Int. J. Human–Computer Studies (1995) **42,** 593–616

A probabilistic approach to determining biological structure: integrating uncertain data sources

RUSS B. ALTMAN

Section on Medical Informatics, MSOB X215, Stanford University Medical Center, Stanford, CA 94305-5479, USA

Modeling the structure of biological molecules is critical for understanding how these structures perform their function, and for designing compounds to modify or enhance this function (for medicinal or industrial purposes). The determination of molecular structure involves defining three-dimensional positions for each of the constituent atoms using a variety of experimental, theoretical and empirical data sources. Unfortunately, each of these data sources can be noisy or not available in sufficient abundance to determine the precise position of each atom. Instead, some atomic positions are precisely defined by the data, and others are poorly defined. An understanding of structural uncertainty is critical for properly interpreting structural models. We have developed a Bayesian approach for determining the coordinates of atoms in a three-dimensional space. Our algorithm takes as input a set of probabilistic constraints on the coordinates of the atoms, and an *a priori* distribution for each atom location. The output is a maximum *a posteriori* (MAP) estimate of the location of each atom. We introduce constraints as updates to the prior distributions. In this paper, we describe the algorithm and show its performance on three data sets. The first data set is synthetic and illustrates the convergence properties of the method. The other data sets comprise real biological data for a protein (the trp repressor molecule) and a nucleic acid (the transfer RNA fold). Finally, we describe how we have begun to extend the algorithm to make it suitable for non-Gaussian constraints.

1. Molecular structure

The determination of molecular structure is critical for many pursuits in biomedicine and industry, including the study of how molecules perform their function and the design of drugs to remove, modify or enhance this function. It is estimated that there are about 100 000 different proteins in the human body, but only a few hundred structures are known and stored in the protein structural data bank (Bernstein, Koetzle, Williams, Meyer, Brice, Rodgers, Kennard, Shimanouchi, & Tasumi, 1977). As the human genome project produces large amounts of information about the atomic makeup of individual molecules, it becomes critical to devise methods for estimating molecular structure—that is, for determining how the atoms within molecules arrange themselves in order to form three-dimensional structures.

Biological macromolecules can be divided into proteins and nucleic acids (Stryer, 1988). Nucleic acids, such as DNA and RNA, encode the genetic blueprints for all living organisms as a linear sequence of four chemical building blocks. Although the structure of nucleic acids was once thought to be uniform and geared only towards compact storage of information, it has become clear that the three-dimensional

1071-5819/95/060593 + 24$08.00/0

structures of these molecules are varied and able to carry out many important functions. Proteins, on the other hand, have long been recognized as the major effectors of function, including signal transduction, locomotion, chemical catalysis, and control of transport across membranes. Macromolecules normally have in the order of 1000–10 000 atoms, and so we must estimate 3000–30 000 coordinates to define a structure. The primary source for structural information has been experimental techniques of X-ray crystallography (Blundell & Johnson, 1976), and more recently, nuclear magnetic resonance (NMR) (Wuthrich, 1986). X-ray crystallography has limited applicability because not all proteins can be crystallized. NMR spectroscopy has technical limitations on the size of proteins that can be studied, and produces data that is somewhat uncertain.

Very often, therefore, structures must be computed with information gathered from multiple sources: experimental, theoretical and empirical/statistical observations. These data provide structural information ranging from geometric distances and angles to global measures of volume, shape and proximity to the surface. The problem of defining a structure from insufficient and noisy constraints is often underdetermined and leads to multiple solutions. It is therefore important to develop methods for combining evidence about structure that can represent the uncertainty explicitly. Moreover, it is critical that such methods produce not merely a single reasonable candidate structure, but also give some idea of the certainty associated with a position of each atom. Although there have been a few efforts to determine structure from *combinations* of experimental, statistical and theoretical data (Crippen & Havel, 1990; Sippl, 1990; Friedrichs, Goldstein, & Wolynes, 1991), not one of these methods is explicitly probabilistic, and the reliability of the solution is sometimes hard to gauge.

We have developed an algorithm that can take a wide range of probabilistic constraints on structure and produce estimates of the mean and three-dimensional variance in the position of each atom (Altman, 1989). The principle advantage of our approach is that data from disparate sources can be combined using the common language of probabilities—either determined objectively through statistical analysis, subjectively by expert estimation, or (most commonly) a combination of both. The goal of this paper is twofold: (1) to describe the methodology, and (2) to show its performance on three different data sets. The ideas used in our work should be useful in a variety of settings where probabilistic algorithms are searching a large space. Our method can be viewed as a nonlinear Bayesian maximum *a posteriori* estimator.

There are two lines of research that are related to the work described here. The first is that of molecular structure determination. *Distance geometry*, is an algorithm which takes as input a set of distances between atoms within a molecule. It employs a clever eigenanalysis of a matrix derived from these distances to estimate the coordinates of the structure (Havel, Kuntz, & Crippen, 1983; Havel & Wuthrich, 1984). It takes as input the min/max boundaries on parameter values, and produces as output a single solution. To estimate the uncertainty in the structure, it is necessary to run the algorithm many times and collect statistics over the resulting population of structures. Some implementations of distance geometry have been shown to sample space in a biased, non-systematic manner (Metzler, Hare, & Pardi, 1989). Distance geometry is prone to local minima, does not have well defined behavior for non-exact distances, and is limited to distance data only. *Energy minimization* and *molecular dynamics* are algorithms for structure determination which are based on the assumption that the proper conformation of a molecule is the one that has the lowest free energy (Levitt & Sharon, 1988; Nemethy & Scheraga, 1990). Energy terms that describe the interactions between all pairs of atoms within a structure can be defined, and optimization methods can be applied to find the conformation of the structure that has the lowest energy. Uncertainty is represented within the energy profiles (which are related to probabilities by the Boltzmann relation). However, these algorithms are based on physical forces, and it is difficult to know how to combine them with statistical and empirical sources of data. Because these algorithms are prone to local minima, they are most commonly

[redacted]

local dynamics. A detailed comparison between our method and these other approaches (varying constraint abundance, precision of constraints, size of molecule) has been published (Liu, Zhao, Altman, & Jardetzky, 1992). The key advantage of our method is the natural representation of constraint information as probability

[redacted]

element of **x**, and off-diagonals that contain the covariances between the elements within **x**:

$$\mathbf{C}(\mathbf{x}) = \begin{pmatrix} \sigma^2_{x_1} & \sigma^2_{x_1y_1} & \cdots & \sigma^2_{x_1z_N} \\ \vdots & \sigma^2_{y_1} & \ddots & \vdots \\ \sigma^2_{z_Nx_1} & & \cdots & \sigma^2_{z_Nz_N} \end{pmatrix} \tag{2}$$

Because the coordinates can be logically grouped into triplets (representing the x, y, and z coordinates for a single atom), we can also consider **C**(**x**) to be a matrix with submatrices.

$$\mathbf{C}(\mathbf{x}) = \begin{pmatrix} C(x_1) & \cdots & C(x_1x_N) \\ \vdots & C(x_2) \quad \ddots & \vdots \\ C(x_Nx_1) & \cdots & C(x_N) \end{pmatrix}, \tag{3}$$

where each of the submatrices represents the variance of a single atom (diagonals), or the covariance between two atoms (off-diagonals).

$$\mathbf{C}(\mathbf{x}_i\mathbf{x}_j) = \begin{pmatrix} \sigma^2_{x_ix_j} & \sigma^2_{x_iy_j} & \sigma^2_{x_iz_j} \\ \sigma^2_{y_ix_j} & \sigma^2_{y_iy_j} & \sigma^2_{y_iz_j} \\ \sigma^2_{z_ix_j} & \sigma^2_{z_iy_j} & \sigma^2_{z_iz_j} \end{pmatrix} \tag{4}$$

Our representation allows us simultaneously to display information about molecular structure and uncertainty. The mean values for the coordinates of each atom can be taken from the vector, **x**, and plotted. In addition, the variance of each coordinate of an atom can be extracted from the diagonal and provides the uncertainty along each axis of the mean estimate. In fact, with the full 3 × 3 variance/covariance information, we can estimate the uncertainty in any direction. Figures 2 and 3(a) illustrate the mean positions and ellipsoidal uncertainties for two molecules. The ellipsoids are drawn at two standard deviations assuming a three-dimensional Gaussian distribution.†

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

† Given the covariance matrix, **C**, for an atom (as in Equation 4) we can compute the ellipsoid of uncertainty assuming a three-dimensional Gaussian in the following manner. We diagonalize $\mathbf{C} = R^T DR$, so that D contains the lengths of the principal axes of the ellipsoid (in units of variance), and R describes the rotation of the ellipsoid in the global coordinate system. If we want to draw an ellipsoid at N standard deviations, we calculate $N\sqrt{\sigma}$ for each of the diagonal elements of D, render an ellipsoid with these semiaxis lengths, rotate the ellipsoid with R, and translate to the mean position.

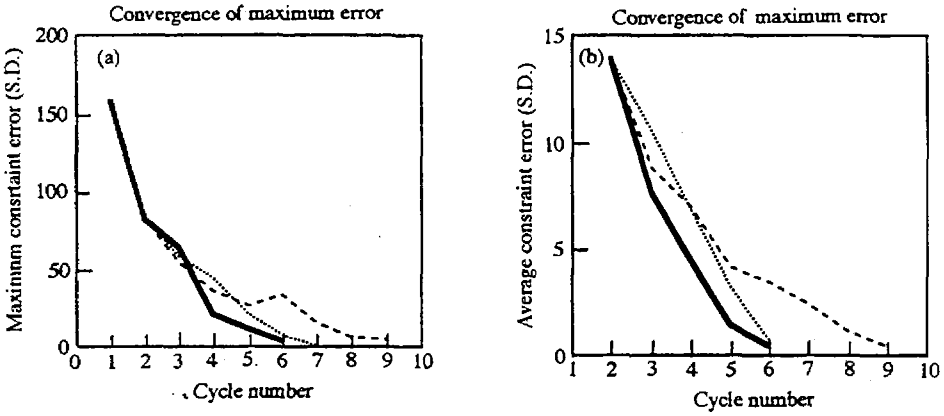


FIGURE 1. (a) Each of three strategies for ordering constraints is compared with respect to the *maximum* error of all constraints as a function of cycle number. *Sorted* constraints were introduced in reverse order of satisfaction at the "reorder" step of the algorithm as presented in the pseudocode summary. *Random* constraints were introduced in random order, and *fixed order* constraints were introduced in the same (arbitrary) order each cycle. This result shows that the sorting step is effective in helping the algorithm quickly find a good solution. Random ordering is also effective, probably because it allows different constraints to rearrange atoms early in each cycle. Fixed order is, as might be expected, less efficient because the same constraints are always used to start each cycle, and so the chance of escaping a minima is lower. Nevertheless, it is reassuring that all three methods do converge. (b) Each of three strategies for ordering constraints are compared with respect to *average* error of all constraints as a function of cycle number. Sorting strategies are the same as in Figure 1(a). Sorted constraints lead to more rapid convergence. (——): sorted; (- - -): random; (– – –): fixed order.

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; a change in any atom position affects the position of other atoms through this first order estimate of their covariation. Thus, the off-diagonal 3×3 submatrices represent a linear summary of how the position of one atom changes as the position of another is modified. There is a strong network aspect to this representation. As more is learned about the relationships between atoms, the network of dependencies grows (for example, see Figure 3(b)). 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 are discussed in the next section.

In practice, we must assign values to the $\mathbf{x}$ and $C(\mathbf{x})$ variables before the introduction of constraints. This represents our *prior* model of the structure. If we have no information about structure, then we can generate random coordinates for the mean positions, and generate an uncorrelated covariance matrix with diagonals that reflect uncertainty in the mean positions (based, for example, on the expected volume of the points in space), and with off-diagonals of zero. On the other hand, if we have information about the general shape of the structure, we may be able to assign reliable starting mean positions, as well as information about the variance at each of these positions. This approach is useful, for example, when modeling an unknown structure that is thought to be similar to a set of previously determined structures. These previously determined structures define the bounds within which the new structure must fall.

Thus, for example, a measurement of distance between two points would be represented as a function of six elements of the mean vector, $\mathbf{x}$:

$$\mathbf{z} = \sqrt{(x_i - x_j)^2 + (x_{i+1} - x_{j+1})^2 + (x_{i+2} - x_{j+2})^2} + \mathbf{v}. \tag{6}$$

If the distance measurement refers to the distance between two carbon atoms in a chemical bond, then the variation in $\mathbf{v}$ is extremely small (the covalent bond distance varies less than 0.1 Å). If the distance measurement refers to an experimental measurent from, for example, a study using NMR, then $\mathbf{v}$ will have larger variation (NMR distances vary as much as 5 Å) (Wuthrich, 1986). For many problems, distance constraints are the primary form of available structural information. We have shown elsewhere (Arrowsmith, Pachter, Altman, & Jardetzky, 1991; Liu *et al.*, 1992), however, that the constraint model (Equation 5) is general and extends to bond angles (a nonlinear function of nine coordinates), dihedral angles (a nonlinear function of 12 coordinates), and any other measurement that is a function only of the atomic coordinates within the vector $\mathbf{x}$.

2.2. INTRODUCING CONSTRAINTS TO UPDATE MODELS

Having established our representation for atomic position, atomic uncertainty, and constraints, we can understand the mechanism for introducing constraints and updating our estimates of the state vector, $\mathbf{x}$, and the covariance matrix, $\mathbf{C}(\mathbf{x})$. The standard Kalman filter employs a static measurement update algorithm of the following form (Gelb, 1984):†

$$\mathbf{x}(\text{new}) = \mathbf{x}(\text{old}) + \mathbf{K}[\mathbf{z} - \mathbf{h}(\mathbf{x}(\text{old}))] \tag{7}$$

$$\mathbf{C}(\text{new}) = \mathbf{C}(\text{old}) - \mathbf{KHC}(\text{old}) \tag{8}$$

where

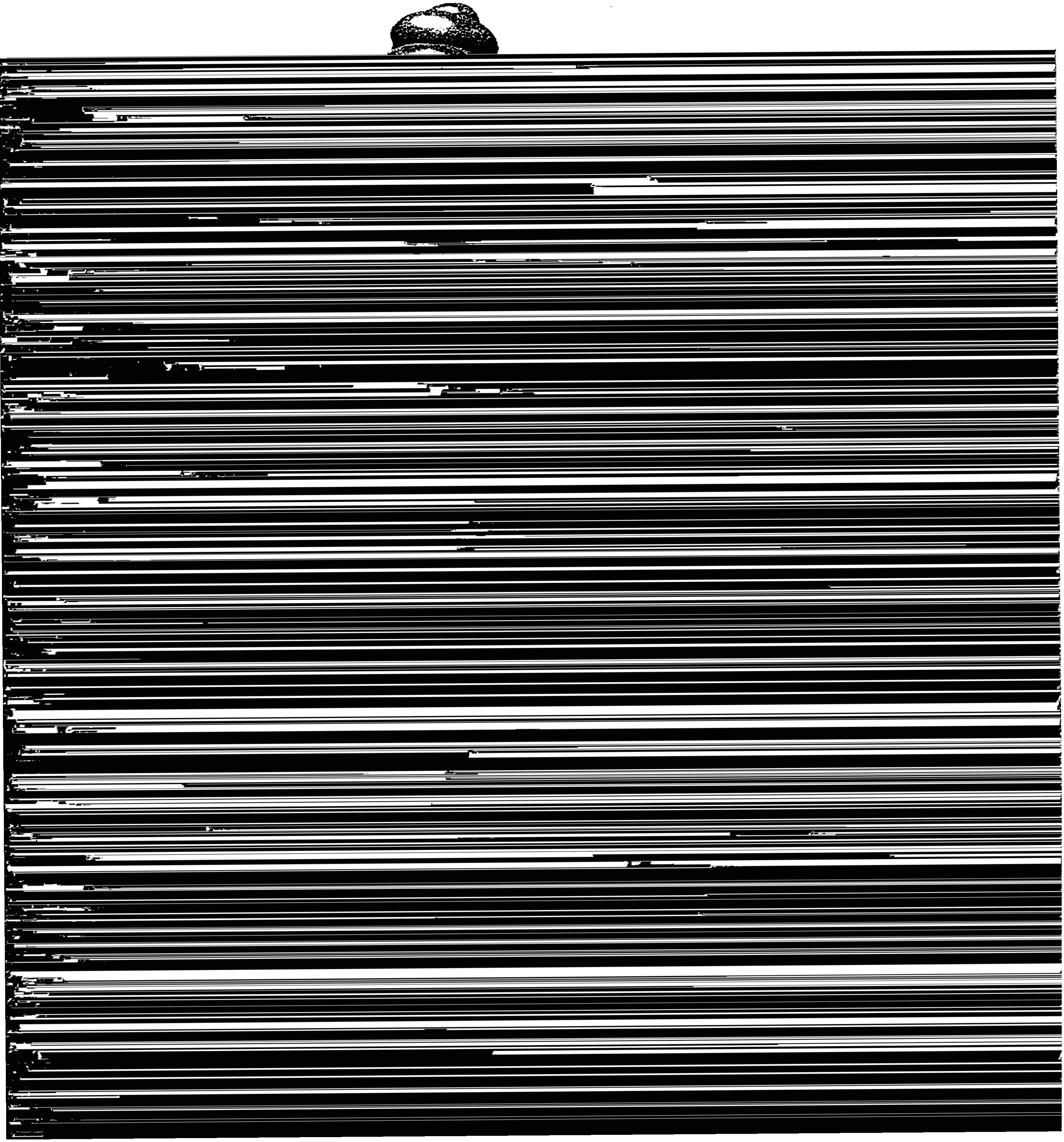
$$\mathbf{K} = \mathbf{C}(\text{old})\mathbf{H}^{\mathrm{T}}[\mathbf{HC}(\text{old})\mathbf{H}^{\mathrm{T}} + \mathbf{C}(\mathbf{v})]^{-1} \tag{9}$$

and

$$\mathbf{H} = \left.\frac{\delta \mathbf{h}(\mathbf{x})}{\delta \mathbf{x}}\right|_x \tag{10}$$

Simply stated, Equation 7 specifies that the new estimate of mean position ($\mathbf{x}$(new)) is based on the old estimate of mean position that is corrected by a weighted difference between the observed value of the measurement, $\mathbf{z}$, and the value that would be predicted from the old model, $\mathbf{h}(\mathbf{x}(\text{old}))$. Note that the matrix, $\mathbf{K}$, depends on the ratio of the uncertainty in the predicted constraint value (in the numerator, which depends on a linearized constraint value, $\mathbf{H}$, and the state vector

† In general, the Kalman filter allows for a time-dependent modeling of how $\mathbf{x}$ and $\mathbf{h}(\mathbf{x})$ change. We assume a static molecule and do not introduce any time-dependent model of change. We therefore are interested in calculating a single estimate that, for example, corresponds to a single point in time. In principle all constraints can be introduced simultaneously by creating a large vector of measurement values. However, this leads to the requirement for a large matrix inversion (as seen in Equation 9), since $\mathbf{v}$ becomes a vector and $\mathbf{C}(\mathbf{v})$, the variance of $\mathbf{v}$, becomes a matrix. We have shown elsewhere (Chen, Singh, Poland & Altman, 1994), that small groups of constraints can be introduced efficiently.



uncertainty, **C**), and the uncertainty in the measured value (in the denominator, a linear estimate of the uncertainty in z). If the measured data, z, have high variance able to respond to data by changing values. This forms the key intuition for our heating strategy discussed in Section 2.4.

2.3. EVALUATING THE MODELS

Given an estimate of mean structure, **x**, its variance, **C**(**x**), and a set of constraints of the form shown in Equation 5, we evaluate the quality of the estimate by comparing [redacted] estimate. We introduce "heat" by resetting the covariances to their initial (large) values (which should allow unsatisfied constraints to have a relatively greater effect on the vector, **x**). We then reintroduce all the constraints once again, but *sorted* such that the constraints that were least satisfied by the previous coordinate estimates are introduced into the solution earliest. We have shown, in experiments described in Section 3.1., that by reheating the covariances and introducing the constraints in reverse order of satisfaction, we maximize the chance that the atoms will be reorganized radically and will jump out of the current minima. Since we have observed a consistent ability of the update equations to improve upon the starting estimates, we simply repeat the cycle of search, reheat, search until the residual errors are acceptable (see outline of procedure). In essence, we are repeating the calculation with an improved *prior* distribution on the mean positions of all the points. We are still being conservative, however, because we use the same variance and not a variance that has been reduced to reflect the increment in information contained in **x**. Although we have not made any formal claims about resistance to minima, there are three forces acting to help the algorithm avoid (or leave) multiple minima. First, we use a covariance matrix to capture the first-order correlation between atomic coordinates; therefore, moving even a few atoms causes changes in the entire molecule (and more coverage of the search space). Second, the reheating allows atoms to move from one local minima to another in a rational way; atoms will current implementation, however, we handle the packing density in a less elegant way. After introducing all input constraints, we check all pairwise distances between atoms and identify pairs of atoms that are too close to one another. When such a pair is found, we introduce a new distance constraint to impose the minimum distance criterion. This new constraint is used only once to "push" the atoms away large protein structure, the trp repressor dimer, from a relatively sparse NMR data set. Third, we describe our use of the algorithm to compute the structure of a nucleic acid molecule, transfer RNA, using constraints derived solely from statistical analysis of sequence.

3.1. EFFICACY OF THE REHEATING STRATEGY: TESTS WITH SYNTHETIC DATA

To test the convergence properties of the method, we chose the problem of defining the topology of a small protein, crambin (Hendrickson & Teeter, 1981). Crambin contains roughly 500 atoms, but for the purpose of this example, we considered only the 46 backbone alpha carbon atoms that define the general topology of the molecule. The structure of crambin is known, so we generated synthetic data sets for these tests. In general, there are 1035 distances between 46 atoms. The minimum number of exact distances required to define the position of N points is $4N-10$, or 526 in the case of crambin.† The state (coordinate) vector, therefore, has 134 parameters and the covariance matrix is 134×134. For all calculations, the starting values for the $\mathbf{x}$ vector were generated randomly between 0 and 50 Ångstroms (an uninformed prior). The covariance matrix was initialized to have all diagonal elements at 100 (that is, a starting variance of 100 Å^2 for each atom, compatible with the expected volume of the molecule), and off-diagonal elements set to 0 (implying independence of all coordinates initially).‡ For all runs, the tolerance for exiting the inner loop of the iterated, extended Kalman filter was 0.01, and the maximum number of cycles, i, was three. The stopping condition for all runs (unless otherwise noted) was an average error for all constraints of 0.3 S.D. or a maximum error of 1.0 S.D. We performed three tests.

(1) We tested the algorithm by providing all possible exact distances (1035), with extremely low variance. The random starting structure had an average error (in S.D. from measured value) of 60, with a maximum error of 175. With all possible exact distances, the algorithm converged to an average error of 0.20 S.D. (maximum error 1.3 S.D.) in three cycles. To test the stability of the solution, we allowed the algorithm to run for a total of 1000 cycles. The solution remained stable, and the ultimate improvement to an average error of 0.0007 S.D. (maximum of 0.002) was achieved at cycle 58. Cycles 59 through 1000 made no further improvement. The structure that resulted was identical to the target solution, as expected.

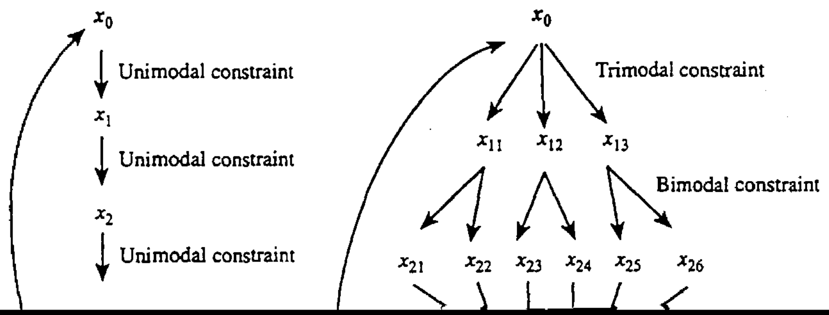
(2) To explore the dependence of convergence on the order of constraints, we

† The $4N-10$ figure is derived from the following argument: given four fixed points to describe a coordinate system, any additional point can be unambiguously positioned by providing the distances to the four fixed points (the distance to the first point provides a shell of possible locations, the distance to the second point provides a two-dimensional circle of possible locations, the distance to the third point selects two points on the circle and the fourth distance disambiguates between these two points). Additional points can then be positioned with four distances to any of the previously fixed ones. Thus, for all points after the first four, we require four distances or $4(N-4)=4N-16$. In order to position the first four points, we need only six distances: one point can be placed arbitrarily at the origin (no distances required), its distance to the second point enables us to place the second point on the x-axis (one distance). The distances of the first two points to the third allow us to place the third point on the positive xy-plane (two more distances). Finally, the fourth point can be positioned in the positive z-hemisphere using the distances to the first three points (three more distances). Thus, the total number of distances required is, minimally, $4N-16+6=4N-10$.

‡ As constraints between atoms are introduced and propagated, the off-diagonals of the covariance matrix become nonzero.

[redacted]

rest were from known chemical bond lengths and angles. A small number of bond angles and dihedral angle constraints were also introduced. After 10 cycles, the shown in Figure 3(b). The final structure satisfies all constraints to less than 1.2 S.D. The crystal structure of two tRNA molecules are known, and are quite similar to the structure produced by our algorithm (10 Å RMSD deviation, which is consistent with the granularity of our bead representation). More importantly, perhaps, the solution produced by our algorithm shows which parts of the molecule have relatively low variance, and are therefore predicted with high confidence. As in the case of the trp repressor, the information about variance can be interpreted


x_0
Unimodal constraint
x_1
Unimodal constraint
x_2
Unimodal constraint
x_0
Trimodal constraint
x_{11}
x_{12}
x_{13}
Bimodal constraint
x_{21}
x_{22}
x_{23}
x_{24}
x_{25}
x_{26}

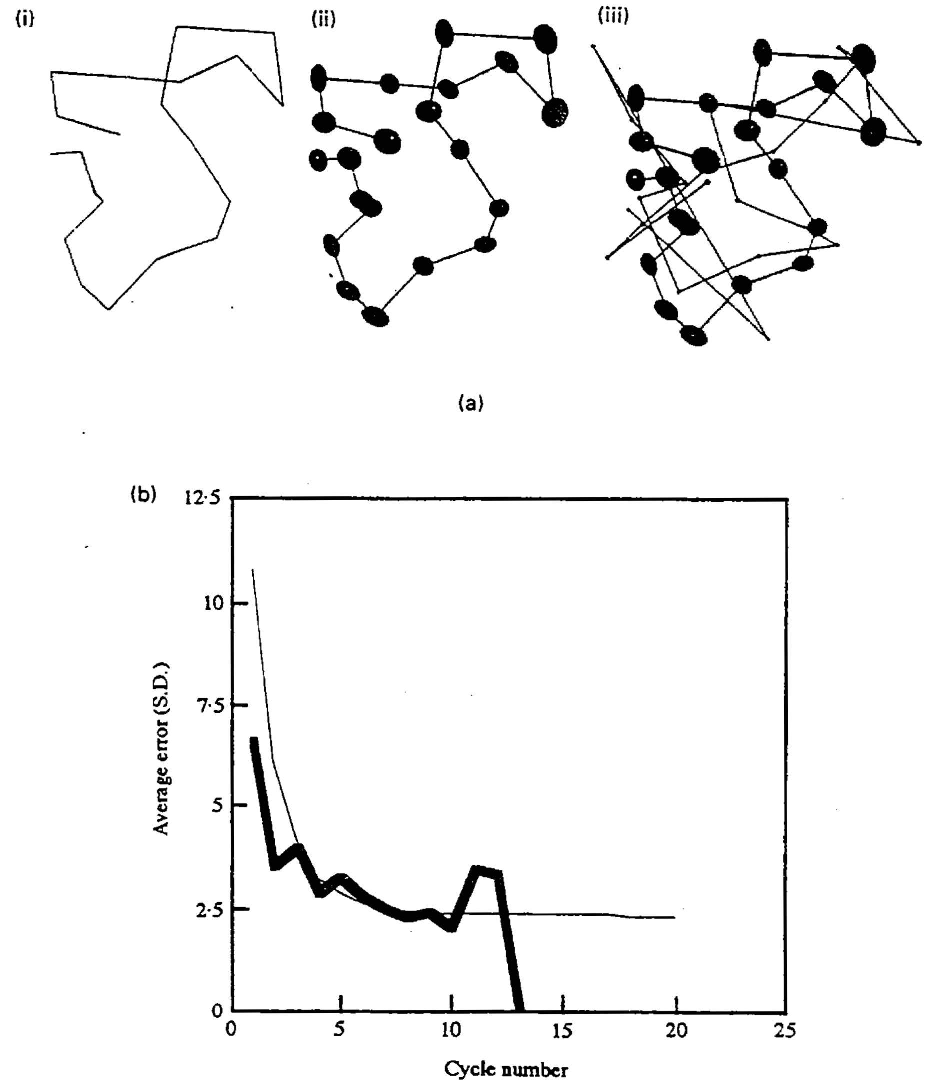


FIGURE 5(a). (i) The gold standard structure used for testing the multicomponent algorithm (a fragment of the molecule crambin). (ii) The multicomponent solution computed using a data set in which each of the actual components (the component corresponding to the actual distance in crambin) is contaminated with between one and three noise components with weights varying betwen 0.1 and 0.5. The multicomponent algorithm is able to identify the correct components and produce the correct structure.

most closely fall. If the errors are large, then the variance matrix is reheated, as described in Section 2.4 and the process is repeated.

The performance of this algorithm can be illustrated with two test calculations. In the first, we have generated a synthetic data set in which we have added to the "real" Gaussian component (describing the actual distance value in the test structure, crambin again) a number of "noise" components. We have shown, as detailed in Figure 5, that the algorithm is able to detect the real components and converge to the correct structure. In the second calculation, we took two molecules with the same number of atoms (structures A and B) labeled all the atoms uniquely, and then provided the program with two equally weighted Gaussian components for each constraint (one drawn from each structure). Thus, we removed all information about which components are associated with structure A and which are associated with structure B. The task of the algorithm was to find the sets of components that are drawn from the same structure, and produce a coherent, low error solution. For N constraints, there are 2^N possible combinations of modes. In our experiments, we capabilities of the Kalman filter. Our method is a member of the class of nonlinear least-squares estimators that seek the most likely set of coordinates that best satisfy the input constraints. It is therefore a (MAP) estimator. The method of posterior mode analysis proposed by Shachter solves a very similar problem, and may be a useful alternative to our method for the case of Gaussian noise (Shachter, Eddy, & Hasselblad, 1990). Our method is Bayesian because it uses an initial probabilistic model of the solution, and updates this model with data. It can be shown that if the prior probability distribution contains no information, then MAP methods are equivalent to least squares estimators (Mikhail, 1976). In these experiments the model had low information content, since it had random starting positions with large variance. Nonetheless, our knowledge of the bounds on the volume of the molecule provided our starting estimate of variance, which was the primary information contained in our prior. Our method uses a first-order approximation to the nonlinearities of the system, and improves its performance by iteration. The idea of combining uncertain data with a least squares criterion and computing explicit estimates of uncertainty dates back to the nineteenth century in early work in *geodesy* (measurements of distances and locations on the surface of the earth) (Bomford, 1960). These methods did not include prior models of parameter values, and solved nonlinear problems by finding (in an unspecified manner) an approximate solution, followed by linearization of nonlinear functions and refinement. They did not maintain covariation information during the refinement.

Our single component algorithm produces a two-moment estimate of atomic location (three-dimensional mean and variance). For purposes of display, we assume that these represent the first two moments of a three-dimensional Gaussian when drawing atomic locations. (Altman, Hughes, & Gerstein, 1995). Of course, it is possible that some atoms will have a bimodal distribution, and we can only capture these distributions with the multicomponent algorithm.† Moving to multimodal representations of atomic position is not a priority in our work for two reasons. First, as more independent data sources are introduced, the three-dimensional Gaussian becomes the most likely final distribution by the central limit theorem. Second, there are few biological examples of significant bimodal distributions. On the other hand, the use of multicomponent constraints is critical, and is the focus of current effort.

Because of the matrix multiplications required for the basic algorithm, it has a computational complexity of $O(N^3)$. However, many of these multiplications are sparse and can be optimized so that the algorithm is able to handle relatively large calculations, such as that of the trp repressor described in Section 3.2. In addition, we have recently reported an implementation of the method that takes advantage of massively parallel supercomputers (Chen *et al.,* 1994). The computational results with the trp repressor and tRNA not only demonstrate that the algorithm scales up to solve biological problems, but also illustrate the biological utility of having estimates of atomic positional uncertainty. In both cases, a key biological observation can be related to the pattern of atomic variation. We anticipate that many other activities, such as the design of drugs or the re-engineering of these molecules for

† If we suspect that there may be two modes for each atom, we can modify the recombine step to produce the best *two* structures instead of the best single structure as an intermediate step before testing more constraints.

other functions, will depend critically on such assessments of the reliability of a structural model.

5.1. EXTENDING THE SYSTEM WITH NEW CONSTRAINT TYPES

Although we have concentrated in this paper on the use of distance constraints between points, the mathematical form of the filter makes it clear that (1) any function of the coordinates can be used as a constraint model, and (2) these functions need not be scalar, but rather can be vector functions. In our applications work so far, we have limited ourselves to distances, angles, and dihedral angles because these types of constraints are sufficient for most structure determinations from NMR data. However, as we collect statistical data on the associations between certain types of atoms and aggregates of atoms, we can use statistical distributions as constraints on our molecule. Since these statistical distributions will not always be Gaussian, we have focused attention on processing non-Gaussian constraints. Our success with such constraints, described in Section 4, is preliminary. We have shown that the algorithm satisfies the necessary conditions of (1) choosing a single structure from a set of noise components and (2) identifying coherent sets of components. We are currently testing the multicomponent algorithm on real biological structures to assess its performance.

The form of equations 5–11 suggest that the measurement **z** can be vector-valued. In principle, we can use this machinery to introduce multiple constraints simultaneously in a single update. Until recently, we have avoided vector-valued measurements (and preferred the serial introduction of scalar constraints) in order to avoid the matrix inversion (required for nonscalar variables) in Equation 9. However, we have recently implemented this algorithm on a massively parallel system, and have shown that the algorithm runs best with the parallel introduction of 50 to 100 constraints (that is, **z** is a vector of 50 distance measurements, and **v** is a vector of their noise) (Chen *et al.,* 1994). With the strategy of introducing many constraints simultaneously for one update, a greater improvement in the solution occurs per cycle, but that the cost per cycle increases.

5.2. EFFICACY OF THE REHEATING AND RESORTING STRATEGIES

Our experiments with different constraint orders confirm our hypothesis that the reheating of the covariance matrix allows the solution space to be explored more effectively. A fixed order of constraints is more likely to explore the same general hypothesis space, and to converge more slowly than either a random order or an order in which the most dissatisfied constraints take the "first shot" at altering the solution. In fact, the trace of the fixed order convergence in Figure 1 shows that it is able to jump out of the local minima of cycle 5 during cycle 6, even without sorting. In this experiment, the sorted run does not seem to fall into local minima.

Simulated annealing is a computational method for assisting optimization by providing a powerful heuristic for efficient search (van Laarhoven *et al.,* 1987; Vanderbilt *et al.,* 1984). Based on an analogy to solid-state physics, simulated annealing protocols add "heat" to an optimization to increase the likelihood that a solution will jump out of a local optima. The solution is then allowed to "cool" slowly such that it settles into a new optimum—as a cooled solid might settle into a new crystalline packing. Our method shares many high level concepts with simulated annealing: by increasing the variances and covariances, we are increasing the range of possible values for each parameter, and by introducing the constraints in reverse order of satisfaction, we give the least satisfied constraints a chance to pull the solution out of a local minima. Although it is heuristic in nature, we have found that this protocol reliably finds low average error structures, as well as low maximum errors (Pachter, Altman, & Jardetzky, 1990; Liu *et al.,* 1992). Whereas simulated annealing protocols for global optimization require a cooling procedure to induce

RBA is supported by NIH grant LM05652, and is a Culpeper Foundation Medical Scholar. This work was supported by computing resources provided by the Stanford University

in dihedral angle space using NMR derived constraints: a new algorithm based on optimal filtering. *Journal of Molecular Biology*, **223**, 299–315.

van Laarhoven, P. J. M. & Aarts, E. H. L. (1987). *Simulated Annealing: Theory and Applications*. Dordrecht: Reidel.

Levitt, M. & Sharon, R. (1988). Accurate simulation of protein dynamics in solution. *Proceedings of the National Academy of Science, USA*, **85**, 7557–7561.

Liu, Y., Zhao, D., Altman, R. & Jardetzky, O. (1992). A systematic comparison of three structure determination methods from NMR data: dependence upon quality and quantity of data. *Journal of Biomolecular NMR*, **2**, 373–388.

Metzler, W. J., Hare, D. R. & Pardi, A. (1989). Limited sampling of conformational space by distance geometry algorithm: implications for structures generated from NMR data. *Biochemistry*, **28**, 7045–7052.

Mikhail, E. M. (1976). *Observations and Least Squares*. New York, NY: Dun-Donnelley.

Nemethy, G. & Scheraga, H. A. (1990). Theoretical studies of protein conformation by means of energy computations. *FASEB Journal*, **4**, 3189–3197.

Pachter, R., Altman, R. B., Czaplicki, J. & Jardetzky, O. (1991). Comparison of the NMR solution structure of cyclosporin A determined by different techniques. *Journal of Magnetic Resonance*, **92**, 468–479.

Pachter, R., Altman, R. B. & Jardetzky, O. (1990). 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. *Journal of Magnetic Resonance*, **89**, 578–584.

Poland, W. B. & Shachter, R. D. (1993). Mixtures of Gaussians and minimum relative entropy techniques for modeling continuous uncertainties. In D. Heckerman & A. Mamdani Eds. *Proceedings of Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 183–190. San Mateo, CA: Morgan Kaufmann.

Shachter, R. D., Eddy, D. M. & Hasselblad, V. (1990). An influence diagram approach to medical technology assessment. In R. M. Oliver & J. Q. Smith, Eds. *Influence Diagrams, Belief Nets and Decision Analysis*, pp. 321–350. New York, NY: John Wiley.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology*, **213**, 859–883.

Smith, R., Self, M. & Cheeseman, P. (1986). Estimating uncertain spatial relationships in robotics. In *Proceedings of the Uncertainty in Artificial Intelligence Workshop*, pp. 1–21. Philadelphia, PA: AAAI Press.

Stryer, L. (1988). *Biochemistry* (3rd ed.). New York, NY: W. H. Freeman.

Vanderbilt, D. & Louie, S. G. (1984). A Monte Carlo simulated annealing approach to optimization over continuous variables. *Journal of Computational Physics*, **56**, 259–271.

Wuthrich, K. (1986). *NMR of Proteins and Nucleic Acids*. New York, NY: John Wiley.

Zhao, D. & Jardetzky, O. (1993). Refined solution structures of the *Escherichia coli* trp holo- and aporepressor. *Journal of Molecular Biology*, **229**, 735–746.