Probabilistic Structure Calculations: A Three-Dimensional tRNA Structure from Sequence Correlation Data

Russ B. Altman, MD, PhD
Section on Medical Informatics
Stanford University, MSOB X215
Stanford, CA 94305-5479
altman@camis.stanford.edu

Abstract

Algorithms based on probability theory can address issues of uncertainty directly through their representational framework and their theory for data combination. In this paper, we discuss the advantages of probabilistic formulations for molecular-structure calculations, describe one implementation of such a formulation, and show its performance on a data set derived from analysis of the statistical correlations within a set of aligned transfer RNA sequences. By assigning reasonable physical interpretations to certain statistical correlations, we are able to calculate three-dimensional structures for tRNA from a random starting structure. The constraints that we use are associated with different variances, and so their effects are not uniform, and must be reconciled by a probabilistic algorithm to yield the most likely structure. As might be predicted, the uncertainty in the position for each base is a function of both the number and strength of the constraints, and is reflected in the variances in atomic position calculated by the algorithm. For example, the hinge region in the tRNA is shown to be the most uncertain. In addition, the algorithm retains information about positional covariation that is useful for understanding the relationships between different parts of the structure. These experiments also demonstrate that we can define a single-sphere representation for each base that is useful for nucleic acid structural calculations in the same way that alpha-carbon representations are useful for protein structural calculations.

Introduction

The determination of molecular structure is one of the key goals in molecular biology. The human genome project is already producing large numbers of sequential data, and there is a need for efficient methods for determining the structures that correspond to these sequences. Experimental methods have been the primary means of structure definition, but are subject to problems of high cost and biased sampling of structure-space[1, 2]. Techniques for sequence analysis provide information about likely structural features based on prediction of secondary structure [3, 4], prediction of structural class [5], and prediction of functional class[6]. There is also an increasing amount of information available based on the detailed analysis of known molecular microenvironments, that shed light, for example, on the ways in which nucleic acid base pairs or amino acid side chains interact[7, 8].

Most of the information derived from analysis of structure or sequence does not perfectly discriminate among the various possibilities; rather, continuums of possibilities exist with relative likelihoods that vary within different contexts. Thus, the information is often not, *in isolation*, sufficient to make a clear structural assignment. Instead, data sources often suggests that certain possibilities are more likely than are others. Similarly, there are often experimental data sets that are not sufficient to define an entire structure clearly, but provide useful information about aspects of the structure. The problem of structure prediction, then, can be viewed as a problem of combining multiple, probabilistic data sources, and arriving at a consensus structure that reflects the information contained within these data sources. An important subproblem is defining a framework within which these data sources can be combined.

Methods based on energy minimization [9] or molecular dynamics [10] are the most commonly used techniques for calculating structures from multiple data sources.[1] Although the fundamental theory behind these techniques is sound (that is, modeling the physical forces, or energies, that act upon atoms within a molecule), the practicalities of the technique require that approximations be made to make problems more tractable. For example, information derived from experiment or theory has been incorporated into these models by adding *pseudoenergy* terms into the general energy equations. One of the earliest uses of energy-based algorithms with nonphysical pseudoenergies was the use of experimental nuclear magnetic resonance (NMR) measurements (which provide information about the distance between protons in a molecule) as separate forces and the combination of these forces with the more traditional physically-based energy terms [13, 14]. Although these approaches worked well in defining protein structures from NMR data, they also illustrated one of the pitfalls of using the energy paradigm for general protein-structure determination: the relative strength of the energies was adjusted arbitrarily to make the algorithms work. It was difficult to predict the weights of the NMR pseudoenergies relative to the weights used for the energies describing van der Waals,

[1] Distance geometry algorithms [11, 12] might also be considered, but they are limited to the use of distance information, and so do not represent a general solution to the problem of combining multiple sources of data.

coulombic and other forces [14]. Essentially, these weights had to be determined empirically, and typically the NMR energies were given a huge weight compared with weights on the other force terms (that is, they dominated the calculation).

The energy-based structure methods may, therefore, not be the best for combining the multiple sources of data that are gathered about protein structure: as nonatomic representations are introduced, these algorithms become further removed from their theoretical bases, and it becomes more difficult to guarantee their performance characteristics. In addition, as these energy-based techniques are used in problems that are underconstrained (that is, a single answer is not expected), they become less useful because of the difficulty of characterizing clearly the full set of structures that is compatible with the data.

We have been working on an alternative formulation of the structural search space and the constraints that are provided to it. Our formulation is a Bayesian probability-based one, in which all constraints are represented as probability distributions over parameters that are calculable from the atomic coordinates. The position of each atom is also represented as a probability distribution. The language of probability is a natural common language for representing the available data sources, and this common language allows the distributions of values (or variance) within these data sources to be used and compared directly. In the case of NMR data mentioned earlier, each measurement has a variance based on the characteristics of the experiment (usually about 3 to 9 $Å^2$), whereas covalent bonds have a variance of 0.01 $Å^2$. The relative strengths of these constraints (as represented by their variance) are used by our algorithm to resolve conflicts within the data set. Thus, for example, bond lengths that deviate from their expected length by 1 Å are totally unacceptable, whereas NMR measurements with the same deviation are acceptable. The other important characteristic of probabilistic formulations for structure calculations is that they can provide explicit estimates of the uncertainty in the atomic locations. These estimates are necessary in cases where the structure in question is underconstrained by the data set (a common situation for structure prediction) and also are useful to programs (such as those based on energetics) that use the structures as starting points for further refinement. In the following sections, we illustrate some of the benefits of these representations and algorithms in the context of calculating the three-dimensional structure of transfer RNA.

The Tertiary Structure of transfer RNA

Our algorithm, as outlined later, was originally created for analysis of protein structure from NMR data, and has been applied successfully to that problem [15]. In this paper, we describe application of our algorithm to the determination of the three-dimensional topology of transfer RNA (tRNA) structure. Recently, investigators have been able to analyze aligned sequences of tRNA and extract correlations that shed light on the structure and function of these molecules. For example, Gutell and coworkers [16] have shown that there are associations within sets of RNA molecules that provide information about base pairing and other structural interactions. These interactions constrain the set of possible conformations which the tRNA can assume. In addition, Klinger and Brutlag [17] have replicated some of the results of Gutell and have framed their work in an explicitly conditional-probabilistic framework. We have taken the constraints inferred by these investigations, and tested the hypothesis that *these constraints provide enough information to define the three-dimensional structure (and perhaps some areas of detailed structure) for tRNA*. Since there are crystal structures known for two tRNA molecules (and their variants)[18, 19], we have a way to validate our structures and to evaluate their utility in cases where the structure is unknown. The problem of estimating structure using statistical constraints is, in many ways, ideally suited to our probabilistic approach: the data are relatively sparse (high-resolution structures are not expected), the uncertainty in the position of each base is of great interest (for example, in evaluating which sections of the molecule are well constrained by the data), and the statistical correlations discovered by these investigators do not correspond to real physical forces or energies. Other approaches to the problem of RNA three-dimensional structure prediction have been based on symbolic constraint satisfaction [30], manual model building [31, energy minimization and molecular dynamics, and have been recently reviewed by Gautheret and Cedergren [32]. None of these approaches are focused on the uncertainties in the resulting structures.

Our data set was provided by Klingler and Brutlag, and it is summarized here and is described in detail in [17]. After aligning 1208 sequences (using the standard numbering conventions from 1 to 76 for bases), they found three types of positional correlations:

1. There were isolated base pair correlations in which a base at position *i* was found to be highly correlated with the identity of a base at position *j*. These correlations took the form of standard base pairing (that is, if base *i* was A, then base *j* was T), and were observed for the following pairs of bases: 8 and 14, 15 and 48, 19 and 56, 54 and 58.

2. There were runs of four or more bases correlated with identical length runs on distant parts of the sequence in an antiparallel, base pairing sense. They found four instances of highly correlated segments of sequence in which the identity of base *i* was correlated with the identity of base *j* (in the base pair sense mentioned), the identity of base *i+1* was correlated with the identity of base *j–1*, and so on for lengths of four to six bases. Such

runs were found for 1-7 with 72-66 (the "stem-loop" helix), 10-13 with 25-22 (D-stem), 27-31 with 43-39 (anticodon stem), and 49-53 with 65-61 (T-stem).

3. There were three-way correlations in which the identity of base *k* correlated to the identity of bases *i* and *j*, and for which *i* and *j* were correlated in a base pairing sense. Base 46 was found to correlate with the 13-22 base pair, and base 9 correlated with the 12-23 base pair.[1]

We postulated that by making assumptions about the physical orientations implied by these correlations, we could calculate the general topology of a typical tRNA molecule from a randomly generated starting structure. Thus, the constraints of type 1 could be interpreted as lone Watson-Crick base pairs, the constraints of type 2 were interpreted to be runs of base paired strands adopting an A-form double helix (the conformation of most often assumed by RNA), the constraints of type 3 were interpreted to be Hoogsteen three-way basepairs in which a third bas hydrogen bonds to each of two Watson-Crick bases [20]. The choice of A-form double helix can be justified based on independent observations that this conformation is the most common one for base paired RNA. The Hoogsteen base interaction is the most likely explanation for the strong correlations among three base positions.

Methods

For the purposes of our initial calculations, we decided to use a simplified representation of the RNA bases. Each RNA base was modeled as a single-sphere, and all constraints between atoms in the bases were transformed into constraints on the position of the spheres representing this bases. Thus, there was a significant loss of precision, since all atoms within the nucleoside, phosphate backbone and ribose moieties (the maximum dimension being roughly 13 Å) were mapped to a single point. We experimented with two different representations: using the center of mass (COM) of the purine/pyrimidine group and using the glycosylic nitrogen (GN). The GN representation was preferable because it is more central than is the COM, and so there was less error (on average) in the constraints. For example, since the GN is closer to the backbone (phosphate and ribose) of the RNA, the constraints that express the covalent connectivity between base *i* and base *i+1* have a much lower variance. Interestingly, there is a larger variance on the distance between the COMs of base paired bases than for the GNs of base paired bases—even though the GNs are a few Ångstroms farther from the actual hydrogen bonds. This observation indicates that the overall width of a base paired, double-helical RNA molecule is, in some sense, more precisely conserved than are the distances between bases. Part of this conservation is probably due to base-twist effects, which are variable, and thus affect the distance between COMs more than they affect distance between GNs.

The advantages of a single-sphere representation include a greatly simplified search problem. Instead of trying to position nearly 1700 heavy atoms (or more than 2000 atoms, if all hydrogens are included), we are positioning only 76 (one for each base). Once a rough topology is defined, we can use the position of the single-sphere to calculate an approximate position for all the atoms, and then perform a refinement of the full-atomic representation. The refinement subsequently explores a much smaller space, benefitting from the structure generated using the simplified representation. Reduced representations of this sort have been used in protein-structure calculations extensively [15, 21, 22].

The disadvantages of the single-sphere representation are clear. There is markedly reduced precision. In this case, we postulated that the relatively small amount of data (covalent chemistry plus some statistical correlations corresponding to distances) would not be sufficient to specify a high-resolution structure, in any case, and so we accepted this limitation. In addition, the ability to use van der Waals constraints effectively is greatly limited. Whereas a full atomic representation of the nucleotide and its backbone would have an irregular van der Waals surface (including a planar nucleoside and more extended and globular backbone segment), the single-atom representation allows only spherically symmetric checks for overlap. Thus, there is no doubt that our representation adds imprecision to the calculation.

Constraints

Using a single-sphere approximation, centered on the glycosylic nitrogen, we generated distance constraints corresponding to each of the correlations described above.

1. Isolated basepairs: We measured the mean distance between base paired bases (*i* and *j*) within the A-Form crystal of RNA (8.83 Å), and increased the variance to 1.0 $Å^2$ to allow for variations in the distance between GNs in the case when a base pair occurs in isolation (in such cases, the backbone can adopt numerous conformations, and so the intrinsic variability is much greater than the variation of the same distance within a base paired double helix). We supplied these values to the program for each of the isolated base pair correlations.

2. Base pair runs: We measured the distances between the GNs within the A-form crystal of RNA [[23]]. The mean distance between base paired bases (*i* and *j*) was 8.83 Å, with a variance of 0.1 $Å^2$. We increased the variance to 0.3 $Å^2$, to account for a relatively small

[1] Klinger and Brutlag also discovered correlations that may correspond to base stacking interactions, but these were not used in this calculation.

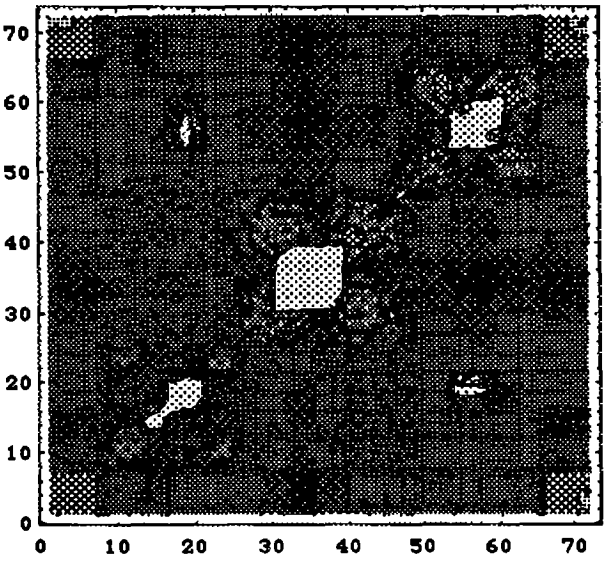

Figure 4. Graphical contour representation of final coordinate covariance matrix. In this plot, darker shades correspond to lower values. The variance of each base is plotted along the diagonal, and corresponds to a contour view of Figure 2. The covariance between the position of bases is represented in the off-diagonals. Thus, for example, there is a strong correlation in the position of bases 19 and 55. On the other hand, there is little correlation between the position of bases 60 and 34.

sample (N=20). We also measured the distances between the GNs of bases within a six base run of helical structure, in order to capture the helical structure of A-form RNA (failure to provide information about twist would lead to the possibility of base paired railroad tracks with no spiral, and so this information was provided as part of our assumption that the RNA adopted a double helical structure). We used these values for all runs of base pair correlation.

3. Base triplets: We measured the theoretical distance between GNs of the third base in a Hoogsteen base triplet from the other two bases to be 9.5 Å. We set the variance to a large value (3.0 $Å^2$) again to account for large potential mobility.

In addition to these constraints, we provided constraints to express the covalent linkage between neighboring bases *i* and *i+1*, *i+2*, *i+3* in an RNA molecule. By manipulation of graphical models of an RNA polymer, and measurement of the GN distances between sequential bases, we estimated the mean distance between GNs to be 5.75 Å with a variance of 3.2 $Å^2$. Finally, after similar manipulation of molecular models of RNA, we imposed a constraint that no two GNs (except those that are neighbors in sequence) could be closer than 4.5 Å. This distance corresponds to the effective van der Waals radii for the glycosylic nitrogens. This minimum distance is closely approached in the A-form conformation, but is otherwise a poor approximation for most contact surfaces between bases.

Outline of the Algorithm

The double iterated Kalman filter (DIKF) is a probabilistic algorithm for combining multiple uncertain measurements (with different variances) and calculating the most likely structure compatible with these constraints. It is described in detail in [22, 24, 25]; we summarize it here.

There are three types of information that our algorithm uses: an estimate of the mean position of each point, an estimate of the variance/covariance between all coordinates of all points, and a representation of the underlying model of the data and its uncertainty. For molecular structure, the parameters to be estimated are the coordinates of atoms in three-dimensional space. We represent the mean positions of each atom as a vector, **x**, of length *3N* for *N* atoms:

$$\mathbf{x} = [x_1\ y_1\ z_1\ x_2\ y_2\ z_2\ \ldots\ x_N\ y_N\ z_N]^T \qquad [1]$$

The second element of our representation is a variance/covariance matrix for vector **x**. This matrix, **C(x)**, contains the autocovariance information for vector **x**: the diagonal elements contain the variances of each element of **x**, whereas the off-diagonals contain the covariances among the elements within **x** (for *N* atoms, **C(x)** is of size *3N* x *3N*) :

$$\mathbf{C(x)} = \begin{pmatrix} \sigma^2_{x_1} & \sigma^2_{x_1 y_1} & \cdot & \cdot & \sigma^2_{x_1 z_N} \\ \cdot & \sigma^2_{y_1} & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ \sigma^2_{z_N x_1} & \cdot & \cdot & \cdot & \sigma^2_{z_N} \end{pmatrix} \qquad [2]$$

The variance of each coordinate of an atom can be extracted from the diagonal and provides the uncertainty in the location of that atom. The off-diagonal elements of the variance/covariance matrix contain information about the dependence between the coordinates of two atoms (that is, the dependence of the position of one atom on the position of the other). Each off-diagonal element is a linear estimate of the relationship between two coordinates. It is related to a correlation coefficient by a normalization term. If the element is positive, then the two coordinates are positively correlated. This information is critical to the search: we can propagate to other atoms a change in the position of a particular atom

using this first order estimate of their relationship. Thus, the off-diagonal *3 x 3* submatrices represent a summary of how the position of one atom changes as the position of another is modified. Our representation, therefore, resembles network data structures which allow linear weights between nodes [26]. As more is learned about the relationships between atoms, the network of dependencies grows (for a graphical example of a covariance matrix, see Figure 3). Eventually, the movement of any atom results in the concerted movement of all other atoms based on this covariance information. The precise mechanisms for updating estimates of the mean vector and covariance matrix using probabilistic constraint are discussed in [22, 24, 25].

Representation of Constraints

We take a constraint to be any information that constrains the possible values of the coordinates. In general, we model constraints in the following form:

$$\mathbf{z} = \mathbf{h}(\mathbf{x}) + \mathbf{v} \qquad [3]$$

z is the measured constraint (that is, the value provided by the experimental, theoretical or statistical source of information), and can be scalar or vector. It is modeled as having two parts: the first part is (in general) a vector function, **h(x)**, which is a function of the mean vector, **x**. The second part of the model, **v**, models the noise in the system. Given a perfect measurements, **v** is zero and the measured constraint takes on the exact value of the model function, **h(x)**. In general, **v**, is a Gaussian noise term that models the degree of certainty in any given measurement.

Thus, for example, a data source with information about the distance between two points would be represented as a function of six elements of the mean vector, **x**:

$$\mathbf{z} = \sqrt{(xi - xj)^2 + (yi - yj)^2 + (zi - zj)^2} + \mathbf{v} \qquad [4]$$

If the distance measurement refers to the distance between two carbon atoms in a covalent bond, then the variation in **v** is extremely small (the covalent bond distance varies less than 0.1 Ångstroms). If the distance measurement refers to an experimental measurement from, for example, NMR, then **v** will have larger variation (NMR distances vary by as much as 5 Å) [2]. For the purposes of this paper, we shall use distance and variance estimates derived from our statistical analyses of model RNA molecules and the correlation information collected by Klingler and Brutlag. We have shown elsewhere [15, 27] that the model is general and accommodates many forms of the function, **h(x)**, including bond angles (nonlinear functions of nine coordinates), and dihedral angles (nonlinear functions of twelve coordinates).

Given a set of probabilistic measurements (such as the distance constraints constructed earlier) and a starting model (such as randomly placed point with high variance an zero covariance), our algorithm calculates the vector of mean positions that best satisfies all the constraints. The satisfaction, or error, **e**, of a constraint is expressed as:

$$\mathbf{e} = \frac{\mathbf{z} - \mathbf{h}(\mathbf{x})}{\sqrt{\mathbf{v}}} \qquad [5]$$

Since the error is normalized by variance, **v**, the algorithm is able to use a uniform criterion to determine whether a constraint is satisfied—without reference to the absolute value of the constraint. This algorithm has been tested extensively and has been applied to problems in NMR data analysis [15, 27]. It has been shown to have good convergence properties, and to reliably find structures with a low average and maximum error [28]. It is therefore well-suited to the task of finding the range of structures compatible with the statistical constraints on tRNA structure that we have extracted.

For the tRNA calculations, each of the 76 GNs were initially placed *at random* with coordinates varying between 0 and 60 Å. The initial variances of each atom were set to 5000 $Å^2$ to approximate the volume of the tRNA molecule (the maximum dimension of the molecule is approximately 70 Å. If we set this to be roughly 1 standard deviation (SD) for the initial uncertainty in each base position, then the variance, SD^2 , for the base will be roughly 5000 $Å^2$). The standard parameters were used for running the DIKF as described in [24]. There were a total of 432 distance constraints, including all the covalent distances and constraints described here.

Results

The initial error for the randomly placed coordinates had an average value of 72 SD with a maximum of 220 SD. The DIKF converged to a stable optimum with an average error of 0.3 SD and a maximum error 1.1 SD. Figure 1 shows a histogram of errors for the constraints at the end of the calculation.

The topology of the tRNA structure as calculated by our procedure is shown in Figure 2. We compared the structure with the available crystal structures [18, 23], by extracting the location of the GNs of the crystal structure. The molecule has an overall dimension of roughly 75Å. The average global RMSD between the crystal structures and our calculated structures ranges from 9.5 Å to 10.1 Å. The global RMSD between different crystal structures of yeast phenylalanine tRNA, and between these and yeast aspartate tRNA, ranges from 1.5 to 4.0 Å. A breakdown of the segment-by-segment RMSDs is given in Table 1. In general, the four double helical elements are well defined by the constraints that were provided. In addition,

extended regions of up to 30 bases match the crystal structure to within 5 Å. The crystal structure itself satisfies the constraints with an average error of 0.9 SD and a maximum error of 5.9 SD. The main area of ill definition, accounting for most of the large deviations, is the region near the L-shaped hinge (residues 7 through 21, and 50 through 60). The RMSD between crystal and calculated structure is large in this region. The significance of and reasons for this discrepancy are discussed below.

The uncertainty in atomic position as calculated by our method is quite variable, as is shown in Figure 3. There are four low-variance regions where the double-helical elements are located. There is a large variation in the connector regions, especially around bases 30 and 70.
The correlations between bases, as represented in the covariance matrix, are shown in Figure 4. The diagonal shows the variance of each individual base position, whereas the off-diagonal regions indicate covariation. Lack of data about the spatial relationship of two bases corresponds to low covariance. As expected, there are regions of high covariation near the diagonal for each of the four helices. In addition, there is large covariation between regions around base 19 and base 55. On the other hand, the regions around base 35 and 52 show relatively little covariation. The relative paucity of data,

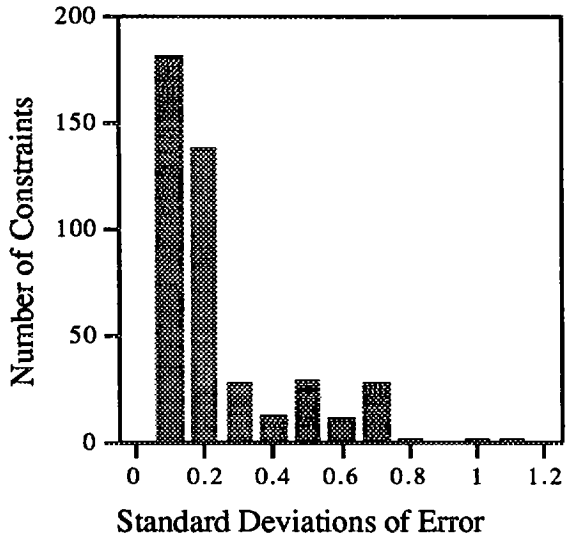

Figure 1. The frequency of errors (as defined in Equation 5 within the text) of the 432 constraints supplied to the program are plotted against the magnitude of the error. All constraints are satisfied to less than 1.2 SD.

and high variances, in the region around bases 50 through 60 may explain the relative disorder and high uncertainty in the hinge region.

Discussion

In an overall comparison, our structure matched the crystals to about 10 Å RMSD. Protein structures with similar resolution are considered very low-resolution structures. Nevertheless, we believe that our result is a quite useful (in particular, as starting point for future refinements) for four reasons. First, the dimensions of tRNA molecules (80 Å in some directions) are generally larger than are the dimensions for the globular proteins that are present in the Protein Data Bank, and so the error relative to size is compatible with medium resolution protein structures. Second, there is considerable variation in the segment-by-segment RMSD (Table 1) and the relatively large deviation is due primarily to the relatively disordered segments (residues 7 through 21, and 50 through 60) for which there are few data. Some regions match the crystal quite well. Third, we are representing with a single sphere all the atoms of each base, as well as the backbone phosphate and ribose. The maximum dimension of these components at a full atomic level of detail is approximately 12 Å. Thus, if we consider a lattice built to contain a single base at each point, our resolution is approximately equal to the grain size of this lattice. We are able, on average, to place the base in the correct lattice point. Finally, there is evidence from molecular dynamics simulations of tRNA that the angle of the L-hinge is somewhat arbitrary in the crystal structures. Tung and coworkers have shown that the angle can vary by as much as 20 degrees [29]. Close examination of the hinge region in the crystal structures confirms that there is a fair amount of unconstrained structure (in the sense that there are fewer base pairs, and base stacking interactions that were picked up by the correlation studies). We therefore believe that, given our coarse representation, the results reported here are reasonably good. The next step in our work is to use multiatom representations for the structure to decrease the variance in certain of our constraints (especially the base-stacking interactions), and to increase the sensitivity of the van der Waals packing constraints on structure.

We have run the algorithm 10 times (with different random starting structures) to check for consistent convergence. Interestingly, we observe that the algorithm always converges on structures that are roughly 10 Å RMSD from the crystal structure and from each other. In addition, we note that the crystal structure of tRNA satisfies the constraints roughly as well as our calculated structure. The uncertainty estimates shown in Figure 2 imply three-dimensional errors in the atomic location that include the crystal structure (within 1 to 2 standard deviations). We conclude that, given our representation of the objects and the constraints, there is an region in the

hyperspace with radius of approximately 10 Å that defines the region of valid solutions. We expect that the radius of such a space will decrease as we increase the precision of the representation (and therefore of the constraints).

Conclusions

In this paper, we have demonstrated the applicability of probabilistic algorithms for generating the topology of tRNA molecules. The constraints used were based on a physical interpretation of statistical correlations. Our algorithm is novel in that it simultaneously calculates estimates of the structure and the uncertainty in the structure. As a first step toward calculating a tRNA structure from statistical data, we have used a simplified representation and a physical interpretation of statistical correlations. We draw four conclusions:

1. Our current probabilistic implementation (using the DIKF) is useful for structure determination in cases of sparse or high-noise data. It is able to combine data sources with different uncertainties using the laws of probability for resolving potential inconsistencies.

2. In the case of tRNA, it is possible to use statistical correlations, knowledge of basic nucleic-acid helical geometry, and basic chemical constraints to calculate a good starting three-dimensional topology. The RMSD match of this topology on a segment-by-segment basis ranges from 1.5 Å to 10.0 Å. The regions of higher deviation correspond to regions of greater structural uncertainty. Most importantly, the shape of the molecule is clearly defined, as seen in Figure 3.

3. A single-atom abstraction of an entire nucleotide is an information-losing abstraction, but provides a great simplification of the conformational search space. We have found that the glycosylic nitrogen is preferable to the center of mass of the purine or pyrimidine ring. Thebenefits of the reduced number of degrees of freedom within the molecule come at the cost of reduced accuracy—especially with respect to the ability to check

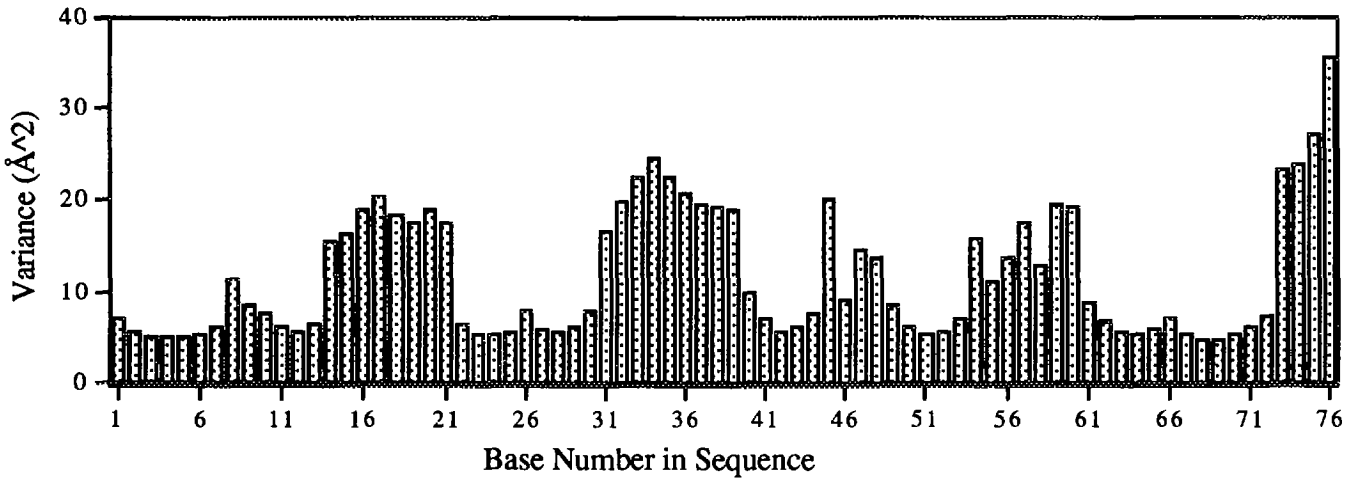

Figure 2. The calculated positional variance of each base in the sequence of the tRNA. The variance is lowest in regions where there are runs of base pairs, implying a helix, and constraining the possible positions for each base. The region at the end of the strand, and in the neighborhood of bases 15 to 21 and 31 to 39 are the most disordered, implying a relative lack of data about their position. This is shown graphically in Figure 3.

Segments Compared	RMSD (Å)
All Bases	10.1
1-6 and 72-66	1.7
10-13 and 25-22	2.8
26-30 and 44-40	1.3
49-53 and 65-61	1.3
49-65	5.5
26-44	3.4

Table 1: The RMS deviations between segments of the calculated tRNA structure and a representative structure (4TNA) from the PDB. Although the global RMSD is large, there are regions of high local agreement.

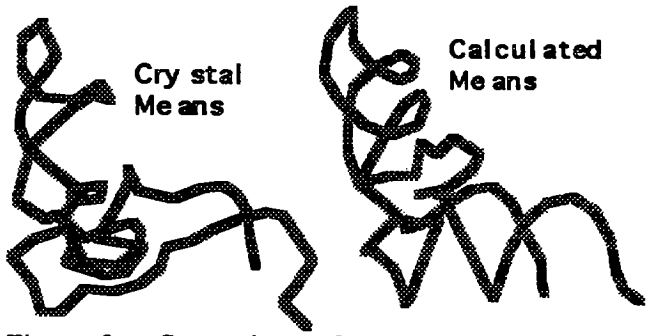

Figure 3. Comparison of the glycosylic nitrogen backbones of the crystal structure of yeast PHE-tRNA (left) with the structure calculated by assigning physical meanings to statistical correlations seen in the set of aligned sequences (right.) The two ends of the tRNA strand (bases 1 and 76) are in the lower right corner of each molecule, with base 76 being the furthermost. The anticodon loop is at the top of the molecule.

accurately van der Waals steric interactions and base-stacking interactions.

4. The probabilistic structure for tRNA indicates the greatest certainty in the conformation of double helices (as expected), with considerable uncertainty about the precise angle of the hinge separating the two major domains. These observations are consistent with molecular-dynamics simulations of the molecule.

Acknowledgments

This work was supported the Stanford University CAMIS project, which is funded under grant number LM05305 from the National Institutes of Health.. RBA received a hardware grant from Hewlett-Packard. Tod Klingler and Doug Brutlag provided valuable comments during this work. Steve Ludtke provided graphical software.

References

1. Blundell, T.L. and L.N. Johnson, *Protein Crystallography*. 1976, New York: Academic Press.

2. Wuthrich, K., *NMR of Proteins and Nucleic Acids*. 1986, John Wiley and Sons.

3. Lim, V.I., Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins. *J. Mol. Biol.*, 1974. **88**: p. 873-894.

4. Chou, P.Y. and G.D. Fasman, Conformational parameters for amino acids in helical, beta-sheet and random coil regions calculated from proteins. *Biochemistry*, 1974. **13**: p. 211-222.

5. Bowie, J.U., R. Luthy, and D. Eisenberg, A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. *Science*, 1991. **253**(July 12): p. 164-170.

6. Nakai, K., A. Kidera, and M. Kanehisa, Cluster analysis of amino acid indices for prediction of protein structure and function. *Prot. Engineering*, 1988. **2**(2): p. 93-100.

7. Zvelebil, M.J.J.M. and M.J.E. Sternberg, Analysis and Predicition of the Location of Catalytic Residues in Enzymes. *Prot. Engineering*, 1988. **2**(2): p. 127-138.

8. McGregor, M.J., S.A. Islam, and M.J.E. Sternberg, Analysis of the Relationship Between Side-chain Conformation and Secondary Structure in Globular Proteins. *J. Mol. Biol.*, 1987. **198**: p. 295-310.

9. Nemethy, G. and H.A. Scheraga, Theoretical studies of protein conformation by means of energy computations. *FASEB J.*, 1990. **4**(November): p. 3189-3197.

10. Levitt, M. and R. Sharon, Accurate simulation of protein dynamics in solution. *Proc. Natl. Acad. Sci. USA*, 1988. **85**: p. 7557-7561.

11. Havel, T.F., I.D. Kuntz, and G.M. Crippen, The Theory and Practice of Distance Geometry. *Bulletin of Mathematical Biology*, 1983. **45**(5): p. 665-720.

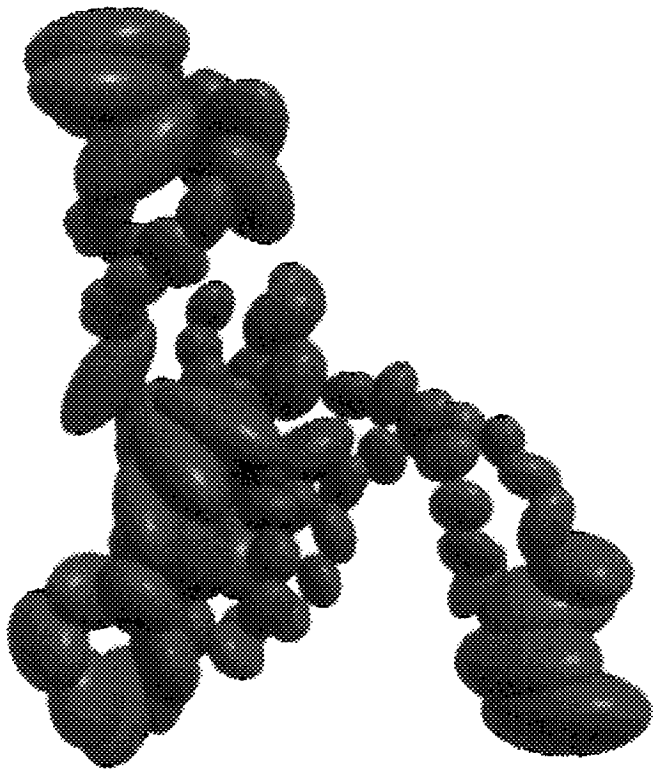

Figure 4. The uncertainty ellipsoids (drawn at 2 standard deviation contour) for the glycosylic nitrogens as calculated by the algorithm. The loop regions have more uncertainty than the regions within double helices.

12. Havel, T. and K. Wuthrich, A Distance Geometry Program for Determining the Structures of Small Proteins and Other Macromolecules from Nuclear Magnetic Resonance Measurements of Intramolecular ^{1}H - ^{1}H Proximities in Solution. *Bulletin of Mathematical Biology*, 1984. **46**(4): p. 673-698.

13. Nilges, M., A.M. Gronenborn, A.T. Brunger, and G.M. Clore, Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. *Prot. Engineering*, 1988. **2**(1): p. 27-38.

14. Gronenborn, A.M. and G.M. Clore, Analysis of the Relative Contributions of the Nuclear Overhauser Interproton Distance Restraints and the Empirical Energy Function in the Calculation of Oligonucleotide Structures Using Restrained Molecular Dynamics. *Biochemistry*, 1989. **28**: p. 5978-5984.

15. Arrowsmith, C., R. Pachter, R. Altman, and O. Jardetzky, The Solution Structures of E. coli trp Repressor and trp Aporepressor at an Intermediate Resolution. *Eur. J. Biochem.*, 1991. **202**: p. 53-66.

16. Gutell, R.R., A. Power, G.Z. Hertz, E.J. Putz, and G.D. Stormo, Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nuc. Acids. Res.*, 1992. **20**(21): p. 5785-95.

17. Klingler, T. M. and D. Brutlag, Detection of Correlations in tRNA Sequences with Structural Implication, *Proceedings of First International Conference on Intelligent Systems in Molecular Biology*, Washington, D.C., 1993. In this volume.

18. Hingerty, B.E., R.S. Brown, and A. Jack, Further Refinement of the Structure of Yeast t-RNA-Phe. *J. Mol. Biol.*, 1978. **124**: p. 523.

19. Westhof, E., P. Dumas, and D. Moras, Restrained Refinement of Two Crystalline Forms of Yeast Aspartic Acid and Phenylalanine Transfer RNA Crystals. *Acta Crystallographica, Section A*, 1988. **44**: p. 112.

20. Hoogsteen, K., Acta Crysallographica, 1959, **12**: p. 822-823.

21. Friedrichs, M.S., R.A. Goldstein, and P.G. Wolynes, Generalized Protein Tertiary Structure Recognition using Associative Memory Hamiltonians. *J. Mol. Biol.*, 1991. **222**: p. 1013-1034.

22. Altman, R. and O. Jardetzky, The Heuristic Refinement Method for the Determination of the Solution Structure of Proteins from NMR Data, in *Nuclear Magnetic Resonance, Part B: Structure and Mechanisms*, N.J. Oppenheimer and T.L. James, Editor. 1989, Academic Press: New York. p. 177-218.

23. Dock-Bregeon, A.C., B. Chevrier, A. Podjarny, J. Johnson, J.S. De Bear, G.R. Gough, P.T. Gilham, and D. Moras, Crystallographic Structure of an RNA Helix. *J. Mol. Biol.*, 1989. **209**: p. 459-469.

24. Altman, R.B., R. Pachter, E.A. Carrera, and O. Jardetzky, PROTEAN-Part II: Molecular Structure Determination from Uncertain Data (Program 596). *QCPE Bull.*, 1990. **10**(4).

25. Altman, R.B., A Probabilistic Algorithm for Calculating Structure: Borrowing from Simulated Annealing. *Uncertainty in Artificial Intelligence: Proceedings of the Ninth Conference*. Washington, D.C., 1993. Eds. Heckerman, Mamdani, Wellman. Morgan Kaufmann Publisher, *in press*.

26. Rumelhart, D.E. and J.L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Computational Models of Cognition and Perception, ed. J.A. Feldman, P.J. Hayes, andD.E. Rumelhart. Vol. 1. 1986, Cambridge, Massachusetts: MIT Press.

27. Liu, Y., D. Zhao, R. Altman, and O. Jardetzky, A Systematic Comparison of Three Structure Determination Methods from NMR Data: Dependence upon Quality and Quantity of Data. *Journal of Biomolecular NMR*, 1992. **2**: p. 373-388.

28. Pachter, R., R.B. Altman, and O. Jardetzky, The Dependence of a Protein Solution Structure on the Quality of the Input NMR data. Application of the Double-Iterated Kalman Filter Technique to Oxytocin. *J. Mag. Res.*, 1990. **89**: p. 578-584.

29. Tung, C.-S., S.C. Harvey, and A.J. McCammon, Large-Amplitude Bending Motions in Phenylalanine Transfer RNA. *Biopolymers*, 1984. **23**: p. 2173-2193.

30. Major, F., Turcott, M., Gautheret, D., Lapalme, G., Fillion, E., and Cedergren, R. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science;* 1991. **253**, p. 1255-1260.

31. Michel, F., and Westhof, E., Modelling of the three-dimensional architecture of group I catalygic introns based on comparative sequence analysis. J. Mol. Biol., 1989. **206**, p. 585-610.

32. Gautheret, D. and Cedergren, R., Modeling the three-dimensional structure of RNA, *FASEB J.*, 1993.7(1): p. 97-105.