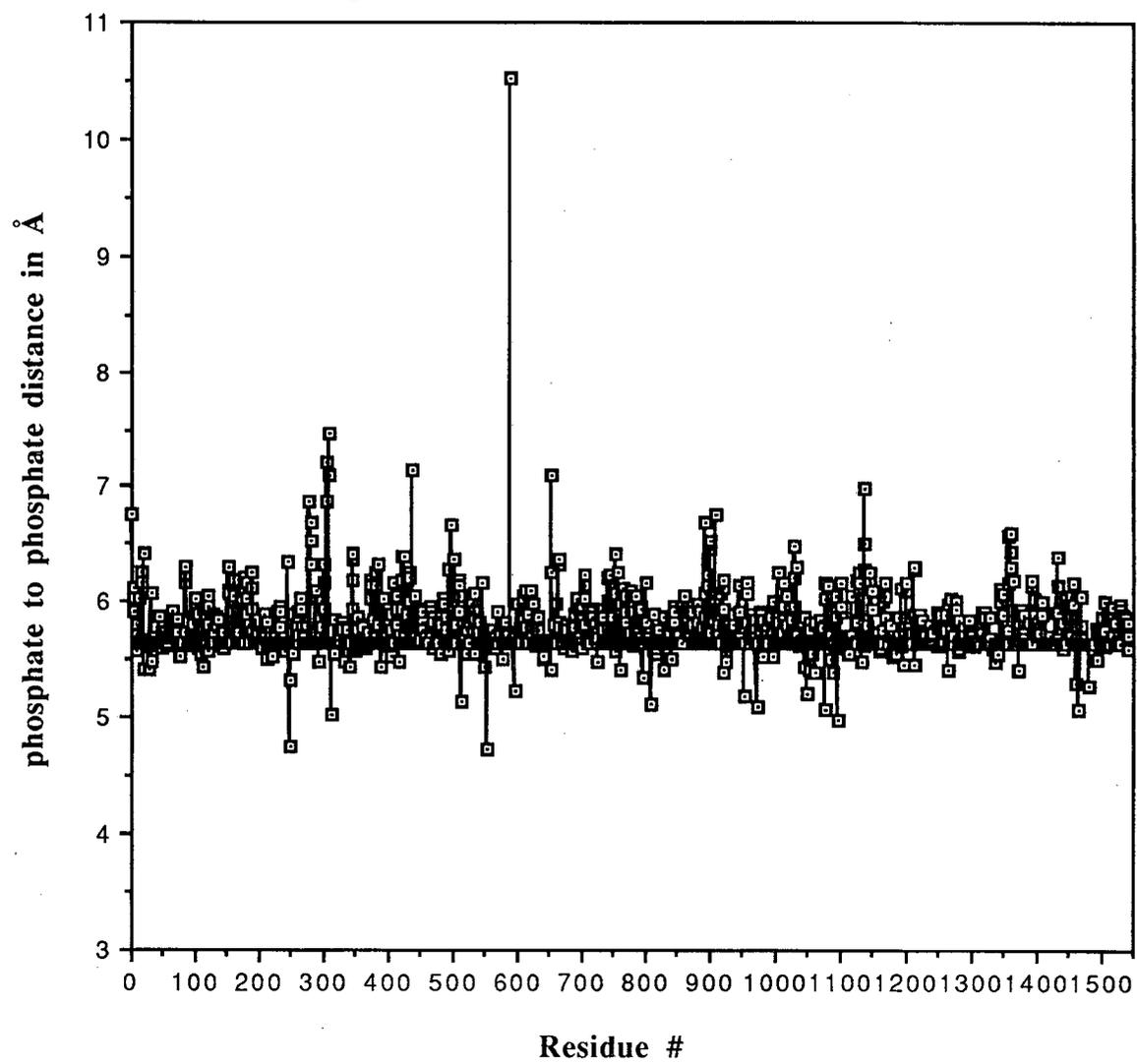
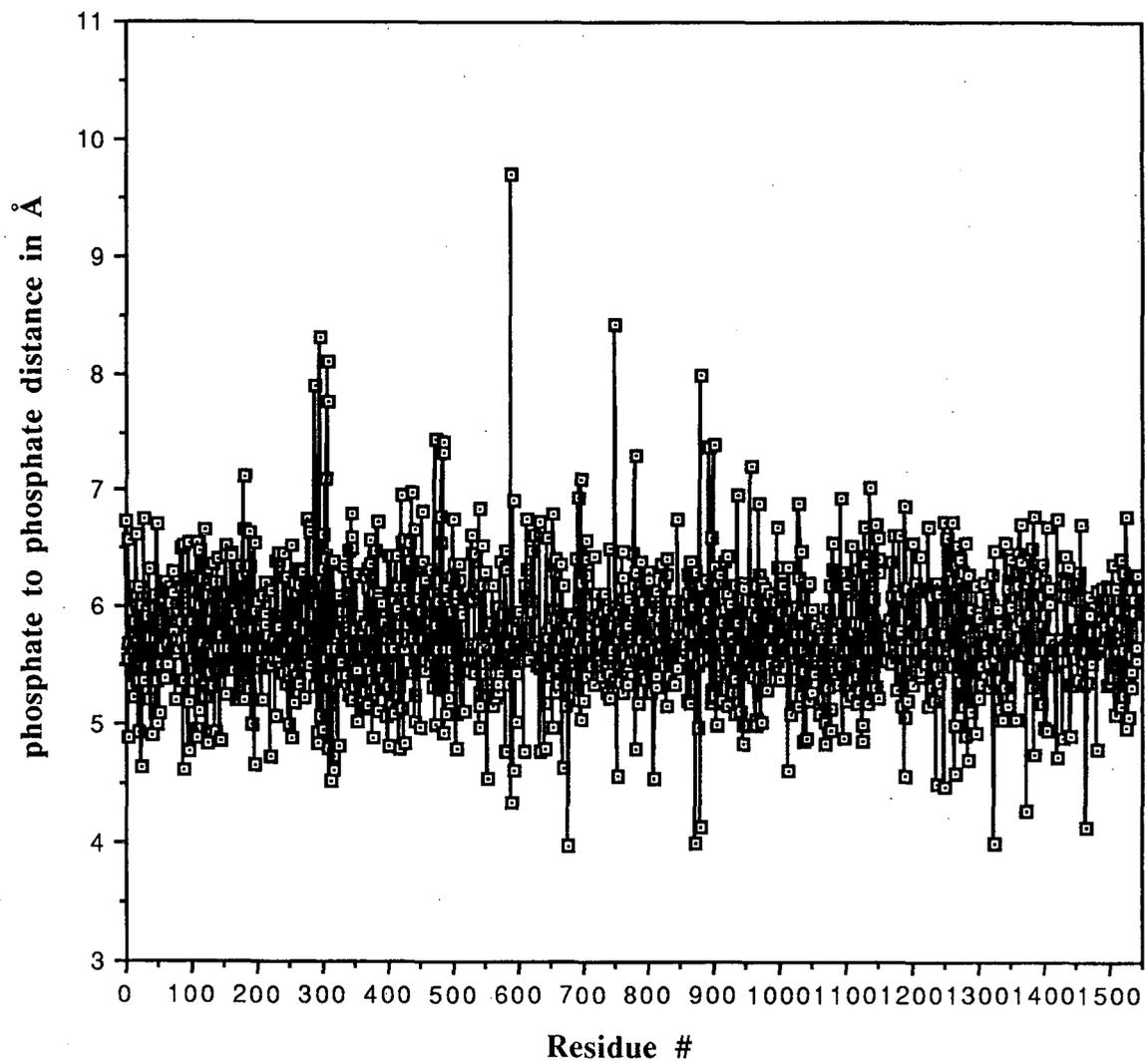


Figure 27. last74 after 10 days of minimization



Display Adjustments

The black on white line drawings are clearly inadequate for more than the most general alignment of the 16S structure. To even begin to locate the various regions and evaluate their relationships would require that the researcher painstakingly display only selected portions of the structure for a particular view of a particular structure. Analyzing even one structure, much less 100, could take weeks. Fortunately color can be used to uniquely label and differentiate the complex internal structure of the folded 16S model.

In deciding which orientations of the folded molecules to display, the longest axis was chosen to be the y-axis. In this orientation it is immediately obvious that the model is tapered and by putting the largest portion at the top, the resemblance between the model and the gross outlines of the 30S particle as determined by Immunoglobulin Electron Microscopy are emphasized. By marking certain key areas of the molecule (5' end, 3' end, S4 binding domain, platform region at ~1400) it is possible to rotate about the y-axis until the faces of the model which best correspond to the 50S and solvent surface are displayed. The vertically oriented presentation made it easier to rotate the molecule in a predictable manner and simplified the production of high quality raster pictures on the PS340 display. Unfortunately when visually comparing the structures to the models of other researchers, this vertical orientation can be confusing as the 16S rRNA appears to be tilted from the vertical axis in the 30S subunit. An additional complication results from the lack of depth in the published pictures of the other models. This flatness is appropriate when the general outline of a structure is compared to the electron micrograph outline of a 30S subunit adsorbed onto a carbon grid. But as the three dimensional statistical reconstructions of the small ribosomal subunit show, there is a wealth of detail that is absent in a simple silhouette. Raster displays of the models are so superior a method of representing the full dimensionality of the various conformers that the use of computer graphics is unavoidable.

An unsophisticated visual evaluation of the structures can be misleading as it is difficult to judge what level of conflict may exist in the raster pictures. The program, MCS,

which creates the tubes uses a smoothing radius to create average graphical toroids. In exploring the order and chirality of the objects produced by MCS, an extremely large version of tRNA, with tube diameters of ten angstroms, was created. At this size the joints between the graphical subunits becomes obvious and the distortions away from the true chain path become pronounced. The smoothing radius for the much smaller phosphate backbone should not produce such egregious errors but a simple visual comparison of the loops in a black and white line drawing and the matching color picture indicates that care must be exercised. For example the purple-striped helix in the lower center of the color version of the byhand model (Fig. 24) appears to have serious van der Waals and kinking problems that are not seen in the line drawing (Fig 25). Both the raster and line drawings make it appear that some regions of the structure are very open and that it is possible to pass through the structure. While this makes it easy to see how the back of the model is folded, it does not convey a true sense of the spacefilling nature of the model. To minimize these weaknesses and to emphasize the helical structures which are the central constructs of all the models of 16S ribosomal RNA, a new raster display scheme was developed. In addition to the colored tube which traces the backbone, cylinders of twenty angstroms diameter were created for each of the well defined helices of the secondary structure map. These correspond in color with the sixteen unique helical regions. Beige cylinders were included for the helices formed by the basepaired segments 122-128:233-239, 511-517:534-540, 923-833:1384-1393, 946-955:1225-1235, 984-990:1215-1221, 1063-1067:1189-1193, and 1168-1073:1102-1107.

On the following page:

Figure 28. The solvent face of last70 displayed with open cylinders denotting helical regions.



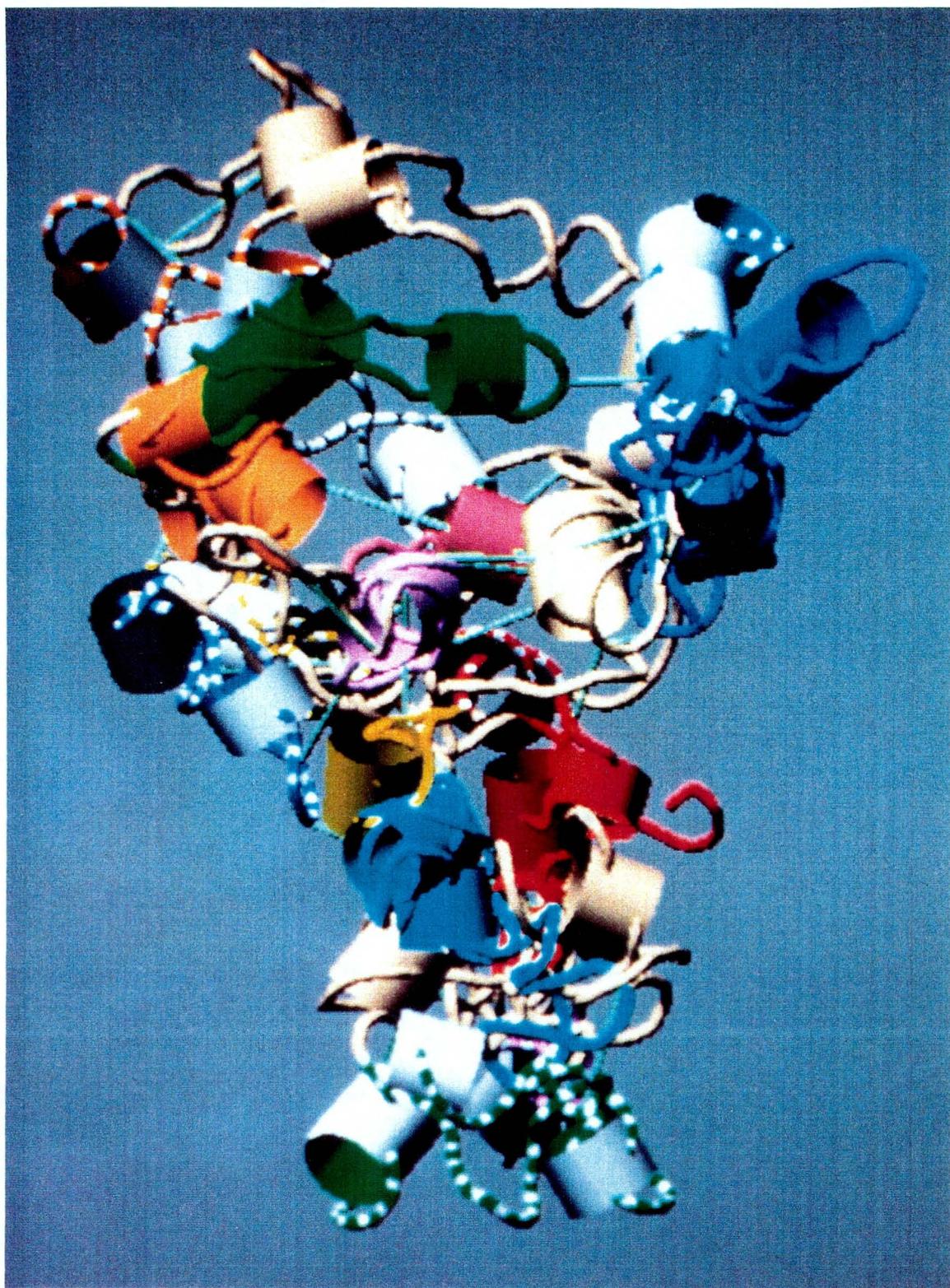
BBC 901 – 830

Figure 28

When the solvent face of last70 is displayed in this manner, the sense of depth possessed by the earlier models is enhanced at the expense of obscuring some regions (Fig. 28). The cyan, yellow, and purple helices are almost completely hidden. On the other hand, the manner in which the yellow-striped helix juts out towards the viewer is sharply realized. Light blue vectors representing tertiary folding constraints have been added to the tube and cylinder display of the solvent face of last74 (Fig. 29). Comparing this picture to the three color and crosslinks version of last74 (Fig. 18) demonstrates that the raster display is becoming overloaded with information. Even with the opposite side of last74 (i.e. the 50S subunit face), it is no longer possible to follow all the crosslink vectors (Fig. 30) and this indicates that a practical upper limit on the amount of information which can be displayed at one time has been reached. When the tube and cylinder version of last70 (Fig. 28) is compared to the near mirror image last74 (Fig. 30), it becomes clear that many of the helices have very similar orientations in addition to their similar locations. In fact it is easier to note that the 5' red helix, the beige helix just below the red helix, and the topmost orange helix have different orientations than it is to note those that are similar.

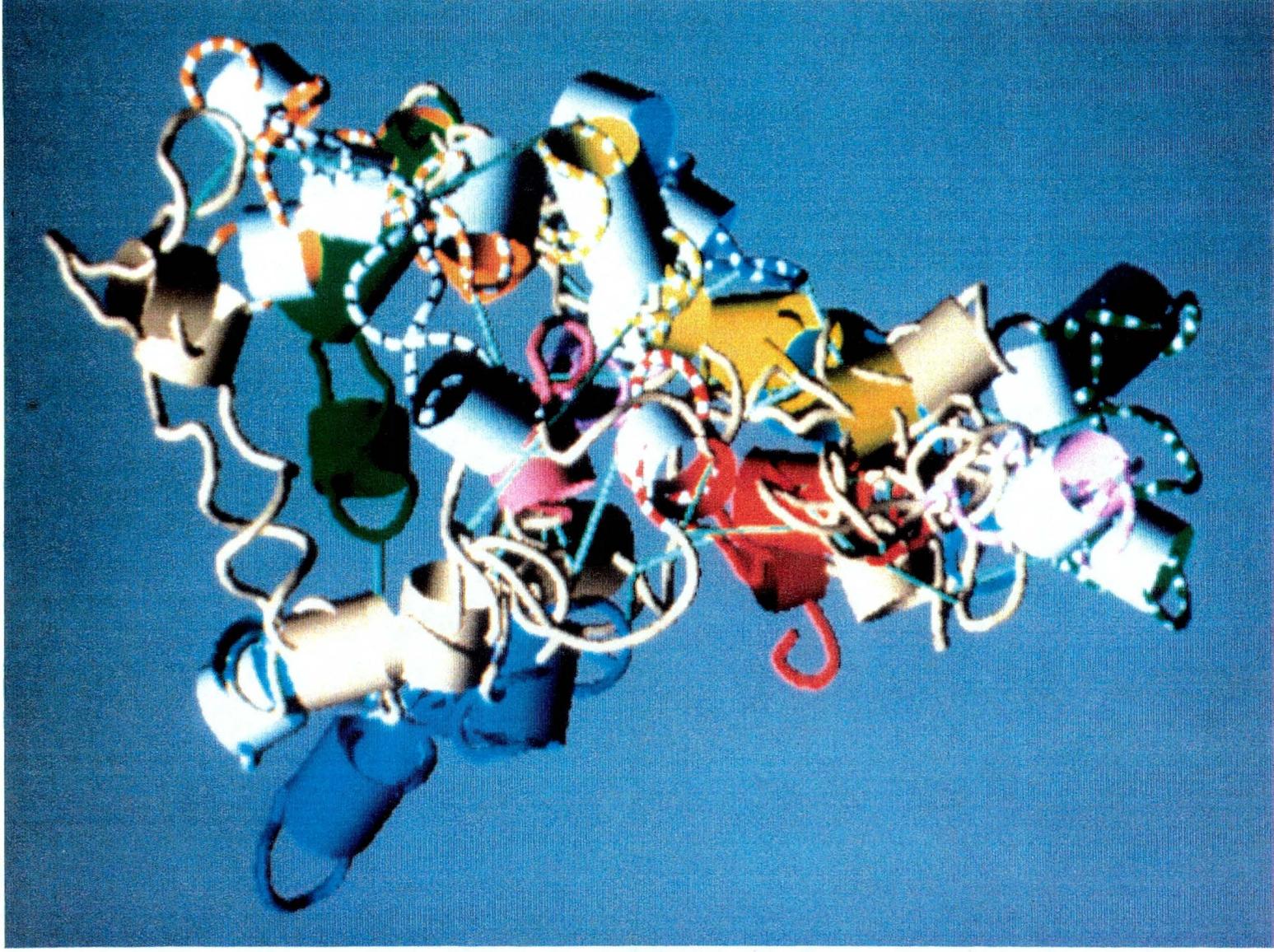
On the followin page:

Figure 29. The solvent face of last74 displayed with open helices indicating helical regions and light green tubes connecting residues with long range relationships.



BBC 901 – 816

Figure 29



BBC 901 - 822

Figure30

On the preceding page:

Figure 30. The 50S subunit of last74 with open cylinders for helices and light green tubes indicating long range relationships.

Evaluation of Models

In evaluating the models produced by the computer protocol it is important to keep in mind the simplicity of the underlying modeling constructs and the limited nature of the tertiary folding data. Attempting to say anything very detailed about a particular nucleotide or helix would be an overextension of the analogy. But the models must meet standards which have a similar or lesser resolution. Comparisons will be made to the overall physical shape determined by electron microscopy and to the other models of 16S RNA. Comparison to data derived from the 30S subunit proteins will be done in the next chapter.

It appears that distance geometry compacts the 'Y' conformer produced by secondary structure information alone by folding the extended helical regions in towards the middle. With respect to the plane which the majority of the molecule inhabits, this folding can occur to one side or the other. As imaged by the electron microscope, the molecule should resemble a cupped hand with the concavity facing the large subunit (Lake, 1985). The enantiomeric structures suggest that the two classes result from the folding of the flat structure toward one side or the other much like that of an ambidextrous oven mitten. In general outline the structures produced by distance geometry correspond very nicely with the shapes seen in electron micrographs. The structure produced by statistical reconstructions from electron micrographs is a much rounder, three dimensional cone which resembles a pitcher plant (Vershoor, et al., 1984). The color raster displays of the

16S models exhibit an even stronger match with this conical image. Special stereo views of both last70 and last74 reveal a structure that looks like the electron micrograph reconstructions, lacking only the central knob. Further folding of the RNA or the addition of the small subunit proteins might well fill in that deficiency. Therefore on the basis of general appearance and the gross physical characteristics as detailed in the Results, the structures produced by the computer modeling protocol are quite reasonable.

The computer generated models are also compatible with the new three dimensional phylogenetic relationships that were published too late for inclusion in the modeling process (Haselman et al., 1989). Most of these relationships involve fine tuning of local interactions but the tertiary basepair between A994 and U1380 should have significant conformational ramifications. A994 is at the base of the cyan-stripped region and U1380 is in the single-stranded region between the rose and beige helices. In the computer generated structures last70 and last74, these bases are found in the same quadrant, with A994 poised above U1380. Although the distance between these two bases is approximately 50 angstroms too large for basepairing to occur there is a direct, unobstructed vector connecting them such that a trivial adjust to both or either region would bring them into proper alignment. This result is in line with the nature of the computer protocol which will not introduce unwarranted structure.

General comparison with other models

The most obvious difference between the pseudohelical models and those of other researches are the helices formed by the residues 588-651 (green) and 1409-1491 (blue). The other models have placed these helices in the lower half of the subunit based on two arguments. The first suggests that the nodes at the bottom of the eucaryotic small subunit are due to the additional eucaryotic nucleotides which extend these helices. This argument is logical and attractive but is not sufficiently forcing. There are other nucleotide insertion sites which may be the cause of these bumps even if we ignore the possibility that these bumps may be due to protein differences. The superior argument is based on the three

crosslinks between the green helix and the S8 small subunit protein that have been found. As the S8 protein appears to bind to the lower third of the 30S particle, this would be sufficient evidence if protein data could be included. Lacking the quantitative data required for the creation of distance bounds, these helices have been left free of constraints. In the absence of such constraints, these helices protrude from the body of the computer generated models. The variable length helices formed by residues 65-104 (purple-striped), 136-227 (green-striped) and 406-496 (cyan), do appear in the bottom of all the computer generated structures and increases in their lengths could produce the nodes at the bottom of eucaryotic ribosomes. The orange-striped (1113-1187) and black-striped (1241-1295) regions are ideally placed in the upper domain of the structure such that increasing their lengths would produce the beak-like structures seen in some ribosomes (Lake, 1985).

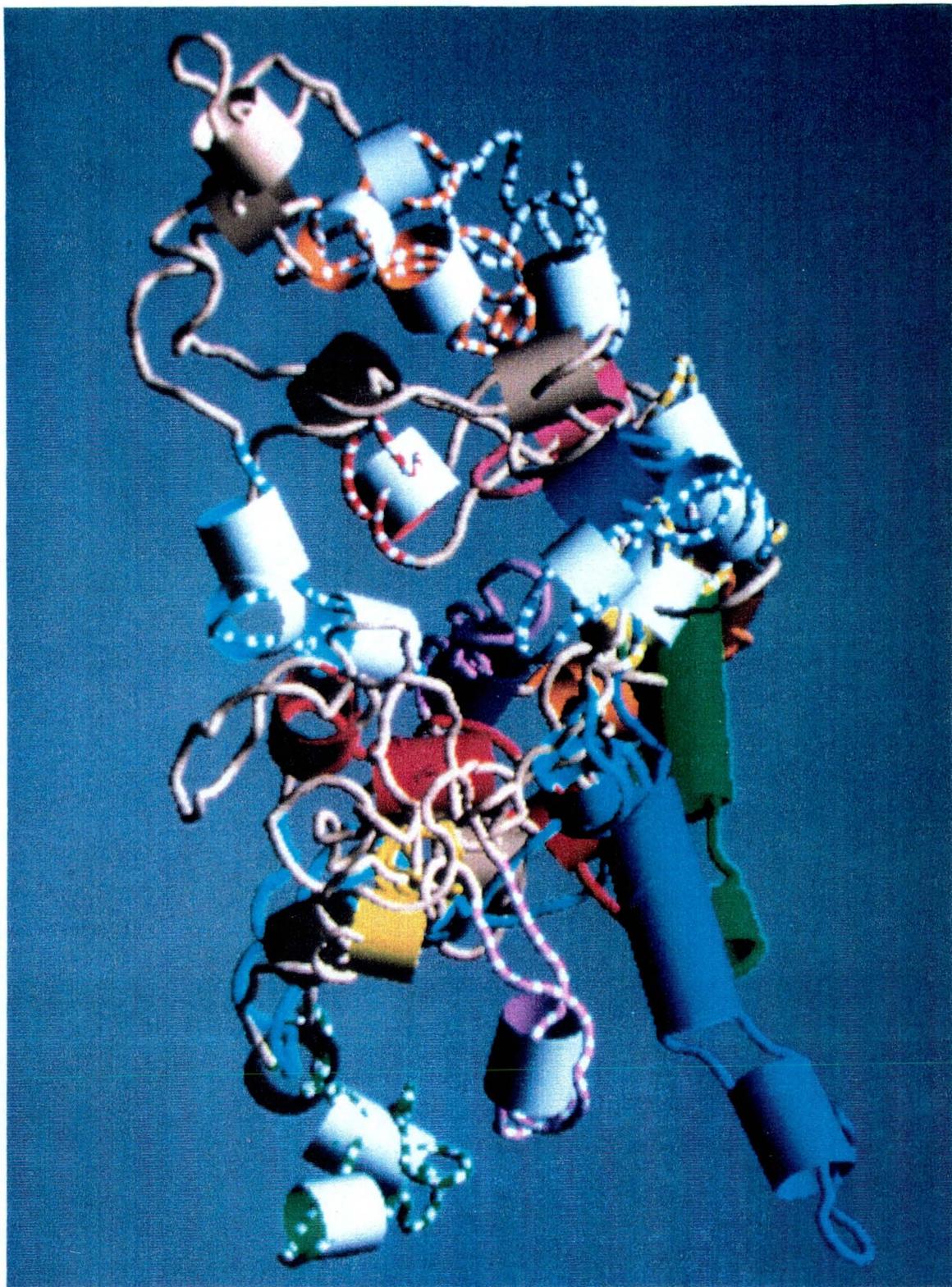
The protrudant character of the green and blue helices does suggest that they should be compacted in some fashion. A-form RNA helices have a diameter of approximately twenty angstroms and are detectable in electron micrographs. As none of the EM studies of the small subunit possess obvious RNA helices jutting from their surface, the green and blue helices must have some other conformation in the ribosome. They could easily be folded down to correspond with their placement in the other models. Using the single-stranded regions at 570-575, 811-819, and 1394-1399 as axes, these helices can be rotated smoothly into a position consistent with protein crosslinking data. The resultant conformational energies are similar to those produced without this manual manipulation. This additional folding would produce a clearly defined platform corresponding to that seen by electron microscopy.

The other major difference among all models appears in the orientation of helical building blocks. Suspected instances of coaxial or parallel helical stacking are important determinants in the production of the models of other researchers. Such orientational information is buried in the pseudohelical constructs employed in this distance geometry protocol but is not specifically invoked to dock any helical region. Consequently even

where the correspondence between another model and the computer generated models appears to be very high, the helical cylinders may have very different orientations. Certainly the variation in helical orientation despite the high degree of similarity in the three dimensional location of some helical segments in the distance geometry structures, argues that an exact helical orientation is neither well supported or necessary in producing reasonable conformers. Therefore the models will be compared on the basis of the location and not the orientation of the important helical determinants.

On the following page:

Figure 31. The 50S interface side of the byhand model displayed with open cylinders indicating helical regions.



BBC 901 - 838

Figure 31

Distance geometry models versus the byhand model

The byhand model occupies a different region of conformational space from that of the structures produced by distance geometry. It differs from the two classes of structures produced by distance geometry to the same degree that those two structures differ from each other. When superimposed on last70, the byhand model has a fit error of 57.96 angstroms and an RMS deviation of 2388 angstroms. When compared to last74 the fit error is 59.41 angstroms and the RMS deviation is 2471 angstroms. Visually the byhand model (Fig. 31) can best be compared to the 50S subunit side of last74 (Fig. 30). The comparison is easier if the byhand raster picture is rotated some 30 degrees about the z-axis in a clockwise manner. In this orientation the agreement between the two models is very good. In the 5' domain the match of the red helices is fair even though the 5' ends point in different directions. The yellow helix is the only miss in this domain, appearing on the left in the byhand model and on the right in last74. All of the helices of the middle domain show fairly good agreement with their counterparts in last74 with the exception of the yellow-striped helix which appears to occupy a mirror position at the back of the byhand model. Of course the green helix is not folded down in the last74 model but there appears to be a simple rotation path that would fold this helix into superposition with the byhand model. In the 3' domain the cyan-striped helix is much nearer the waist of the molecule in the byhand model than it is in last74. The distance between the predicted tertiary paired bases, A994 and U1380, is 80 angstroms too large. Adjusting the model to accommodate this interaction would not be trivial as it was in last70 and last74 and would require even more compression of the 3' domain. The match of the blue helical regions is even worse than anticipated. There does not seem to be any means of bringing these helices into similar conformations, short of a major refolding of much of the structure. The rest of the helices show good agreement although it must be remembered that this is the domain of the byhand model that was most affected by the refinement process.

The comparison of the distance geometry and byhand models with those produced by other researchers is very difficult. Even with access to more explicit representations than those which were published, finding common orientations and referents is very time consuming. Copies of the published models were colored in an attempt to expedite the process, but the lack of depth possessed by these pictures is a serious impediment.

Physical Models

The physical modeling protocols of Wollenzien and Brimacombe resemble the packing of a suitcase, where the overall shape seen by electron microscopy is the form into which the linked helices of 16S rRNA are stuffed. The constraints on the packing come from attempting to match particular portions of the RNA with the positions on the surface of the small subunit that have been mapped by Immunoglobulin Electron Microscopy. RNA/protein associations and crosslinks are also used to link portions of the RNA with the protein maps of the small subunit developed by IEM and neutron diffraction studies. Flexible tubing or plastic cylinders are used to represent the RNA at some fixed ratio and styrofoam balls are used to represent the proteins of the small subunit.

Wollenzien Model

The model produced by Expert-Bezaccon and Wollenzien was constructed from a later secondary structure map than that used for the byhand model and a data set that included the GbzCynAc crosslinks, but without the complete set of RNA/RNA crosslink data or the completed map of the 30S proteins. The published version of this model approximates the 50S subunit side of the molecule and it most resembles that side of last74 (Fig. 30). The red helices of the 5' domain are not visible from the 50S side in the Wollenzien model making comparison difficult although the positioning of associated helices makes it clear that the position of the 5' end will not be irreconcilably different from that of last74. The purple-striped and green-striped helices appear in the foot of the structure as they do in last74. Some rotation about the y-axis might bring the cyan helices into better agreement, but the yellow helices are on opposite side of the waist in the two

models. In the middle domain, the blue-striped and yellow-striped helices are similarly placed. If the substructure formed by the green, orange, and purple helices of last74 is folded straight down, it will bring the middle domain into very good alignment. The correlation of the helices of the 3' domain is good excepting the jutting blue helix of last74. But as in the middle domain, there is an unimpeded rotation path about a flexible single-stranded region that would bring it into agreement with the Wollenzien model.

Comparing the Wollenzien model with the byhand model requires a tilting of the byhand display to maximize the fit. In addition a slight rotation about the long axis of the byhand model is also necessary. Once these adjustments have been made it is clear that the agreement between these two models is very high. All the helices of the 5' domain are similarly positioned. In the middle domain only the purple helix is differently placed. The Wollenzien model places this helix on the back right side of the waist of the molecule while this helix appears in the in the middle of the back of the byhand model. The red-striped and rose helices are folded tightly down into the middle of the Wollenzien model. These two helices were partially compacted towards such a position during the distance geometry refinement in the byhand model. Further folding in this manner would bring them into close agreement with the Wollenzien model. All other helices of the 3' domain show good superposition.

Brimacombe Model

The model constructed by Brimacombe and coworkers (Stern et al., 1988) used a secondary structure map developed in his lab that differs only slightly from the latest version created by Noller and coworkers. The Brimacombe model had access to the protein map lacking only protein S21 and the latest RNA/RNA and RNA/protein relationships. Only information which was derived from the intact 30S small subunit of the ribosome were used, specifically the psoralen and GbzCynAc crosslinks were excluded. The gross dimensions of the Brimacombe model (220 X 140 X 90 angstroms) are similar to those seen in electron micrographs and the computer generated structures. Pictures of

both the solvent and 50S subunit side of the model were published and again it appears that last74 is a better match than last70. Comparison of this model with the ones generated by computer are most charitably characterized as mixed. In the 5' domain, the red, green-striped, and yellow helices occupy similar positions, but the purple-striped and cyan helices are on opposite faces of the model in mirror positions. In the middle domain the blue-striped and yellow-striped helices are closely related and the usual downward folding of the green helix might bring both it and the orange helix into the same area of last74 that they occupy in the Brimacombe model. Postulating what such changes might produce is dangerous because there is a temptation to slightly rotate the molecule to one side or the other to improve the fit. In this case improving the fit of the green and orange helices would seriously degrade the fit of the red helices of the 5' domain. Of the 3' domain helices only the black-striped and cyan-striped helices of last74 show any correspondence with their counterparts in the Brimacombe model. Although the other helices of the 3' domain, save the blue helices, do form the head of the Brimacombe model their folding is completely at variance from last74.

The byhand model shows a much greater resemblance to the Brimacombe model. Unlike previous comparisons the match between these two models is better when the byhand display is left with its longest axis parallel to the y-axis. The red and the purple-striped helices of the 5' domain appear to occupy mirror positions in the two models while the other three helices of this domain have similar positions if not similar orientations. In the middle domain only the purple helix is placed differently, being in the middle of the byhand model and on the upper right hand side of the Brimacombe model. In the 3' domain the Brimacombe model placed the red-striped and cyan-striped helices at the very top of the structure. These helices are packed beneath the black-striped and orange-striped helices in the byhand model. In the Brimacombe model the green helix of the middle domain faces the 50S subunit and the blue helix of the 3' domain is packed just behind it. These helices

occupy the same region of the byhand model but the blue helix faces the 50S subunit and the green helix is behind it. The other helices of the 3' domain are in agreement.

Computer Models

A computer representation of the Wollenzien model has been constructed (Nagano & Harel, 1986) but it does not attempt to analyze the model or adjust it to conform to new data. Instead the computer model is used to explore the relationships of the other pieces of the protein synthesis machinery. Therefore a comparison with this model would merely recapitulate the analysis of the Wollenzien model. An independent model of 16S RNA has been constructed using a computer modeling approach (Stern et al., 1988). Noller and coworkers have used the computer as a replacement for the wires, tubes, and spheres of the physical models. The actual modeling still relies on the researcher to issue the commands, spin the knobs, and supply the logic for every manipulation. Both of these computer models use display schemes that may be adequate when viewed on a terminal in a darkened room but are fairly uninformative when reproduced as pictures.

Noller model

The protocol used to assemble this model resembles a sequential folding of the RNA. Starting from the well established helices of the 5' domain, computer graphic representations of helices are stacked or packed onto the growing body of the RNA structure. RNA linkage and helical stacking are used to dock the regular A-form helices. As in the Brimacombe model, the psoralen and GbzCynAc crosslinks were excluded from the input data set that included protein/RNA and RNA/RNA crosslinks, protein induced changes in protection of the RNA, and the three dimensional protein map derived from neutron scattering. This level of data is sufficient to restrict the placement of more than half of the 16S RNA helices. The positions of the purple-striped, green-striped, cyan-striped, red-striped, rose, and blue helices are ill-defined under this protocol and were added to the model based on little more than the intuition of the researchers. Only the solvent face of this model was published in a format which allowed it to be color coded and this view best

fits the solvent side display of last74 (Fig. 29). The 'best' of the previous sentence is an entirely relative term as serious conflicts between the Noller and last74 models exist and the Noller model can be compared almost as well to the solvent surface last70. As the key alignment feature the red helices are perforce superimposed. Of the other 5' domain helices only the yellow helix is close to the same position in both models. The purple and the blue-striped helices occupy the same area in both models although their positions are reversed with the purple helix being on the right waist of the Noller model and in the middle of last74. The green and orange helices are placed in the bottom of the Noller model. Although it might be possible to rotate the extended green helix of last74 down to match the Noller model, there are no simple changes that could bring the orange and the yellow-striped helices into alignment. In the 3' domain only the black-striped helices correspond in the two models though the cyan-striped helix could be made to fit with a simple folding. The orange-striped helix is on the upper right in the Noller model and on the opposite side in last74. Nothing short of a complete refolding of the model could bring the blue helices into agreement.

The byhand model shows a little better match with the Noller model. The comparison must be made to the solvent side of the byhand model that is not pictured in this document. Beyond the benchmark red helix, none of the 5' domain helices have the same positions in both models. A little imaginative rotation might bring the yellow and cyan helices in the same neighborhoods in the Noller and byhand models but the purple-striped and green-striped helices are in the bottom of the byhand model, far from their positions in the Noller model. The purple and yellow-striped helices of the middle domain have irreconcilably different positions in the two models. The orange helix is packed beneath the green and blue helices in the Noller model while it is on the outside in the byhand model even though it is very close to the same area of the structure. The green and blue-striped helices are very much closer to superposition with their counterparts in the two models. Only the cyan-striped and red-striped helices are seriously misplaced when the 3' domain

helices are compared, being on the right and center waist of the byhand model and forming the central and left top of the Noller model. The orange-striped helix is near the uppermost right in both models although their cylindrical orientations vary by ninety degrees. The black-striped and rose helices are well aligned and the blue helices are almost perfectly superimposed.

In summary, all the models agree on the positioning of the blue-striped and black-striped helices. All but one model agree on the placement of the red, green-striped, yellow-striped, cyan-striped, and rose helices. Although all the models save those generated by the distance geometry place the green and blue helices vertically in the bottom of the molecule, there are major differences over the exact position and order of these helices. There is no clear consensus on the conformation of the other helices. The alignment of last74 with the other models is superior to that of last70 in every case and this indicates that the last74 family of structures is the correct 16S RNA enantiomer. It appears that since the three domains of 16S can be folded fairly independently that it is possible to distribute the chirality problem throughout the molecule. The comparison with the models of other researchers reveals that these structures are diastereomers.

Both the physical and computer modeling approaches of other researchers suffer from chaotic effects. Slight changes in the early stages of the modeling process can produce dramatically different final conformations. For example the Noller and Brimacombe models differ on whether the 565-810 and 310-350 helices face the 50S ribosomal subunit or the solvent side of the assembled ribosome. In contrast the pseudohelical models seem to be fairly stable with respect to small changes in the data set. This probably results from the derivation of the predominant mass of distance constraints from the primary and secondary structure. This weights the data against the less certain long range relationships and should be resistant to poor data as was seen in the modeling of transfer RNA. To the extent that this principle holds true, it may be possible to include less precise crosslinks in future studies. It has been well established that some of these crosslinks are exclusively associated

with the inactive form of the 30S subunit (Ericson & Wollenzien, 1989). For this reason Brimacombe and Noller have chosen to exclude psoralen crosslinks from their data sets. The crosslinks have been included here as there do not seem to be any major energy or kinetic barriers between the various forms of 16S. Consequently there must be a smooth transitional path among these forms and they must have similar free energies. Including these crosslinks will insure compatibility of the model with the conformational variability of 16S. The resilient flexibility of the computer modeling protocol is also evident in its ability to incorporate new data. None of the other models are adjustable in this manner are therefore doomed to obsolescence as more data becomes available.

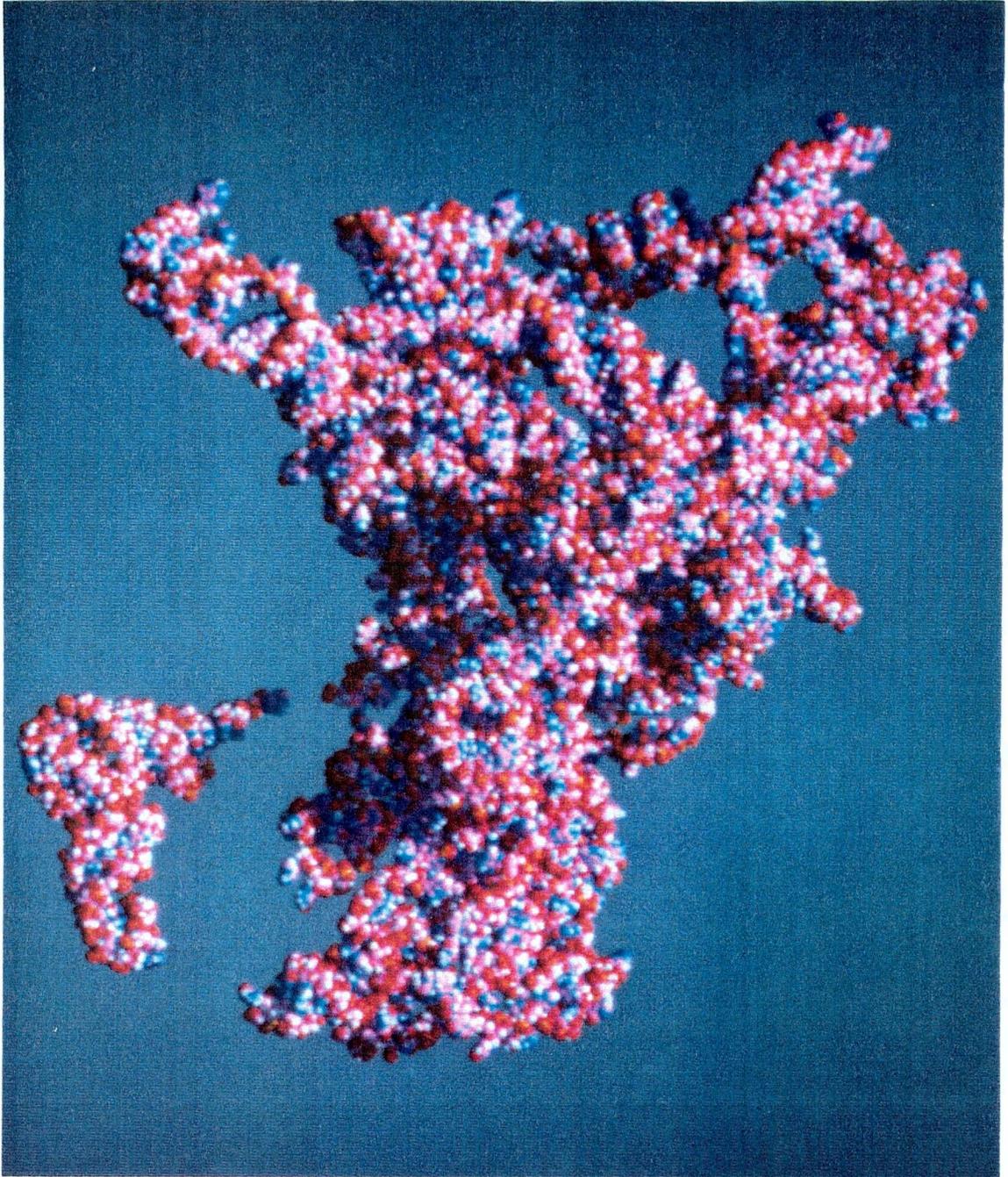
It is well established that the 16S undergoes further compaction upon addition of the ribosomal proteins S4, S7, S8, S15, S16, and S17 (Serdyuk, 1983). As only quantifiable 16S rRNA crosslinks have been used to fold the molecule, the resultant structures may reflect a stage in the assembly process. If the computer generated structures represent a partial folding of 16S RNA, then the additional folding of the RNA may be protein dependent. This would be consistent with the decrease in the hypochromicity of the fully folded state which indicates a decrease in RNA base stacking. It should also be noted that the computer models allow helices to become strained or less helical as required by the folding of the entire molecule, unlike the other models which rigidly maintain the integrity of all the phylogenetically indicated helices. If unstacking does occur, absolute adherence to the secondary structure map could be a source of error in the other models.

The fact that both the byhand model and the DSPACE models are able to satisfy the modeling constraints suggests that the point at which the structure of 16S ribosomal RNA can be uniquely determined has yet to be reached. Certainly the serious conflicts with the models of other researchers indicate that the computer generated results must be viewed with caution. But theoretical arguments (Crippen, 1987) and the success in modeling transfer RNA with this protocol suggest that the computer generated structures will resemble the global conformation of 16S ribosomal RNA. Considering the simple nature of

the modeling constructs, the selective structural data set, and the limited computing resources, the resultant structures show a high degree of correspondence with physical characteristics of 16S ribosomal RNA. The compatibility with new phylogenetic and crosslink data and the ability to quickly and seamlessly incorporate such new tertiary information is especially noteworthy. Even if some the details are wrong, the improvements in display, discrimination, and understanding which derive from this protocol are substantial. When the all atom CPK version of a 16S conformer is simultaneously displayed with a similar transfer RNA model (Fig. 32), the instinctive visual judgement that an individual has accumulated over a lifetime integrates all the abstractions used into concrete questions about just what will be involved in making the ribosome work.

On the following page:

Figure 32. An all atom representation of the crystal structure of Yeast Phenylalanine transfer RNA and a model of the solvent face of 16S ribosomal RNA. Spheres colored according to atom type; blue = nitrogen, red = oxygen, purple = carbon, white = hydrogen, and orange = phosphorous.



BBC 901 – 834

Figure 32

References

- Atmadja, J., Stiege, W., Zobawa, M., Greuer, B., Osswald, M., & Brimacombe, R. (1986) *Nucleic Acids Research* 14, 659-673.
- Beauclerk, A.A.D., & Cundliffe, E. (1987) *Journal of Molecular Biology* 193, 661-671.
- Bowman, C.M., Dahlberg, J.E., Ikemura, T., Konisky, J., & Nomura, M. (1971) *Proceedings of the National Academy Science* 68, 964-968.
- Brimacombe, R., Atmadja, J., Stiege, W., & Schueler, D. (1988) *Journal of Molecular Biology* 199, 115-136.
- Brosius, J., Palmer, M.L., Kennedy, P.J., & Noller, H.F. (1978) *Proceedings of the National Academy of Science* 77, 201-204.
- Capel, M.C., Kjeldgaard, M., Engelman, D.M., and Moore, P.B. (1988) *Journal of Molecular Biology* 200, 65-87.
- Crippen, G.M. (1987) *Journal of Physical Chemistry* 91, 6341-6343.
- Dahlberg, A.E. (1989) *Cell* 57, 525-529.
- Dams, E., Hendriks, L., Van de Peer, Y., Neefs, J.-M., Smits, G., Vandenbempt, I., and De Wacter, R. (1988) *Nucleic Acids Research* 16, r87-r173.
- Denman, R., Negre, D., Cunningham, P.R., Nurse, K., Colgan, J., Weitzmann, C. & Ofengand, J. (1989) *Biochemistry* 28, 1012-1019.
- Ericson, G. & Wollenzien, P. (1989) *Journal of Biological Chemistry* 264, 540-545.
- Expert-Bezancon, A. & Wollenzien, P.L. (1985) *Journal of Molecular Biology* 184, 53-66.
- Folkhard, W., Pilz, I., Kratky, O., Garrett, R. and Stoeffler, G. (1975) *European Journal of Biochemistry* 59, 63-71.
- Gutell, R.R., Weiser, B., Woese, C.R., & Noller, H.F. (1985) *Progress in Nucleic Acid Research and Molecular Biology* 32, 155-216.
- Haselman, T., Camp, D.G., & Fox, G.E. (1989) *Nucleic Acids Research* 17, 2215-2221.

- Higo, K., Held, W., Kahan, L., & Nomura, M. (1973) *Proceedings of the National Academy of Science* 70, 944-948.
- Hill, W.E., Thompson, J.D., & Anderegg, J.W. (1969) *Journal of Molecular Biology* 44, 89-102.
- Helser, T.L., Davies, J.E., & Dahlberg, J.E. (1972) *Nature New Biology* 235, 6-8.
- Kan, L.S., Chandrasegran, S., Pulford, S.M., & Miller, P.S. (1983) *Proceedings of the National Academy of Science* 80, 4263-65.
- Krzyzosiak, W., Denman, R., Nurse, K., Hellmann, W., Boublik, M., Gehrke, C.W., Agris, P.F., & Ofengand, J. (1987) *Biochemistry* 26, 2353-2364.
- Lake, J.A. (1985) *Annual Review of Biochemistry* 54, 507-530.
- Melancon, P. Lemieux, C., and Brakier-Gingras, L. (1988) *Nucleic Acids Research* 16, 9631-9639.
- Nagano, K. and Harel, M. (1986) *Progress in Biophysics and Molecular Biology* 48, 67-101.
- Noller, H.F., & Chaires, J.B. (1972) *Proceedings of the National Academy of Science* 69, 3115-3118.
- Noller, H.F., Stern, S., Moazed, D., Powers, T., Svensson, P., & Changchien, L.-M. (1987) *Cold Spring Harbor Symposia on Quantitative Biology* 52, 695-708.
- Noller, H.F., & Woese, C.R. (1981) *Science* 212, 403-411.
- Nomura, M. (1987) *Cold Spring Harbor Symposia on Quantitative Biology*, 52, 653-663.
- Prince, J.B., Taylor, B.A., Thurlow, D.L., Ofengand, J., & Zimmerman, R.A. (1982) *Proceedings of the National Academy of Science* 79, 5450-5454.
- Ramakrishnan, V. (1986) *Science* 231, 1562-1564.
- Santer, M. (1963) *Science* 141, 1049-1050.
- Serdyuk, I.N., Agalassov, S.C., Sedelnikova, S.E., Spirin, A.S., & May, R.P. (1983) *Journal of Molecular Biology* 169, 409-425.

- Shine, J., & Dalgarno, L. (1974) *Proceedings of the National Academy of Science* 71, 1342-1346.
- Spitnik-Elson, P., Elson, D., Avital, S., & Abramowitz, R. (1985) *Nucleic Acids Research* 13, 4719-4738.
- Stiege, W., Atmadja, J., Zobawa, M., & Brimacombe, R. (1986) *Journal of Molecular Biology* 191, 135-138.
- Stiege, w., Kosack, M., Stade, K., & Brimacombe, R. (1988) *Nucleic Acids Research* 16, 4315-4329.
- Stern, S., Weiser, B., & Noller, H.F. (1988) *Journal of Molecular Biology* 204, 447-481.
- Thompson, J.F. & Hearst, J.E. (1983) *Cell* 32, 1355-1365.
- Vasiliev, V.D., Selivanova, O.M. and Koteliansky, V.E. (1978) *FEBS Letters* 95, 273-276.
- Vasiliev, V.D., Serdyuk, I.N., Gudkov, A.T. and Spirin, A.S. (1986) In *Structure, Function and Genetics of Ribosomes* (Hardesty, B. & Kramer, G., eds.), pp 128-142, Springer-Verlag, New York.
- Verschoor, A., Frank, J., Radermacher, M., Wagenknecht, T. and Boublik, M. (1984) *Journal of Molecular Biology* 178, 677-695.
- Weiss, R.B. Dunn, D.M., Atkins, J.F., and Gexteland, R.F. (1987) *Cold Spring Harbor Symposia on Quantitative Biology* 52, 687-693.
- Woese, C.R., & Fox, G.E. (1977) *Proceedings of the National Academy of Science* 74, 5088-5090.
- Wollenzien, P.L., Expert-Bezancon, A., Murphy, R.F., Cantor, C.R. & Hayes, D.H. (1985) *Journal of Molecular Biology* 184, 67-80.
- Yonath, A., Glotz, C., Gewitz, H.S., Bartels, K.S., von Boehlen, K., Makowski, I., & Wittmann, H.G. (1989) *Journal of Molecular Biology* 203, 831-834.

Chapter 6

30S RIBOSOMAL SUBUNIT MODELING

Introduction

In the previous chapter the results of the computer modeling protocol were described and proved to be compatible with the physical characteristics of the small subunit of the ribosome. The comparison with the models of 16S ribosomal RNA constructed by other means was less informative. Some ribosomal data, most prominently RNA/protein relationships, were excluded from the distance geometry parameter set as insufficiently quantitative. Now that the first phase of the modeling is complete, it is time to evaluate how compatible the computer generated structures are with other aspects of ribosomal structure.

Restricting the input data to the primary, secondary, and tertiary structure of 16S RNA was consistent within the context of an RNA modeling protocol. But as 16S RNA is part of a ribonucleotide/protein particle, there exists a substantial amount of information about its relationship to the particle and its protein parts. At a resolution of ~20 angstroms, electron microscopy has revealed the general overall shape of the ribosome and its subunits (Lake, 1985). Isolated ribosomal components can be used as antigens for the production of monoclonal antibodies. The ribosomal proteins can then be generally located within the small subunit silhouette by determining the attachment site of the immunoglobulins. Antibodies to the naturally occurring modified bases of 16S RNA and the haptenized 5' and 3' ends of the molecule can be used to generally position the RNA within the 30S subunit (Woese et al., 1983). Of course these studies are all limited by the resolution of the electron microscope. The sample preparation required by this technique is also fairly harsh and may adversely affect the structure of a highly hydrated complex like the ribosome. Thus the maps constructed from electron micrographs must be considered qualitative in nature.

Several more quantitative techniques have been used to probe the quaternary structure of the small subunit. Protein/protein crosslinking and fluorescent energy transfer between pairs of labeled proteins were used to establish an initial protein map of the 30S subunit (Hardesty et al., 1986). Since the shapes of the proteins are not known, both of these techniques have very substantial levels of uncertainty. Reconstitution of the ribosome

with pairs of deuterated proteins allows the separation of the centers of mass of the protein pairs to be determined with low angle neutron scattering. Fifteen years of work have produced a set of distances that can be used to construct a complete mapping of the ribosomal proteins (Capel et al., 1987). These neutron scattering experiments also produce good estimates of the radius of gyration of the individual proteins, although the details of their structures remain a mystery. Protection of 16S ribosomal RNA from chemical and enzymatic probes when it is part of the small subunit or bound to a subset of the 21 small subunit proteins, has revealed a set of RNA/protein relationships (Noller et al., 1987). Some of these interactions correspond with the relationships deduced from reconstitution of the ribosome from its parts. Finally, RNA/protein crosslinks have been developed which provide a firm linkage between the RNA structure and the protein map (Brimacombe et al., 1988b). The identity of the crosslinked nucleotides can be determined but very little can be said about the protein attachment site. This severely undermines the ability of this data to uniquely locate ribosomal structure elements.

With the RNA/protein crosslinks as guides, the computer generated models of 16S RNA can be aligned with the proteins of the 30S ribosomal particle to produce a model of the small subunit. The protein map and other data concerning the position of 16S RNA within the small subunit are used as an independent check on the quality of the structures produced by the new modeling technique. This final evaluation indicates that the protocol has successfully met the initial research goals. Improvements suggested by this pioneering work should greatly increase the facility with which RNA structures are built and understood.

Materials and Methods

Materials

Computer Resources

All the calculations were performed on a MicroVAX II workstation. The raster pictures were displayed on an Evans & Sutherland PS340 hosted by a VAX 11/785. The Protein Data Bank formatted coordinate data files were transformed into graphics with the raster modeling programs written by Michael Connolly and obtained from the Scripps Research Institute in San Diego, California.

Structural Data

RNA Structures

The computer generated structures used, last70 and last74, and the composite physical/computer model, byhand, were produced during the development of the computer modeling protocol. They were left in the orientations based on alignment with the Immune Electron Microscopy map of the small subunit (Lake, 1985).

Protein Map

The ability to disassemble and then reconstitute the ribosome has allowed researchers to construct ribosomal subparticles with specific differences. *E. coli* will grow in deuterium saturated media and produce ribosomal proteins that are approximately 85% deuterated. The labeled proteins can then be recovered and purified in sufficient quantity for reconstitution experiments. These proteins have a low angle neutron scattering profiles that differ significantly from unlabeled proteins. Comparison of the scattering curves from unlabeled 30S subunits, subunits with only protein A labeled, subunits with only protein B labeled, and subunits with both A and B labeled will yield an interference scattering curve for the two proteins. From this a quantitative estimate of the distance between their centers can be determined. When enough interprotein distances have been accumulated, a complete mapping of the small subunit can be constructed. If the proteins are treated as spheres, a simple geometric approach can be used to produce a self-consistent three dimensional

arrangement which satisfies the distance data (Peter Moore, personal communication). The first protein is assigned to be the origin of reference for subsequent placements. Adding the next two proteins requires all three interprotein vectors to form the appropriate triangle. The fourth protein to be added to the map will require three distances to the previously placed proteins to insure proper chirality. Each subsequent protein will require at most four interprotein distances to be uniquely located in the growing structure. With this scheme, it can take as few as 78 distances to place the 21 proteins of the 30S subunit. Peter Moore and coworkers have completed a mapping of the small subunit in this fashion (Capel et al., 1988) and the coordinates of the proteins and their radii of gyration will be used for comparison with the computer generated models.

RNA/protein crosslinks

A total of 29 crosslinks between 16S RNA and the proteins of the 30S subunit have been produced with a variety of crosslinking reagents (Fig. 1). The resolution of the RNA crosslink site varies from one crosslink to the next. As the structures of the individual proteins are unknown and the crosslinked protein residues are not identified, the protein end of the crosslinks can only be specified in a general way.

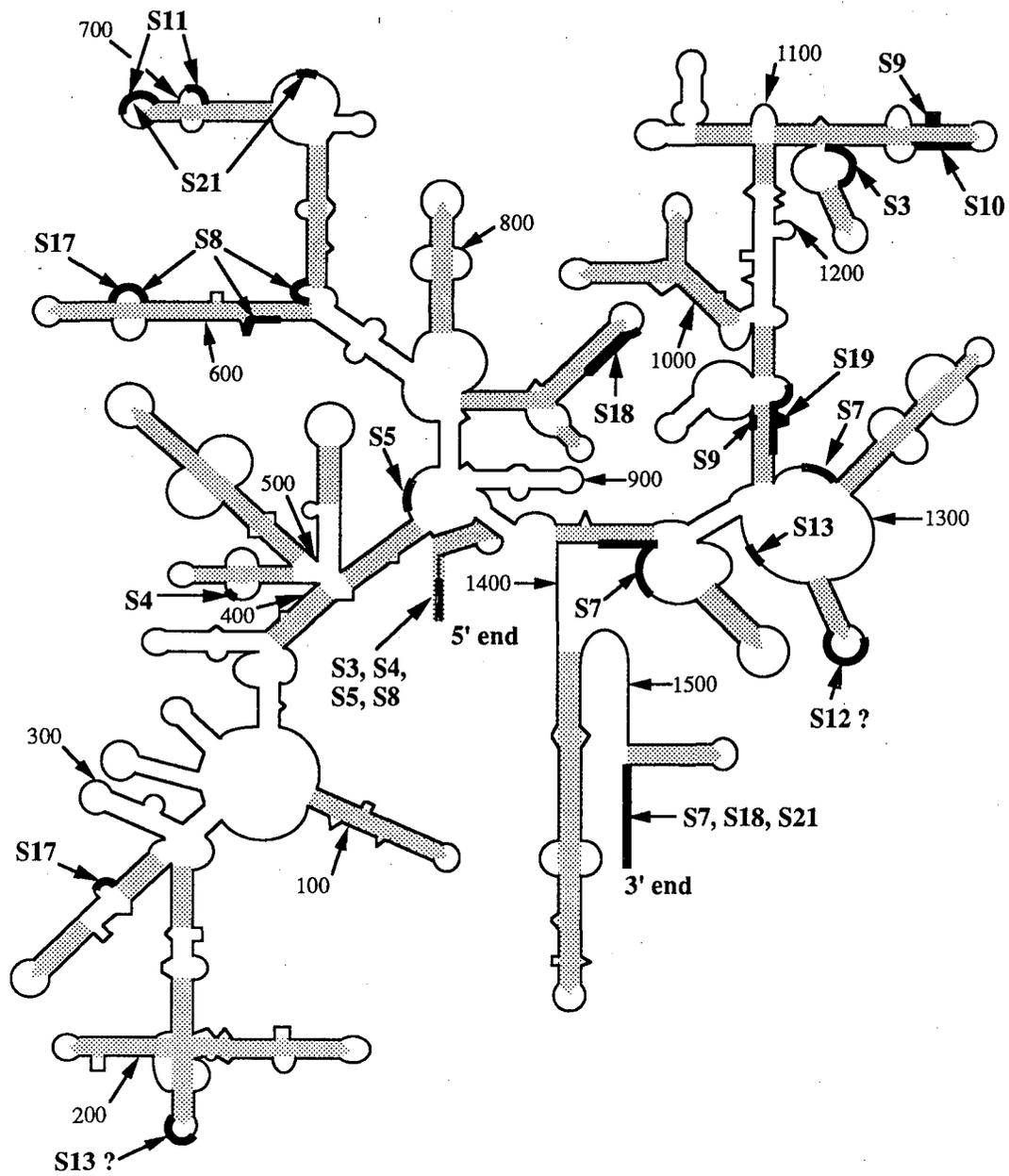


Figure 1. The positions of RNA/protein crosslinks mapped onto the 16S RNA secondary structure.

2-Iminothiolane crosslinks

2-Iminothiolane is a heterocyclic, bifunctional reagent which will crosslink RNA and proteins when exposed to ultraviolet irradiation. Eight crosslinks between 16S RNA and the small subunit proteins have been identified. The crosslinks S7 X 1238-1240, S7 X 1377-1388, S8 X 629-633, S8 X 651-654, and S8 X 593-597 were the first to be determined (Wower & Brimacombe, 1983). The three crosslinks S17 X 629-633, S21 X 723-724, and S21 X 1531-1542 were identified in a later study (Kyriatsoulis et al., 1986).

Nitrogen mustard crosslinks

Bis-(2-chloroethyl)-methylamine is such a reactive molecule that it will crosslink almost anything. RNA/RNA crosslinks made with this chemical were used earlier in this work to produce the 16S RNA model structures. Nine RNA/protein crosslinks, S4 X G413, S3 X 1155-1158, S7 X 1531-1542, S9 X G954, S10 X 1139-1144, S11 X 693-697, S17 X 278-280, S18 X 845-851 and S21 X 693-697, have been isolated (Greuer et al., 1987). A new technique was used in this study which made it possible to absolutely identify the S4 and S9 crosslinks. It is unfortunate that the protein side of the crosslink is so nebulous, as the short range (~5 angstroms) of this linker could place a severe limit on the number of possible alignments of 16S RNA and the proteins.

P-azidophenyl acetimidate crosslinks

This reagent is an imido ester which first bonds to a lysine residue of a protein. Subsequent irradiation photoactivates the azide group which reacts with the RNA to produce a crosslinks. Ten RNA/protein interactions, S3 X 1-4, S4 X 1-4, S5 X 1-4, S5 X 559-561, S8 X 1-4, S7 X 1238-1240, S9 X 1130-1131, S11 X 702-705, S13 X 1337-1338 and S19 X 1223-1231, have been identified (Osswald et al., 1987). Immune electron microscopy and RNA protection experiments indicate that the 5' domain of 16S RNA which forms the base of the small subunit is not bound by protein. Thus the S13 X 189-191 crosslink which was also seen raises some questions about the reliability of any RNA/protein crosslink.

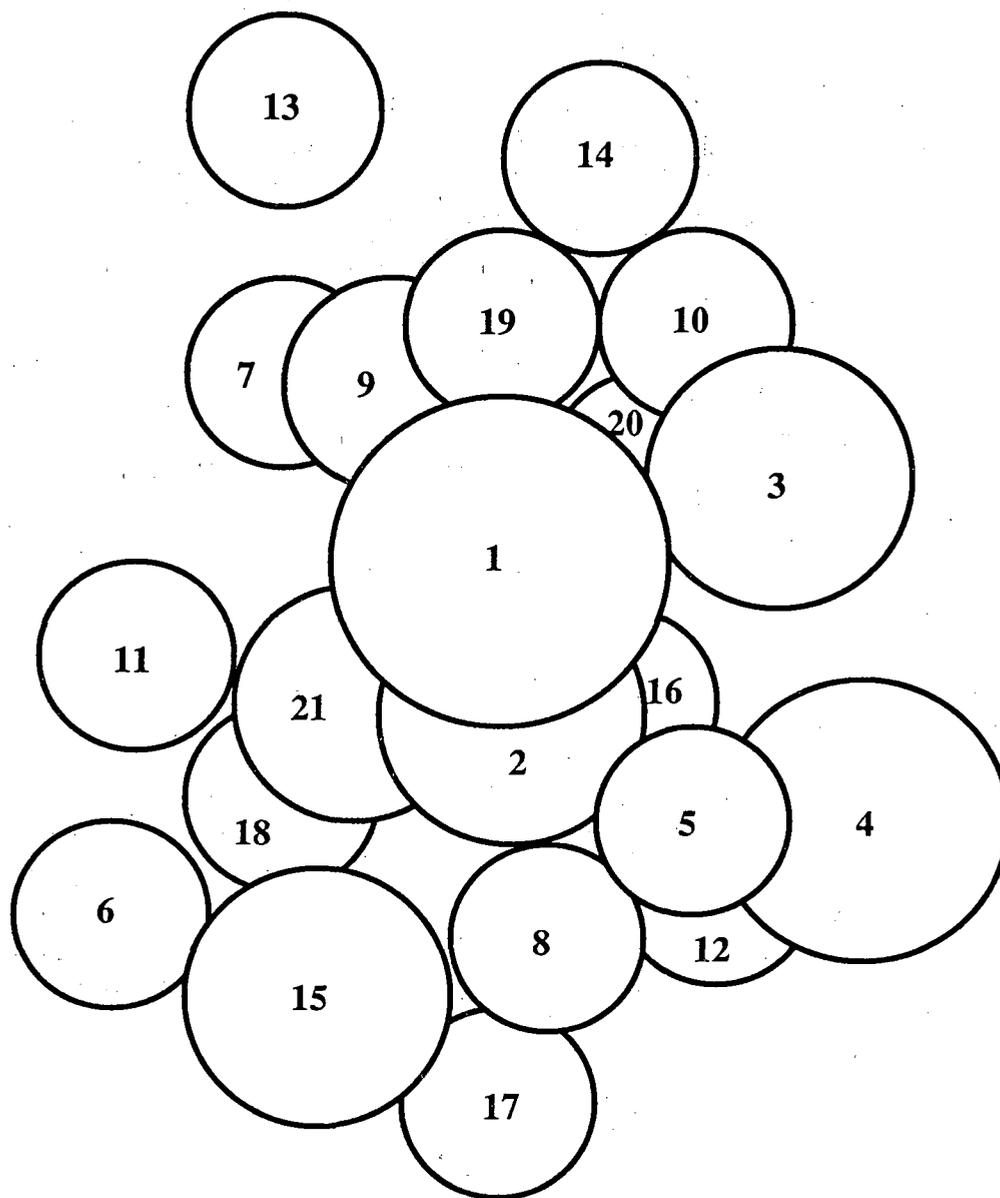
1-Ethyl-3-dimethyl-aminopolycarbodiimide

The crosslink between S12 and C1322 of the 1316-1322 segment of 16S RNA produced by this reagent, is included in the interest of completeness even though subsequent experiments by the same research group have been unable to repeat the result (Chiaruttini et al., 1982).

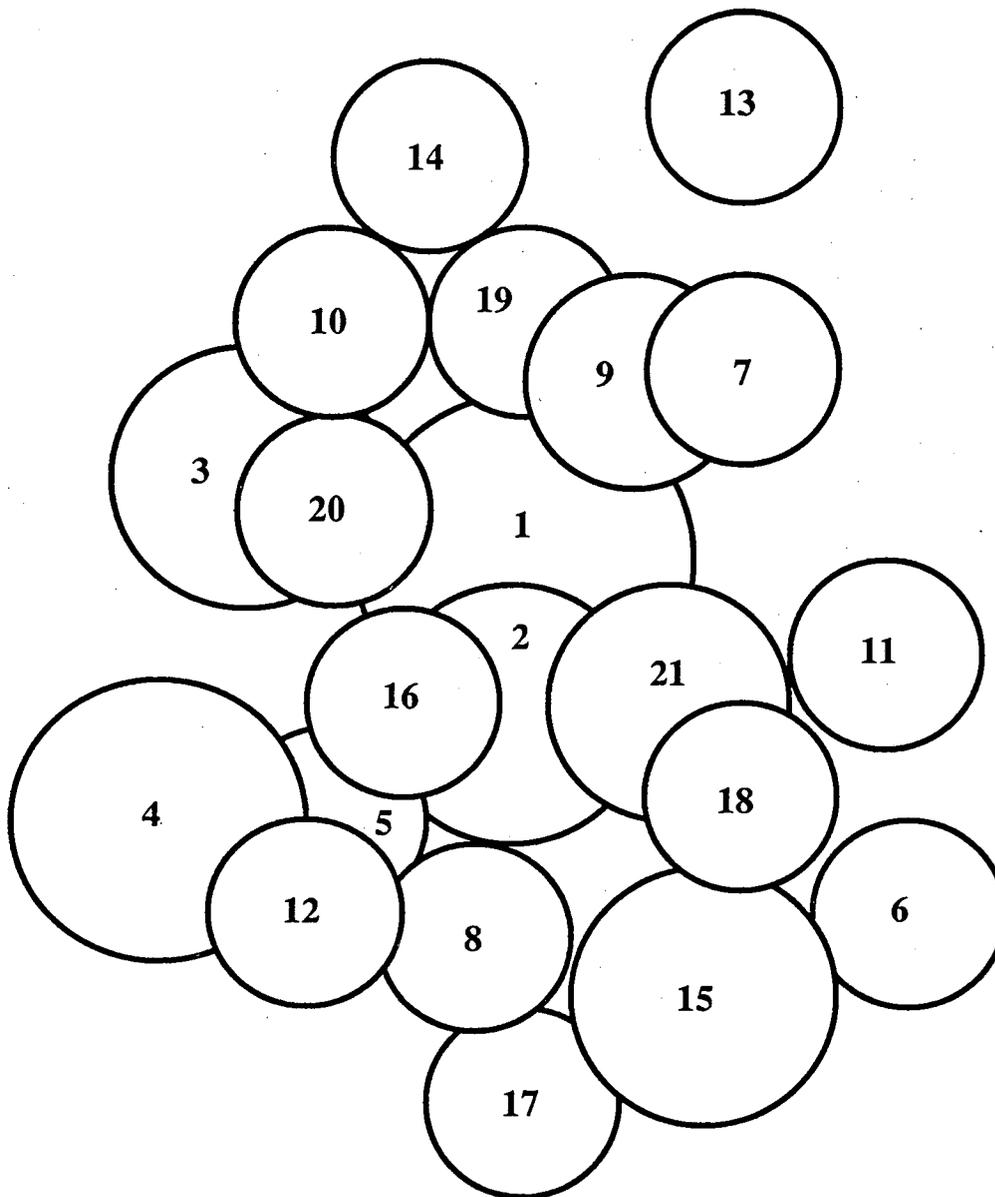
Methods

RNA/protein superposition

As neither the crosslinking data nor the RNA models have a high degree of precision, no special procedure was used to superimpose the protein map and the RNA structures. The protein map coordinates were first translated so that the centroid of the map became the origin. The protein map was then rotated so that the views identified as the solvent-facing side (Fig. 2) and the 50S subunit interface (Fig. 3) were parallel to the xy plane. The RNA models were previously transformed so that their centers of mass coincided with the origin and the putative solvent and 50S subunit sides paralleled the viewing plane. Protein center to center of crosslinked RNA segment vectors were calculated using locally written FORTRAN software. The variation in the length of these vectors was monitored as the 16S RNA model was rotated in 30 degree increments about the y-axis while the proteins were left in their starting orientation.



**Figure 2. Small Subunit Proteins
Solvent Face
(proteins spheres not to scale)**



**Figure 3. Small Subunit Proteins
50S subunit interface
(protein spheres not to scale)**

Raster graphics

The Molecular Cross Section (MCS) program was used to produce the raster displays of the RNA/protein superpositions. The RNA models are displayed as a tube of one angstrom radius which approximates a tracing of the phosphate backbone. Proteins are shown as spheres with radii equal to their radius of gyration as determined by neutron diffraction. It was necessary to make some of the proteins transparent because of their large size. This was especially true for S1 which almost completely obscures the solvent side of the map. The protein coloring scheme was derived from the sections of the 16S RNA that appear to be influenced by a respective protein (Noller et al., 1987). The color pictures are 35mm photographs of the results as displayed on the raster attachment of an Evans & Sutherland PS340 graphics workstation.

Protein	Color	Rg(angstroms)
S1	LIGHT GREEN	55.4
S2	TAN	32.8
S3	ORANGE	25.4
S4	CYAN	30.2
S5	RED	13.5
S6	GREEN	13.1
S7	ROSE	13.7
S8	GREEN	21.1
S9	BEIGE	20.8
S10	ORANGE	12.6
S11	PURPLE	12.5
S12	RED	12.6
S13	ROSE	12.5
S14	TAN	12.0
S15	BLUE	26.8
S16	CYAN	25.7
S17	YELLOW	13.2
S18	YELLOW	10.9
S19	BEIGE	11.6
S20	LIGHT GREEN	11.3
S21	BLUE	24.8

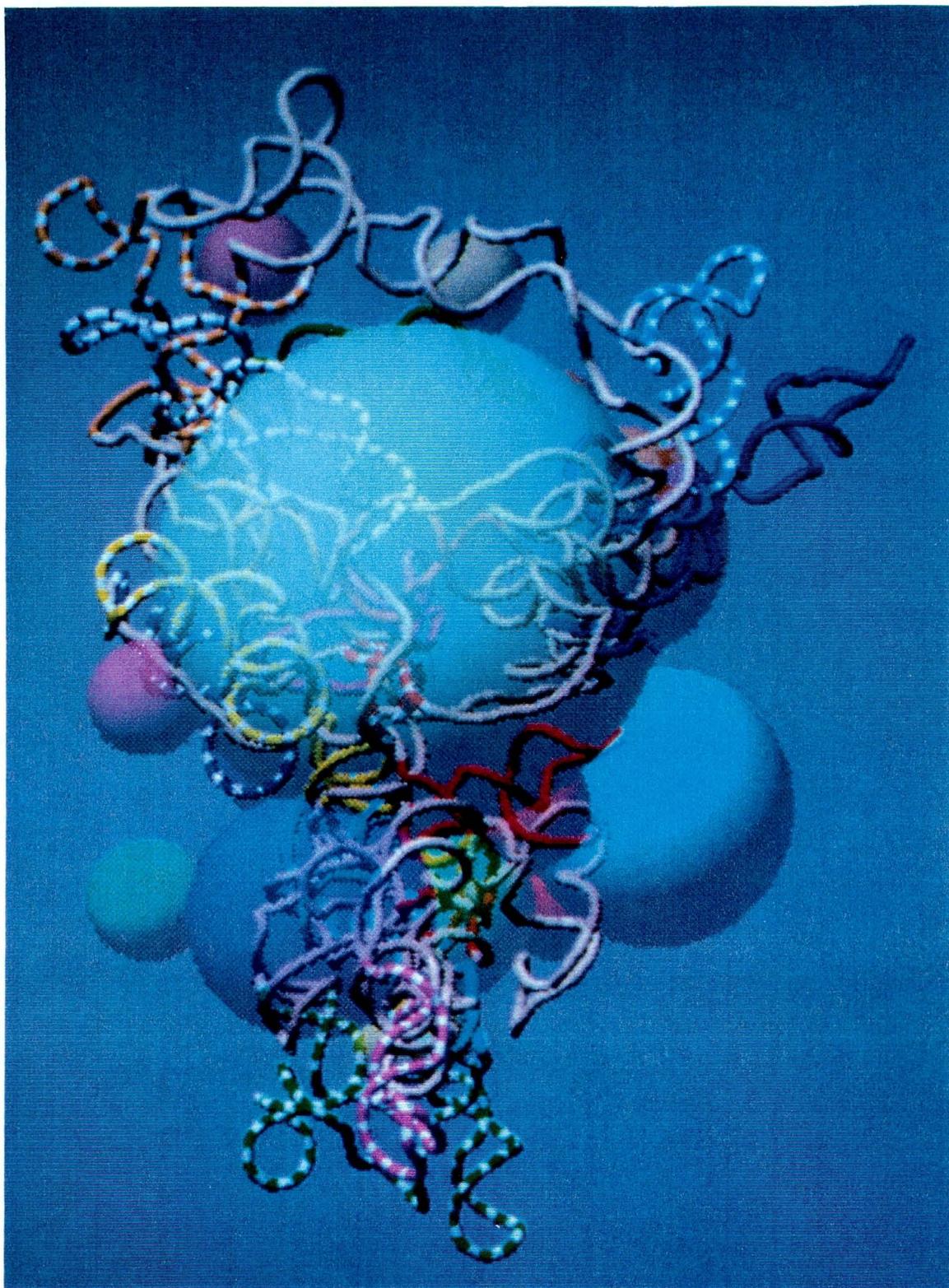
Results

30S Ribosomal Subunit Models

The initial orientations of the 16S RNA models were based on a straightforward manual alignment with the map of the 30S subunit of the ribosome produced from Immunoglobulin Electron Microscopy. This simple docking procedure was good enough to allow the computer generated models to be compared to the models of other researchers. These orientations also proved to be fair alignments of the 16S RNA structures with the protein map developed from neutron scattering studies. When the small subunit proteins are simultaneously displayed with the computer generated models, it is apparent that the RNA models and the protein map have similar dimensions and volumes. The solvent face of the last70/protein composite is dominated by the light green sphere representing the S1 protein (Fig. 4). The proteins have been made semitransparent so that those portions 16S RNA which intersect with proteins can still be seen. Proteins S6 (green sphere on the left) and S4 (cyan sphere on the right) form a broad waist that leaves them some distance from the RNA backbone. Moving the RNA vertically down, or equivalently, moving the proteins up,

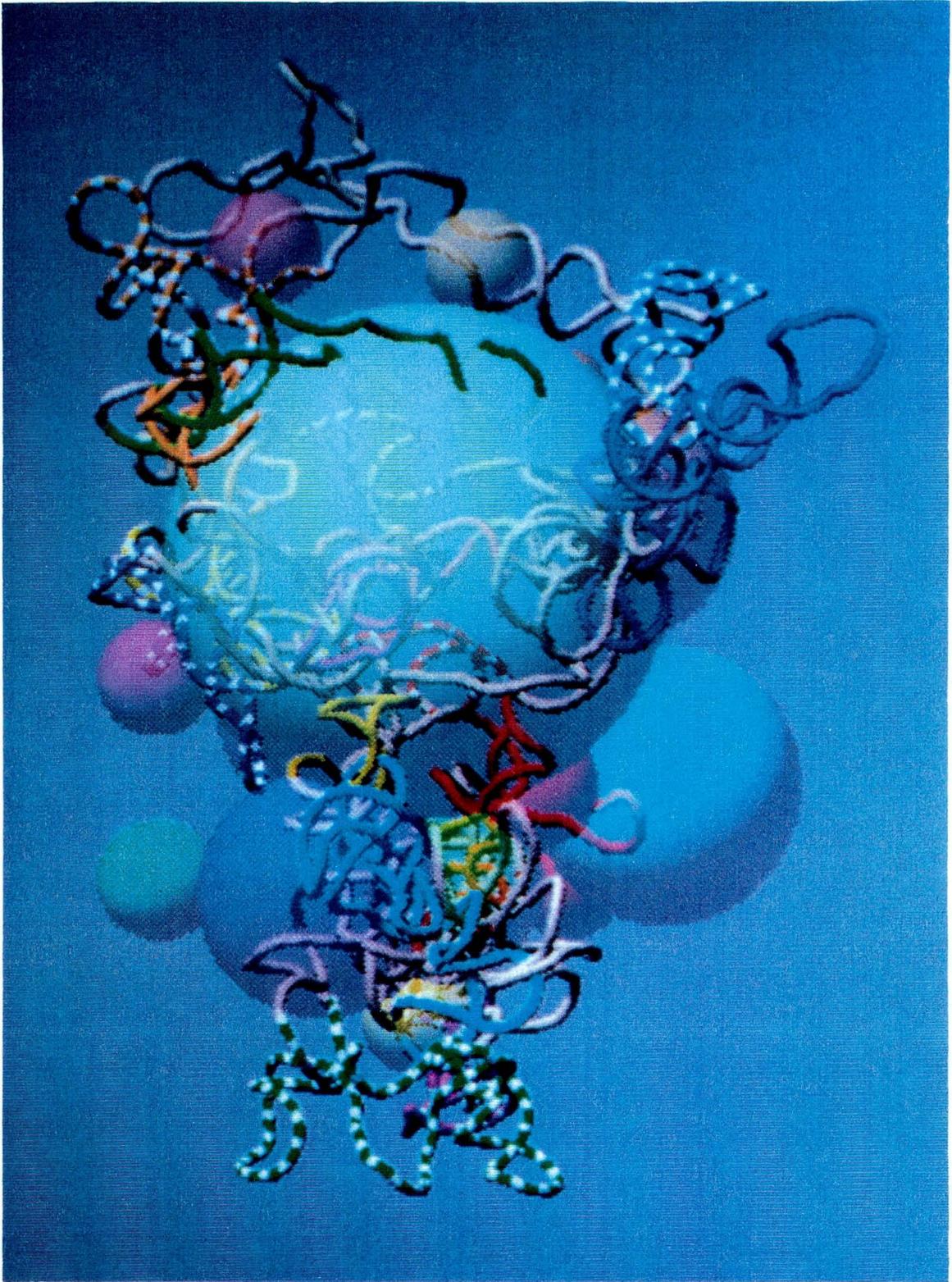
On the following page:

Figure 4. The 30S subunit proteins superimposed on the solvent face of last70.



BBC 901 - 832

Figure 4



BBC 901 – 818

Figure 5

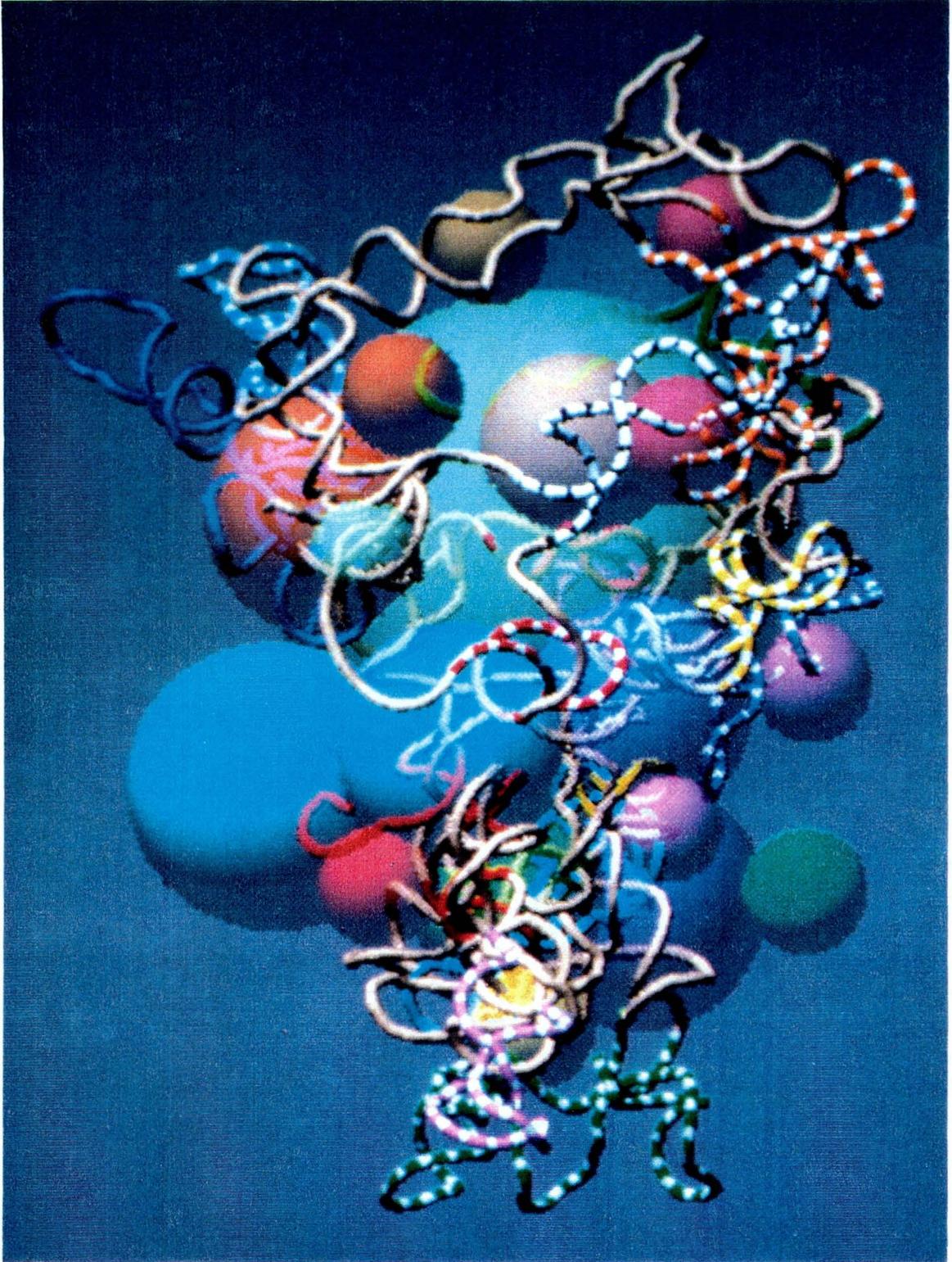
On the preceding page:

Figure 5. The small subunit proteins superimposed on the solvent face of last74.

would improve the fit of the 16S RNA model and the protein map. The solvent face of the last74/protein composite has similar features (Fig. 5). The fit of last74 to the protein map could also be improved by a vertical displacement of the RNA and protein centers of mass. This composite has several provocative features that the last70 30S subunit model lacks. The closer alignment of the yellow helix (247-277) with the yellow sphere (S17) near the bottom of the structure and the red helices (5' prime end) with the red sphere (S5) on the right side of the waist indicate that the RNA and protein models have similarities at a deeper level. The manner in which the protruding green and blue helices reach out through and around the light green S1 protein is highly suggestive. This display also illustrates that a simple pivoting rotation of these two helices will bring them into close proximity with the proteins S6 (green) and S15 (blue) at the bottom of the map. The view of the 50S subunit interface of the last74 30S subunit model reveals a complex and highly interdigitated superposition of RNA and protein (Fig. 6). This side of the complex is more informative because the overbearing S1 protein has been rotated to the rear.

On the following page:

Figure 6. The 50S interface of the 30S subunit formed from the small subunit proteins and last 74.



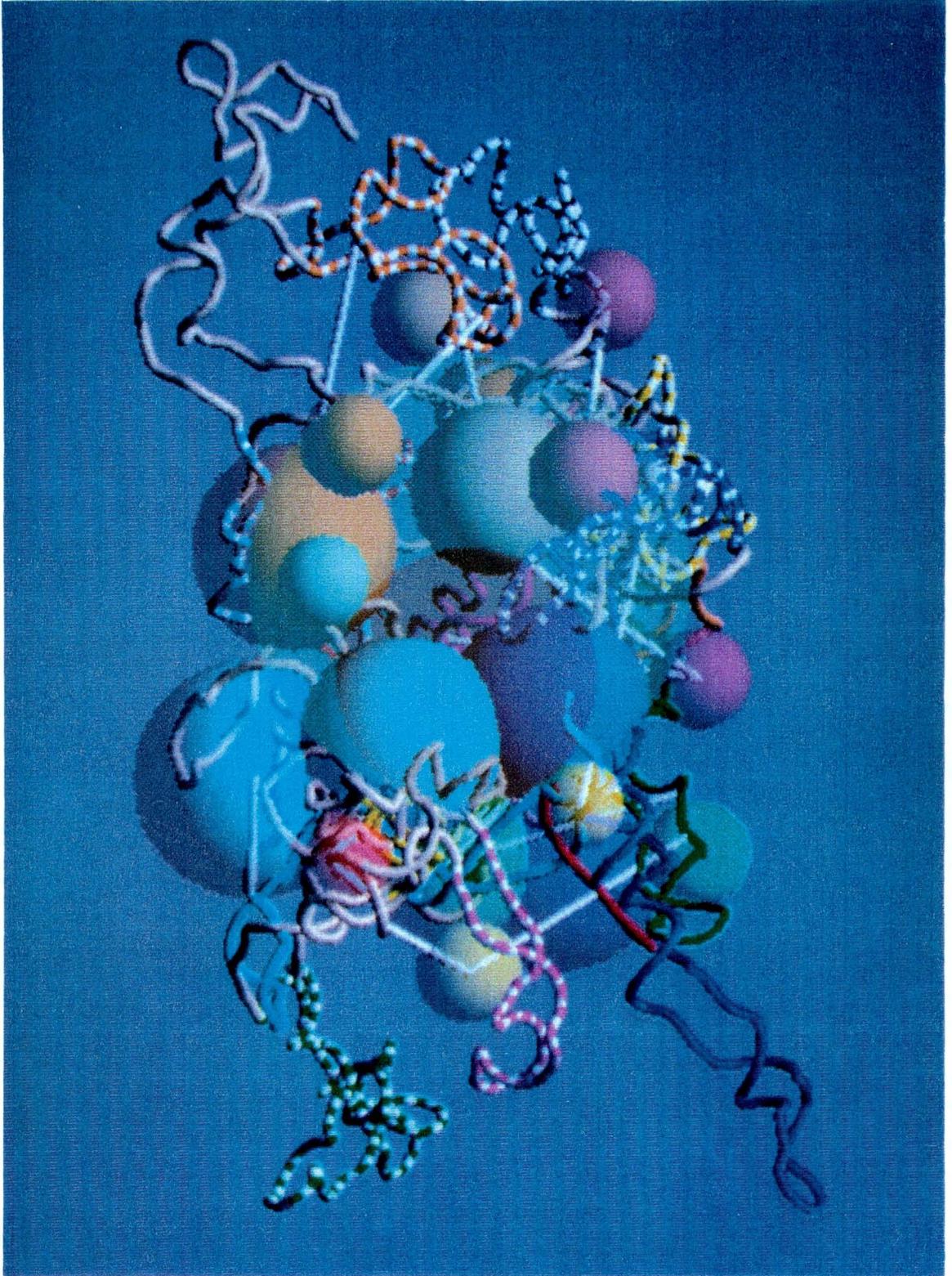
BBC 901 - 824

Figure 6

The 50S subunit interface of the byhand/protein composite shows a better match in horizontal dimension but a poorer match in vertical dimension than either last70 or last74 (Fig. 7). The superior alignment of the cyan, green, and blue helices with their matching proteins spheres may indicate that these relationships were part of the data set which was used to construct the byhand model. The green and blue correlations demonstrate that trivial pivoting rotations of these helices in the 16S RNA models, last70 or last74, might improve their fit to the protein map. As in the other composites, the byhand 30S subunit model could be improved by translating the proteins vertically upward. This adjustment has a physical correspondence with the structure of the 30S ribosomal particle. Determination of the centers of mass of the RNA and of the protein components indicates that they are separated by approximately 25 angstroms (Ramakrishnan, 1986). That the proteins should be higher in the vertical direction is confirmed by experiments which show the bottom third of the subunit to be free of protein (Lake, 1985).

On the following page:

Figure 7. The 50S interface of the small subunit model formed by the superposition of the 30S proteins and the byhand model.



BBC 901 - 840

Figure 7

In order to examine the correspondence between the 16S RNA models and the protein map, the vectors connecting the center of the protein to the adducted RNA residue for the known RNA/protein crosslinks were monitored as the RNA model was rotated about the vertical axis relative to the protein map. Considering the large uncertainties which exist in the crosslink data, this unsophisticated analysis should be as revealing as a more complex approach. The suspicious crosslink between S13 and the 5' domain has been omitted because no one, not even the researchers who isolated it, believe that it is a legitimate part of the 30S subunit structure (Osswald et al., 1987).

When the crosslink vector lengths for the last70 30S subunit model are plotted it is clear that no orientation is preferable to the original display. A simplified but representative graph of the results shows that, as expected, the three S8 and one of the S17 crosslinks to the unrestrained green helix are very bad and establish an upper limit on poorness of fit (Fig. 8). The fit of the S5, S8, S9, and S17 crosslinks is relatively independent of the last70 rotation because these proteins are near the vertical axis. The S4 crosslink vector shows a similar invariance because the crosslinked nucleotide is near the y-axis. Proteins S3, S7, S10, and S11 should provide the best discrimination as they are widely distributed. Of these four proteins only the S11 crosslinks are minimized in an alignment other than the original. The purple helix (687-713) that is crosslinked to S11 may be adversely affected by its close association with the protrudant green helix and further folding in this region should improve the value for this interaction.

The analysis of the last74 small subunit model produces crosslink displacement magnitudes that are similar to those seen for the last70 30S particle (Fig. 9). But the variation in the fit of the four key proteins is very disturbing. S3 best fits the last74 30S subunit when the 16S RNA is rotated by 180 degrees. S7 matches a last74 30S subunit that

Figure 8. RNA to Protein Center Distances for last70

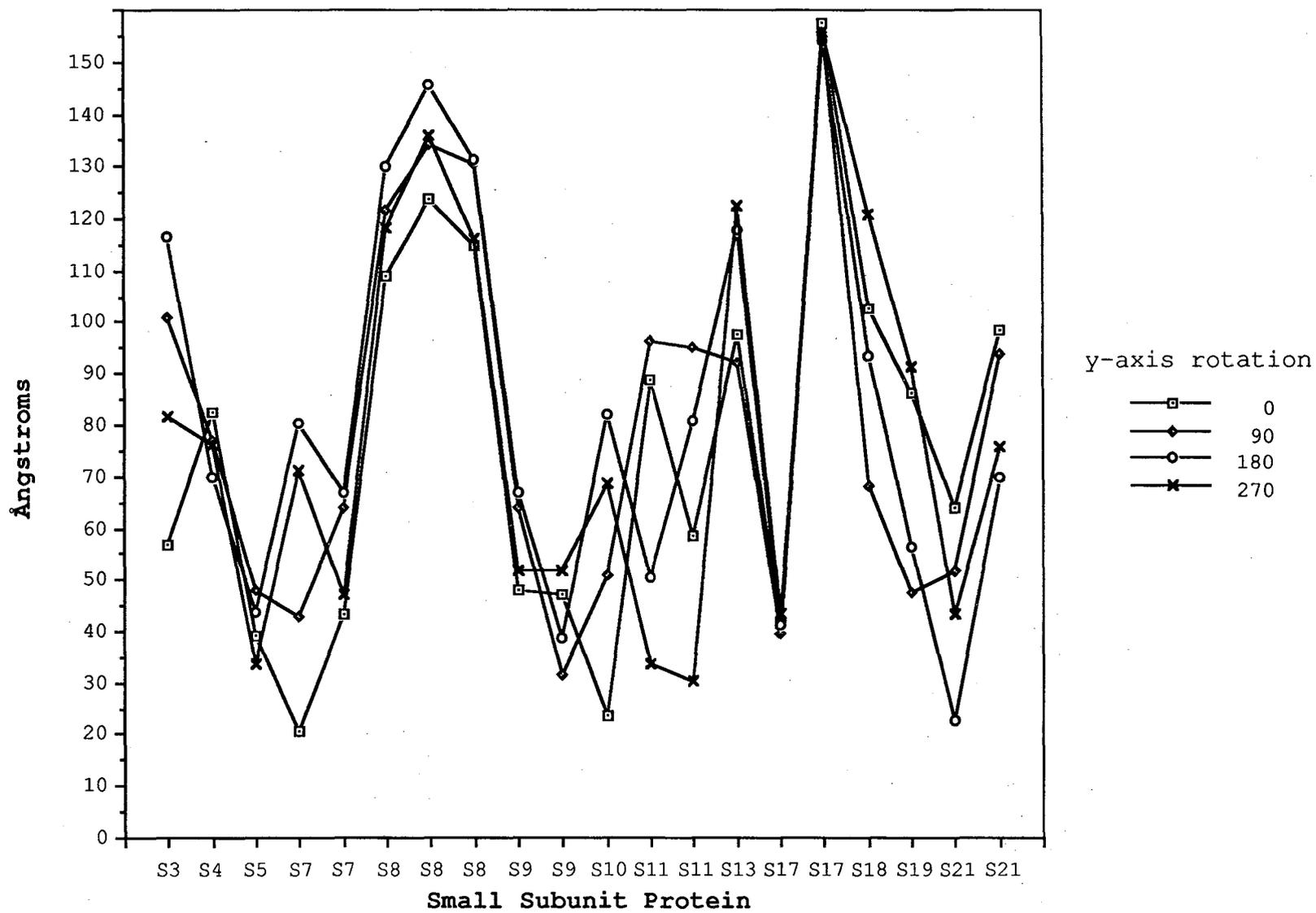
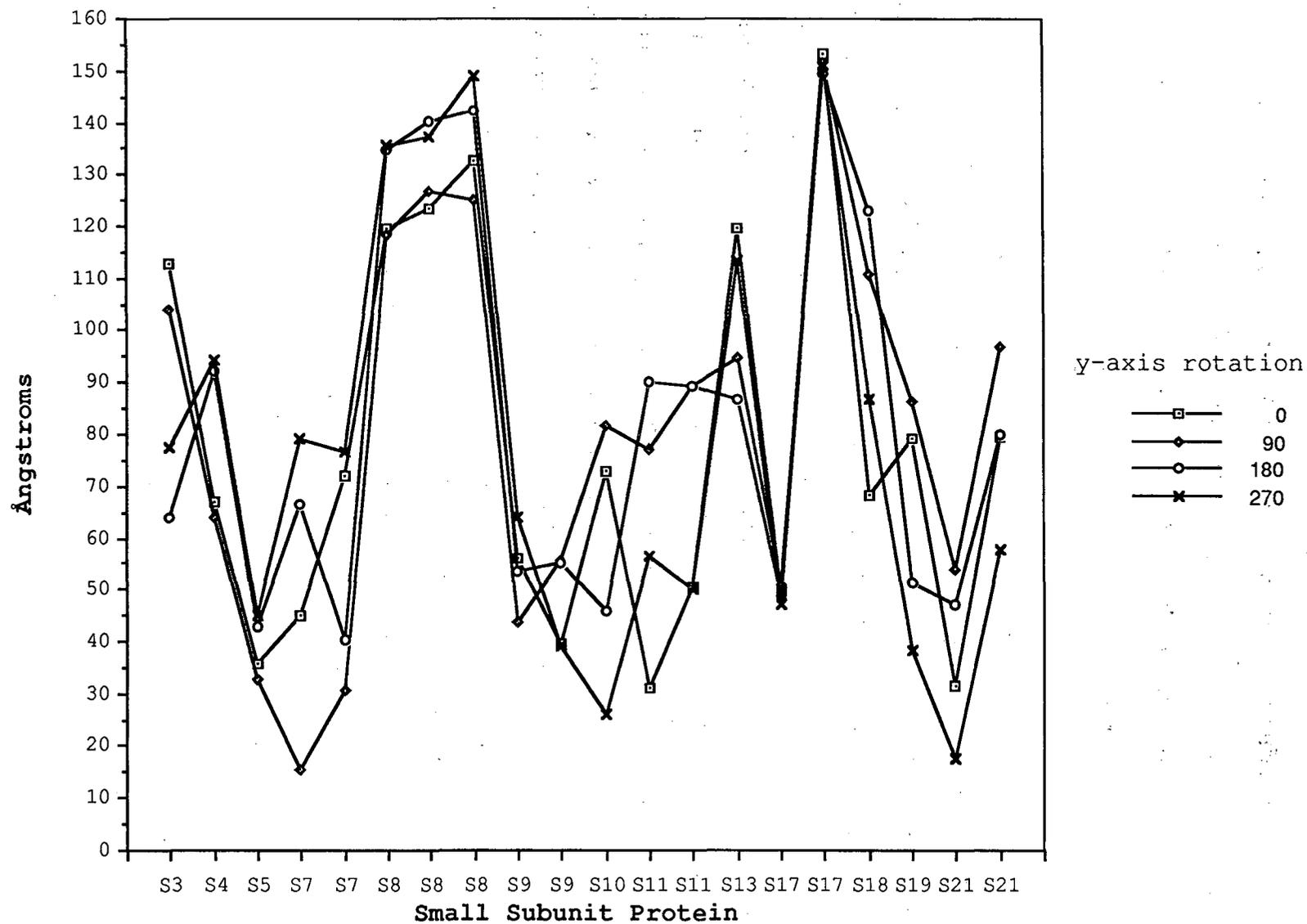


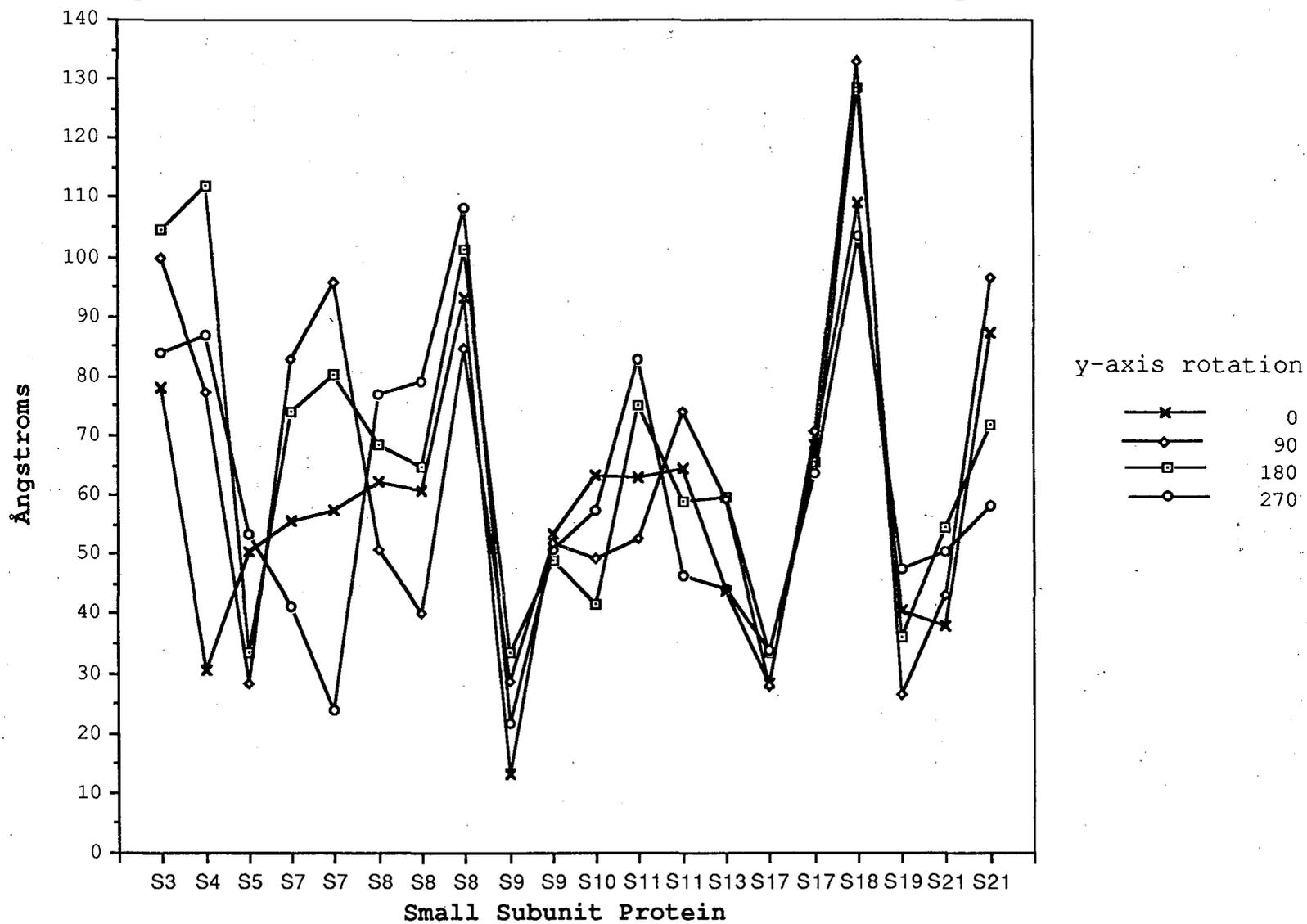
Figure 9. RNA to Protein Center Distances for last74



has the 16S RNA rotated by 90 degrees. The S10 vector is minimized with an RNA rotation of 270 degrees. And the S11 crosslink is less troublesome when the last74 30S subunit is left in the original orientation. These results resurrect the unpleasant chirality problem that seemed to be solved in favor of the last74 16S RNA model in the previous chapter. Of course chirality concerns also exist for the IEM and neutron scattering protein maps.

The byhand/protein superposition data reveals a correlation between this model of 16S RNA and the protein map that is similar to those of last70 and last74. The graph of the magnitudes of the crosslink vectors appears on first glance to be better because of the improvements in the S8 and S17 crosslinks (Fig. 10). But on the whole the values for byhand are very close to those of the other models. The spread between the best and worst values for a particular crosslink is generally smaller and the original orientation shows an acceptable fit for all the proteins except S8 and S18. Even with the green helix folded into the lower third of the structure, the three S8/RNA crosslinks indicate that a further compaction of this region of 16S RNA would be helpful. The crosslink vector between S18 (the uppermost yellow sphere) and the yellow-striped helix (821-879) is very poor and passes through a portion of the byhand 16S RNA model that was compressed during the conjugate gradient refinement of the 16S RNA structure. As the yellow-striped region on the right side is almost directly above S18, the translation of the protein map some twenty-five angstroms vertically will reduce this crosslink vector by a similar amount.

Figure 10. RNA to Protein Center Distances for byhand



30S Subunit RNA/Protein Correlations

The numerical evaluation of the RNA/protein crosslinks alone cannot reveal the structural implications that are the primary motive for model building and so some specific interactions will be used to examine the computer models of the 30S subunit. As none of the RNA/proteins superpositions would be dramatically improved by rotating the 16S RNA, the 30S subunit models have been left in their original orientations.

In the raster display of the byhand 30S subunit, the large S1 and S2 proteins have been made almost totally transparent (Fig. 7). This allows the interior of the structure to be seen and makes it possible to include the crosslinks between 16S RNA and the proteins of the small subunit. The vectors connecting the centers of proteins to the crosslinked RNA bases are shown as white tubes of one angstrom diameter. These vectors were not included for the last74 and last70 30S subunits because the links to the green and blue helices dominated and overloaded the pictures. The byhand model of the small subunit illustrates both how good and how poor the fit to the protein map is. The crosslink between S4 (cyan) and G413 (cyan helix) on the left waist of the structure is too good to be true. The crosslinks between the purple helix (677:713) and S11 (the purple sphere on the right waist) do not appear to be a great problem on the data graph (Fig. 10). But in the color figure it is clear that in the present configuration, the crosslink vectors pass through a crowded section of the structure. More common are the crosslinks from S8 and S17 to the green helix on the lower right of the picture. The length of these vectors is longer than expected but the green helix could be easily adjusted to improve the fit by folding it into the empty region at the bottom of the model. Generally the byhand model of the 30S subunit would be improved if the proteins were moved upward along the y-axis relative to the RNA by a distance comparable to the separation of the centers of mass of the RNA and proteins.

The crosslink between S4 (outermost cyan sphere) and G413 (cyan helix) is a particularly good example of the problems in attempting to evaluate the fit of RNA and protein on the basis of crosslink length. In the present configuration the surface of the S4

sphere is just touching the RNA to which it is crosslinked. As the crosslink is most certainly to the base and not the phosphate backbone, the present alignment may put the base inside the radius of gyration of the protein. Since the protein may not be a sphere and the single-stranded base is free to rotate about the sugar-phosphate backbone, any translation or rotation of the byhand 16S RNA structure which leaves the cyan helix in the same quadrant cannot be excluded based on this crosslink.

The flexibility of the ends of 16S RNA is emphasized by the number and range of crosslinks to both the 5' and 3' termini. The S3, S4, S5, and S8 proteins which have been crosslinked to the 5' end and the S7, S18, and S21 proteins associated with the 3' end, form clusters on opposite sides of the waist of the protein map. But these loose neighborhoods span a significant volume with the S3 and S8 protein centers being separated by 90 angstroms and the space between them being densely occupied with both protein and RNA. The distance between the centers of proteins S18 and S7 is 81 angstroms but traverses an exterior and more open path. Immunoglobulin electron microscopy has directly mapped the 5' and 3' ends of the molecule to opposite sides of the 50S subunit face of the 30S subunit and confirms their location near the proteins to which the 5' and 3' ends are crosslinked. The byhand small subunit is in agreement with this data. The last74 and last70 16S RNA structures appear to have the 5' end directly under the 3' ends of the molecule. Viewed in stereo the ends are on adjacent vertices of the boxy waist of the models. Given the flexibility of the ends of the molecule, this conflict is not irreconcilable but suggests that some rotation of the RNA relative to the proteins should be made to improve agreement with this well mapped region. But as the crosslink graphs show, any such improvement will be made at the expense of another crosslink.

The S12 protein is the most highly conserved of the small subunit proteins and is implicated in the proper folding of 16S RNA. A crosslink between this protein and the oligonucleotides 1316-1322 of 16S RNA was reported although it has proven difficult to confirm (Chiaruttini et al., 1982). This crosslink would connect the loop of the red-striped

helix (1308-1329) to the red sphere in the lower left of the protein map. In the last74 30S subunit model, shifting the proteins up will bring these two objects into very close proximity (Fig. 6). The red-striped helix is positioned in the middle of the 50S subunit interface. Proteins S12 and S5 have been shown to protect this region plus bases near G530 and A900 (Stern et al., 1988b). These regions are represented in the last74 16S RNA structure as the beige helix just below the red-striped helix and the beige helix just to the right and behind the red-striped helix. Some researchers have been puzzled by these effects, but the computer generated models have clustered these regions so closely that it is easy to see that a pair of proteins could directly affect their reactivity.

By replacing the uridines which naturally occur in RNA with 4-thiouridine, it is possible to crosslink 16S RNA and proteins in functionally active 70S ribosomes. A low resolution mapping of these crosslinks reveals an unexpected relationship between the proteins S7 (rose) and S9 (beige) and bases 504-615 (beige backbone connecting the cyan and green helices) which form the junction between the 5' and middle domains of 16S RNA (Hajnsdorf et al., 1989). Most models have difficulty explaining this interaction between proteins found in the head of the 30S subunit and the red/beige/green section of 16S RNA which is usually sequestered in the base of most models. But as this region of 16S RNA forms the hinge upon which the protruding green helix must pivot in the computer generated models, it is not in the least surprising that such crosslinks could be formed. In the last74 30S subunit, these proteins are close together in the upper middle of the structure (Fig. 6). Although obscured by S1 (light green), the base of the green helix (residue 600) can be seen just to the right and behind these proteins.

Additional 30S Subunit Structural Relationships

The fit of the computer models of 16S RNA and the protein map is not a complete surprise. As the models of other researchers were made with reference to the protein map, the comparisons made between their models and the computer generated structures were comparisons to the neutron scattering map at one remove. What is significant, is that the

computer generated models were constructed without any reference to the protein data and yet show strong similarities to the protein map. There are other interesting ribosomal relationships which were not used to build the computer models because of the difficulty in defining the interactions in quantitative terms.

Some of the residues of 16S RNA have special properties that cause them to be considered as a class. The most obvious class is the group of nine modified bases that the 16S ribosomal RNA of *E. coli* normally contains (Brimacombe et al., 1988b). They have been shown to be involved in resistance to antibiotics and in translational control and proofreading. These bases (G527, G966, C967, G1207, C1402, C1407, U1498, A1518, and A1519) are scattered throughout the primary structure of 16S RNA. Another class of bases are formed by those that have different reactivities with chemical probes when a transfer RNA is bound to the ribosome (Moazed & Noller, 1986). The majority of these bases are clustered in the single-stranded regions near G530, G693, G791, A909, G926, G966, G1338, U1381, C1400, and A1492. A similar class of bases show changes in chemical reactivity when the ribosome is bound to an antibiotic (Moazed & Noller, 1987). Some of these bases overlap with those of the tRNA protection class and are clustered near G530, G693, G791, C912, G926, C1054, G1064, A1408, and A1494. When these bases are located in a stereo view of last74 a very striking relationship is revealed. All of the nucleotides lie in an arc along the bottom of the bowl formed by the junction of the protrudant green and blue helices and the waist of the rest of the 16S structure. This is the cleft/platform that should bind the messenger RNA that will be translated into protein. This makes it easy to explain the interplay of modified bases, tRNAs, and antibiotics on the basis of simple steric interaction. Perhaps too easy an explanation since other researchers find that it is necessary to invoke allosteric interactions in their models of ribosomal function. In the byhand model most of these special nucleotides are clustered in a tight knot just above the right waist. The exceptions appear in the irregular helix which connects the

orange-striped helices to the rest of 16S RNA. This region is visible as the grayish backbone section in the upper left of the model.

Several attempts have been made to determine what portions of the 16S RNA are involved in the association of the 30S and 50S subunits. A DNA probe complementary to the nine bases in the loop of the blue-striped helix (residues 787-795) was used to demonstrate that this region is solvent accessible and involved in the association of the 30S and 50S ribosomal particles (Tapprich & Hill, 1986). The blue-striped helix is the rightmost helix on the waist of last74 and the loop is extended toward the 50S interface free of any protein (Fig. 6). In addition to this region, it has been shown that the bases A673, A702, A802, C817, C1063, C1389, C1404, C1496, and G1515 have altered reactivities with chemical probes in the assembled ribosome (Herr et al., 1979). Bases A802 and C817 are part of the blue-striped region on the righthand side of the last74. Bases C1063 and C1389 are part of the beige helix which forms the top front of the 50S interface. Bases C1389, C1404, C1496, and G1515 are at the base of the protruding blue helix on the left side. The bases A673 and A702 are the only residues that could not directly interact with the 50S subunit in this model. They are part of the purple helix which is in the middle of the opposite side of the model, directly behind the red-striped helix. The fact that almost the exact same comments can be made about the fit of this data set and the byhand model, demonstrates the similarity between the computer generated models and the hybrid byhand model. The one difference is in the positioning of the residues at the base of the penultimate blue helix. In the byhand model the residues have disappeared into the blue S21 protein sphere on the right waist of the structure. A slight rotation of the 16S RNA might bring this region directly into the interface with the 50S subunit.

Mitochondrial ribosomal RNA is much smaller than other ribosomal RNAs, with many ribosomal functions apparently assumed by additional small subunit proteins (de la Cruz et al., 1985). When the 9S rRNA from the small subunit of a mitochondrial ribosome is mapped on last74, only the small section in the 5' stem thought to be involved in

association with the 50S subunit and the platform region which traverses the waist of the structure remain. Despite the drastic reduction from 1542 to 610 nucleotides, the remaining structural elements are closely packed into a rudimentary 'Y'-shaped conformation in the heart of the last74 model of 16S RNA.

These examples indicate that the computer models of 16S RNA and the 30S subunit are as successful as the models developed by other researchers. They may in fact be even more successful since the physical dimensions, the suggested displacement of the RNA and protein centers of mass, and the correlations between 16S RNA and the proteins and other portions of the ribosomal process, were developed in an objective manner. There is no possibility that any of the interesting features of the computer generated structures were used, even unconsciously, as parameters in developing the models.

Discussion

Some researchers have excluded the psoralen and GbzCynAc data because the nucleic acid attachment sites were identified by electron microscopy. The good results obtained in modeling 16S RNA demonstrate that an understanding of the chemical preferences of these crosslinkers and a cautious evaluation of the data derived from such studies, make it possible to derive quantitative constraints which can be used. The 30S subunit computer models constructed from the supposedly more precise RNA/protein crosslinks and the neutron scattering protein map appear to be less successful. That may be because the presence of numbers with decimal points and the single nucleotide resolution of crosslinks disguises the serious inadequacies of the protein data set.

The Immunoglobulin Electron Microscopy data must be considered a qualitative source which requires interpretation for two reasons. The first relates to the inherent limitations of electron microscopy. To successfully image a molecule it must be affixed to a support grid and placed in a vacuum. If the molecule does not contain strong electron scattering atoms, as neither proteins nor nucleic acids do, it must also be positively stained by application of heavy atoms to the molecular surface or negatively stained by being surrounded on the grid by strong scatterers. Uneven staining can produce severely distorted images. The data collected is a transmission image which best describes the silhouette of the molecule. This can produce mirror image problems as it can be difficult to tell whether an image is the X-ray outline of the palm of a right hand or the back of a left hand. Statistical reconstruction which combines the images of the various molecular orientations is necessary before three dimensional information can be obtained.

Using immunoglobulins as probes introduces another level of uncertainty which compounds the problems of electron micrograph interpretation. It is known that a hexa- or octapeptide is sufficient for IgG binding and that the N-terminal and C-terminal residues are especially antigenic (Westhof et al., 1984). Consequently even precise localization of the antibody binding site by EM may be 50-100 angstroms removed from the center of a

protein. Furthermore the proteins are not required to be symmetrically shaped and the radius of gyration may be only a fraction of the distance from the protein center of mass to its most extreme extension. Such errors might lead to attempts to place a particular protein or nucleotide sequence in the wrong domain or even on the wrong side of the 30S particle. Despite these uncertainties there is good general agreement among the Immune Electron Microscopy map, the neutron scattering map, and the computer generated models.

Testing the accessibility or protections of 16S RNA with chemical or enzymatic probes is not particularly incisive. Such experiments do illustrate the kinds of practical problems that make the study of the ribosome so difficult.

The mapping of the interactions between 16S RNA and the small subunit proteins demonstrates the unusual indirect effects that are sometimes present. The binding site of the protein S4 was narrowed to a region in the 5' domain of the molecule spanning the bases from the 5' end to U47 and U437 to the junction with the middle domain (C556) (Stern et al., 1986). Later work with oligoribonucleotides demonstrated that only the cyan helices (437-497, 500-545) were required to achieve full S4 binding affinity (Vartikar & Draper, 1989). Therefore the other chemical and protection effects are not due to direct protein binding but some secondary effect. S17 binds and can be crosslinked to the helix formed by nucleotides 240-286 (Greuer et al., 1987). S20 binding also increases the protection of this same region from attack by chemical probes (Stern et al., 1988a). Yet the neutron scattering map indicates that these two proteins are separated by more than 100 angstroms. Similarly, chemical probing indicates that both S16 and S20 protect the penultimate 3' helix (Stern et al., 1988a) while immune electron microscopy suggests that the 3' end of the molecule is on the opposite side of the 30S subunit (Lake, 1985). Finally protein S8 has protection effects throughout the sequence of the middle domain of 16S RNA (Svensson et al., 1988) (Stern et al., 1988b) but appears to bind directly only to the hairpin formed by oligonucleotide sequence 588-651 (Mougel et al., 1987). This limited range of direct binding is confirmed by three protein/RNA crosslinks (Wower & Brimacombe, 1983).

Oligonucleotide DNAs have been used to examine the accessibility of several regions of 16S RNA that appear to be single-stranded in the phylogenetic secondary structure map (Lasater, 1988). DNA/RNA hybridization was assayed by filter binding and rated as constant, conformation-dependent but good, conformation-dependent but weak, and weak. The computer models show excellent agreement with the results of these experiments with one exception. But because of the qualitative nature of the data and the large number of possible interactions, including steric conflicts, protein binding, conformational strain, RNA mobility, structural allostery, almost any result could be accommodated. The data do not require specific, verifiable RNA conformations, while at present the models are too imprecise and have too low a resolution to rule out any experimental result.

These experiments do indicate that there are dynamic rearrangements of the RNA folding pattern. There may be cooperative binding effects which increase the affinity of a protein for RNA. There are also significant changes in the gross characteristics such as the radius of gyration of the small subunit depending on how carefully the particle is washed after reconstitution. (Ramakrishnan, 1986) Such nonspecific protein binding may explain the crosslink that has been found between green-striped helices at the bottom of the structures and the protein S13 which appears at the top of the neutron map (Osswald et al., 1987). The most daunting possibility in modeling 16S RNA is that the data are correct and the RNA is undergoing constant rearrangement. To the extent that this is true it should not be expected that any one model will be able to reconcile all the data from immune electron microscopy, neutron diffraction, and RNA protection studies.

The map of the proteins derived from neutron scattering experiments appears to be quantitative and reliable. Although there are some differences between this map and the earlier ones derived from fluorescence energy transfer and immunoglobulin electron microscopy, it does appear that these discrepancies are due to weaknesses in the other techniques (Capel et al., 1988). It is also worth noting that the large radii of gyration for

some of the proteins allow almost anything to be rationalized. Some of the proteins, specifically S7 and S9, are placed differently in the latest map than they were in an earlier map derived from the nearly complete set neutron scattering data. These proteins jump around in the protein map with each data addition to the data set while most of the other proteins are relatively fixed. This may be due to a combination of the uncertainties in the interprotein distances and their sizes. It might be caused by the subtle differences between 30S particles which may occur unless great care is exercised during their reconstitution (Ramakrishnan,1986). Some problems may derive from the simple geometric algorithm used to construct the three dimensional mapping. As in distance geometry modeling, it is possible to create an enantiomer of the map which will satisfy all the neutron scattering data. This problem applies not just to the global conformation but to local structures as well. Combined with the uncertainties in some of the interprotein distances, the protein mapping may produce diastereomers as well. The present map has been checked against the IEM map developed by Lake and coworkers (Lake, 1985) in an attempt to avoid this problem.

Distance geometry was used to create a map of the ribosomal proteins of the small subunit from a much smaller set of interprotein distances than is presently available (Kuntz & Crippen, 1980). This early work was hampered by two major problems. First was the ill-defined shape of the proteins. Nothing is known about them beyond their sedimentation velocities and the apparent radii of gyration. A unique three-sphere representation for the proteins was devised which attempted to convey the uncertainties involved. This approach was so unusual that no other researcher has attempted to repeat or continue the work. Given these amorphous constructs, the problems in assigning reasonable distance constraints between the proteins based on neutron scattering and IEM data were exacerbated. Despite these problems, the resultant structure is comparable to the present protein map. The strong similarity of this early 30S subunit protein map and the complete protein map derived from neutron scattering studies demonstrates that distance geometry

can be used with some success even in the very early stages of data collection. Perhaps it would be appropriate to use the improved distance geometry algorithms and the larger neutron scattering data set to create a protein map of the 30S subunit.

Although all the compounds used to crosslink 16S RNA and the proteins of the small subunit are small enough to indicate that there is direct contact between a section of the RNA and the protein, the unavoidable uncertainties of the present data set turn what appears to be hard quantitative results into qualitative indicators. As the phosphate to base distance can be ten angstroms and the smallest protein has a radius of gyration of eleven angstroms, a crosslink vector length of twenty angstroms should be considered the minimum separation and may well indicate that the RNA and protein are too close together. The crosslink vector for the 3' dodecanucleotide and S21 could be very long when displayed in the present manner. If the crosslinked RNA residue lies at either end of the dodecamer, the base used to calculate the distance could be as much as 25 angstroms away. It is doubtful that the ribosomal proteins are perfectly spherical and the radius of gyration may not indicate just how extended a protein is. If S21 has an unusual shape, perhaps like that of transfer RNA which has a similar radius of gyration, the crosslinked amino acid residue could be as much as 40 angstroms distant from the protein center of mass. In this case a crosslink vector length of 65 angstroms would indicate an exact alignment of the 16S RNA structure and the protein map. Therefore the RNA/protein crosslink distances cannot be uncritically used to evaluate the models of the 30S subunit.

The goal of the computer modeling protocol is not the production of models but an improvement in understanding the ribosome. On this basis the new protocol is successful in facilitating the coordination of disparate facts into a plausible explanation of some aspects of the ribosome. Although it can be seen most dramatically in a stereo view that cannot be included in this manuscript, the last74 model suggests an elegant system for the positioning and translocation of tRNA and mRNA on the small subunit of the ribosome during protein translation (Fig. 5 and Fig. 22 of the preceding chapter). Immune Electron Microscopy

studies indicate that messenger RNA and transfer RNA interact with the cleft/platform region of the 30S subunit (Olson et al., 1988). One model of the interaction between tRNA, mRNA, and the ribosome shows a tRNA lying on its side on the edge of the cleft of the 30S subunit. The tRNA is positioned with anticodon end on the solvent face and the aminoacylated end reaching around into the cleft of the small subunit which faces the 50S subunit (Lake, 1985). The identification of the 50S/solvent faces of the last 74 30S subunit model suggests a different conformation and placement of the cleft/platform structure that resembles the German IEM model (Stoeffler & Stoeffler-Meilicke, 1986). In this model the platform that would be formed by folding down the green and blue helices, would be facing the S1 protein. The tRNAs and mRNA would bind in the space between 16S RNA and S1, protected from solvent interactions. The charged ends of the tRNA molecules would then reach over the top of the small subunit, perhaps as one reaches over the counter at a delicatessen, placing the charged ends of the tRNA's in the gap between the two ribosomal subunits. Experimental data which indicate that thiolated aminoacyl-tRNAs can be crosslinked to both subunits and both 16S RNA and 23S RNA, could be produced by tRNAs in a sidepocket as reasonably as tRNAs sandwiched between the two subunits (Barritault et al., 1981).

The models of the small subunit reconstructed from electron microscopy (Verschoor et al., 1984) show the platform forming a lip beneath the head which extends almost completely around the entire subunit and the computer generated structures strongly resemble such a shape. This extended shape for the cleft also correlates nicely with the large number of transfer binding sites found in a recent set of experiments (Moazed & Noller, 1989). In this context, the size and shape of tRNA take on physical explanations. Incoming tRNAs would bind to the mRNA in the cleft far to one side with the aminoacylated end of the tRNA reaching around into the interface between the two subunits. The nascent protein is transferred to this tRNA. Then as the ribosome advances along the mRNA to the next codon, the tRNA slides over the face of the head, maintaining

contact with the mRNA. This motion might require the retrograde pulling of the growing protein back up through the extrusion hole in the 50S subunit. This dynamic tension would provide excellent opportunities for proofreading and translational control. The tRNA carrying the peptide then snaps down into the site on the other side of the small subunit where it can easily reach around into the interface. Such a mechanism would require that the shape and size of the head of the 30S subunit be highly conserved and indeed the knobs and projections seen in the ribosomes of other species occur at the side or bottom of the 30S subunit. This process resembles a catch/release clockwork which could rapidly and reliably proceed along the mRNA pullchain because it relies on simple physical mechanisms and not the subtleties of wobble basepairs. The electron micrographs of the 30S subunit/mRNA complex show antibodies in just these positions (Olson et al., 1988).

This model would explain how transfer RNAs can simultaneously interact with the messenger RNA and the growing protein without interrupting the tight 30S/50S couple. A more mechanical linkage of mRNAs, tRNAs, ribosome and the protein product also has a kinetic appeal. Translational proofreading that relied on the hydrogen bonding information in the major or minor groove of a nucleic acid helix or the codon/anticodon basepairing of the messenger RNA and the transfer RNA would be very slow. The ribosome and other polymerases must operate at very high speeds while maintaining accuracy. Detecting larger conformational features would be much easier. The detection of some irregularity could then act as a signal for more precise proofreading.

The computer models of 16S RNA and the 30S subunit offer some provocative explanations of the characteristics and operation of the small subunit of the ribosome. It is unrealistic to think that the computer generated models will prove to be the correct structure in any detail. In fact the RNA/protein crosslink data suggest that substantial changes in the models will need to be made. But when the shortcomings and difficulties of other attempts to decipher the ribosome are considered, it is clear that the computer protocol is a substantial improvement.

Conclusion

Final Considerations

This work has demonstrated that a significant part of the structure of the small subunit of the ribosome can be derived from the primary and secondary structure of 16S RNA with a minimum of tertiary structure information. The assumption that helices should be emphasized simplified the structure and allowed the double-stranded constructs to dominate the folding process. The inclusion of distance geometry in the modeling protocol makes it possible to produce a reasonable three dimensional structure in an objective manner. When modeling a complex system like the ribosome, there is a danger that the researcher will claim that his model predicts an interaction or conformation which was used directly or by implication to construct the model (Brimacombe, 1988a). The possibility of becoming trapped in such circular arguments has been eliminated with computer modeling protocol. The more subtle effects that might influence the modeling process which result from the inclinations or amorphous conjectures of a human modeler have also been avoided. The input data for the construction of computer models is required to be discrete and clearly defined. However care must be exercised in the selection and processing of the input data lest the dreaded "Garbage In, Garbage Out" syndrome appear. It is also important to maintain a clear perception of what the strengths and weakness of the modeling programs are within the present context.

A series of papers have criticized the manner in which distance geometry searches conformational space (Levy et al., 1989). In a test case, the same distance geometry program used in this study, DSPACE, produced only one structure and its enantiomer from a set of distance constraints. The researchers did a lot of Monte Carlo searching and restrained molecular dynamics to show that there are many possible neighboring conformations which distance geometry does not find. But their extensive analysis revealed that the best structure is the one produced by distance geometry. This lends support to the theory that distance geometry preferentially finds the proper conformation (Crippen, 1987).

More troublesome is the result that the program does not reproduce the structure from which a set of NMR distance constraints are derived. Perhaps this indicates a tacit weighting of the final structure by the large mass of the template data as opposed to the smaller number of NMR distance constraints. As the modeling of 16S RNA also found essentially one structure, this criticism of distance geometry cannot be denied, but is it a weakness?

Despite these reservations, the success in predicting the folding of transfer RNA, the ability to form a reasonably correct small subunit protein map from minimal data, and the strong correspondences of the computer generated models of 16S RNA with data outside the input parameters, argue that distance geometry can be reliably applied to the search for the global conformation of RNA structures. The ability to produce extended structures from a partial data set may be applicable to the attempts to understand multistep folding processes. A better appreciation of the possible intermediate structures will make it easier to comprehend the forces which are responsible for the final folded form.

The computer protocol was designed to introduce objectivity into the folding step of the modeling process. To compensate for the loss of human intelligence in the early stages, computer graphics would be employed to facilitate the use of intuitive visual skills in evaluation the models. Subjective judgements have been moved from the initial stages where their effect on the final structure may be hidden, to the end of the process where it belongs. When converted to raster pictures, the computer models convey a strong sense of three dimensional structure of the molecule and suggest possible working relationships. In fact the graphics images are so powerful that there can be a tendency to over interpret the structures. As the replacement set of pseudoatoms is so severely reduced and the number of long range constraints is so low, any discussion of atomic resolution structure is mere guesswork. Attempting to predict specific hydrogen bonds or the orientations of particular nucleotides is not possible. It is practical to search for a region of conformational space that represents the global preference. Finding an area of conformational space that contains

many closely related structures is also significant, especially in light of the flexible structure of the ribosome.

Future Directions

Now that the feasibility of this modeling approach has been demonstrated, some improvements in the protocol should improve its effectiveness. The easiest change would be to add a new type of pseudophosphate to the distance geometry library of atoms. This new atom type would be used to give single-stranded phosphates a longer and more flexible bonding character. To improve the spacefilling characteristics of the pseudohelices, it might be desirable to add a few pseudoatoms in the middle of the longer helical strands. Finally, if the five pseudoatom per nucleotide replacement constructs which were used in the early stages of transfer RNA modeling were added to the AMBER library, they would provide an intermediate minimization step to smooth the transition from the fully reduced structure to the all atom representation.

With the modeling protocol established there are a large number of projects to which it could be profitably applied. The simplest would be to repeat the modeling of 16S RNA including the changes in the secondary structure map (Noller et al., 1987) and the improved tertiary phylogenetic relationships (Haselman et al., 1989) in the input parameter set. Less trivial but more interesting would be to include the proteins in the structure construction process. This will simultaneously bring three important additions to the modeling of the small ribosomal subunit. First it would provide an independent verification of the protein map constructed from neutron scattering data. Next it would provide the necessary distance constraints which should produce a further folding of 16S RNA, especially of the two protruding helical domains. Finally it should promote an automatic alignment of the protein and RNA structures and make it easier to eliminate the improper enantiomer. A less attractive but important project would be to convert the other models of 16S RNA into the pseudohelical representation for direct, quantitative comparison. This could lead to a consensus on the three dimensional structure similar to

that which was found when the secondary structures were finally compared on an even footing.

Repeating the modeling procedure for the 23S rRNA of the large ribosomal subunit should be possible although the number of known 23S sequences and long range constraints are much smaller than that available for 16S rRNA. Finally the techniques developed by this project should be applied to the determination of the structures for the rapidly expanding class of catalytic self-cleaving RNAs.

The End

In the final analysis there can be no doubt that development of an RNA modeling protocol has been successful. It is simple, elegant, and powerful. The modeling rationale predicts that helical subunits will predominate, it reduces the size of the problem, and it produces objective models. Having been constructed with tools and procedures that are well-defined and easy to obtain, these methods can be moved to other computer systems and applied to other molecules. The evaluation of the model of 16S RNA is less certain. The comparison with physical data and the models developed by other researchers reveals some significant but not irreconcilable differences. Even if the model should eventually prove to be a poor reflection of the real molecule, it can play a valuable role in coordinating and stimulating the study of the ribosome.

References

- Barritault, D., Buckingham, R.H., Favre, A., & Thomas, G. (1981) *Biochimie* 63, 587-593.
- Brimacombe, R. (1988) *Biochemistry* 27, 4207-4214.
- Brimacombe, R., Atmadja, J., Stiege, W., & Schueler, D. (1988) *Journal of Molecular Biology* 199, 115-136.
- Capel, M.S., Engelman, D.M., Freeborn, B.R., Kjeldgaard, M., Langer, J.A., Ramakrishnan, V., Schindler, D.G., Schneider, D.K., Schoenborn, B.P., Sillers, I.-Y., Yabuki, S., and Moore, P.B. (1987) *Science* 238, 1403-1406.
- Capel, M.S., Kjeldgaard, M., Engelman, D.M., & Moore, P.B. (1988) *Journal of Molecular Biology* 200, 65-87.
- Chiaruttini, C., Expert-Bezacon, A., Hayes, D., & Ehresmann, B. (1982) *Nucleic Acids Research* 10, 7657-7676.
- Crippen, G.M. (1987) *Journal of Physical Chemistry* 91, 6341-6343.
- Dahlberg, A.E. (1989) *Cell* 57, 525-529.
- de la Cruz, V.F., Lake, J.A., Simpson, A.M., & Simpson, L. (1985) *Proceedings of the National Academy of Science* 82, 1401-1405.
- Greuer, B., Osswald, M., Brimacombe, R., & Stoeffler, G. (1987) *Nucleic Acids Research* 15, 3241-3255.
- Hajnsdorf, E., Favre, A., & Expert-Bezancon, A. (1989) *Nucleic Acids Research* 17, 1475-1491.
- Hardesty, B., Odom, O.W., & Deng, H.-Y. (1986) In *Structure, Function, and Genetics of Ribosomes* (Hardesty, B. & Kramer, G., eds) pp. 495-508, Springer-Verlag, New York.
- Haselman, T., Camp, D.G., & Fox, G.E. (1989) *Nucleic Acids Research* 17, 2215-2221.

- Herr, W., Chapman, N.M., & Noller, H.F. (1979) *Journal of Molecular Biology* 130, 433-449.
- Kuntz, I. D., & Crippen, G.M. (1980) *Biophysical Journal* 32, 677-695.
- Kyriatsoulis, A., Maly, P., Greuer, B., Brimacombe, R., Stoeffler, G., Frank, R., & Bloecker, H. (1986) *Nucleic Acids Research* 14, 1171-1186.
- Lake, J.A. (1985) *Annual Review of Biochemistry* 54, 507-30.
- Lasater, L.S., Olson, H.M., Cann, P.A., & Glitz, D.G. (1988) *Biochemistry* 27, 4687-4695.
- Levy, R.M., Bassolino, D.A., Kitchen, D.B., & Pardi, A. (1989) *Biochemistry* 28, 9361-9372.
- Moazed, D. & Noller, H.F. (1986) *Cell* 47, 985-994.
- Moazed, D. & Noller, H.F. (1987) *Nature* 327, 389-394.
- Moazed, D. & Noller, H.F. (1989) *Nature* 342, 142-148.
- Mougel, M., Eyermann, F., Westhof, E., Romby, P., Expert-Bezancon, A., Ebel, J.-P., Ehresmann, B., & Ehresmann, C. (1987) *Journal of Molecular Biology* 198, 91-107.
- Mougel, M., Philipe, C., Ebel, J.-P., Ehresmann, B., & Ehresmann, C. (1988) *Nucleic Acids Research* 16, 2825-2839.
- Noller, H.F., Stern, S., Moazed, D., Powers, T., Svensson, P. & Changchien, L.-M. (1987) *Cold Spring Harbour Symposia on Quantitative Biology*, 52, 695-708.
- Olson, H.M., Lasater, L.S., Cann, P.A., & Glitz, D.G. (1988) *Journal of Biological Chemistry* 263, 15196-15204.
- Osswald, M., Greuer, B., Brimacombe, R., Stoeffler, G., Baeumert, H., & Fasold, H. (1987) *Nucleic Acids Research* 15, 3221-3240.
- Powers, T., Changchien, L.-M., Craven, G.R., & Noller, H.F. (1988) *Journal of Molecular Biology* 200, 309-319.

- Powers, T., Stern, S., Changchien, L.-M., and Noller, H.F. (1988) *Journal of Molecular Biology* 200, 697-716.
- Ramakrishnan, V. (1986) *Science* 231, 1562-1564.
- Stern, S., Changchien, L.-M., Craven, G.R., & Noller, H.F. (1988) *Journal of Molecular Biology* 200, 291-299.
- Stern, S., Powers, T., Changchien, L.-M., & Noller, H.F. (1988) *Journal of Molecular Biology* 201, 683-695.
- Stern, S., Wilson, R.C., & Noller, H.F. (1986) *Journal of Molecular Biology* 192, 101-110
- Stern, S., Powers, T., Changchien, L.-M., & Noller, H.F. (1989) *Science* 244, 783-790.
- Stoeffler, G. & Stoeffler-Meilicke, M. (1986) In *Structure, Function, and Genetics of Ribosomes* (Hardesty, B. & Kramer, G., eds.), pp. 28-46, Springer-Verlag, New York.
- Svensson, P., Changchien, L.-M., Craven, G.R., & Noller, H.F. (1988) *Journal of Molecular Biology* 200, 301-308.
- Tappich, W.E., & Hill, W.E. (1986) *Proceedings of the National Academy of Science* 83, 556-560.
- Vartikar, J.V. & Draper, D.E. (1989) *Journal of Molecular Biology* 209, 221-234.
- Verschoor, A., Frank, J., Rademacher, M., Wagenknecht, T., & Boublik, M. (1984) *Journal of Molecular Biology* 178, 677-695.
- Westhof, E., Altschuh, D., Moras, D., Bloomer, A.C., Mondragon, A., Klug, A., & Van Regenmortel, M.H.V. (1984) *Nature* 311, 123-126.
- Woese, C.R., Gutell, R., Gupta, R. & Noller, H.F. (1983) *Microbiology Reviews* 47, 621-669.
- Wower, I., & Brimacombe, R. (1983) *Nucleic Acids Research* 11, 1419-1437.

LAWRENCE BERKELEY LABORATORY
TECHNICAL INFORMATION DEPARTMENT
1 CYCLOTRON ROAD
BERKELEY, CALIFORNIA 94720