A Surface Measure for Probabilistic Structural Computations

Jeanette P. Schmidt†, Cheng C. Chen^, Jonathan L. Cooper^, Russ B. Altman^*

†Polytechnic University & Incyte Pharmaceuticals
^Section on Medical Informatics, Stanford Universty
jschmidt@incyte.com, cchen@smi.stanford.edu, rba@smi.stanford.edu
*Corresponding Author

Abstract

Computing three-dimensional structures from sparse experimental constraints requires methods for combining heterogeneous sources of information, such as distances, angles, and measures of total volume, shape, and surface. For some types of information, such as distances between atoms, numerous methods are available for computing structures that satisfy the provided constraints. It is more difficult, however, to use information about the degree to which an atom is on the surface or buried as a useful constraint during structure computations. Surface measures have been used as accept/reject criteria for previously computed structures, but this is not an efficient strategy. In this paper, we investigate the efficacy of applying a surface measure in the computation of molecular structure, using a method of probabilistic least square computations which facilitates the introduction of multiple, noisy, heterogeneous data sources. For this purpose, we introduce a simple purely geometrical measure of surface proximity called *maximal conic view* (MCV). MCV is efficiently computable and differentiable, and is hence well suited to driving a structural optimization method based, in part, on surface data. As an initial validation, we show that MCV correlates well with known measures for total exposed surface area. We use this measure in our experiments to show that information about surface proximity (derived from theory or experiment, for example) can be added to a set of distance measurements to increase significantly the quality of the computed structure. In particular, when 30 to 50 percent of all possible short-range distances are provided, the addition of surface information improves the quality of the computed structure (as measured by RMS fit) by as much as 80 percent. Our results demonstrate that knowledge of which atoms are on the surface and which are buried can be used as a powerful constraint in estimating molecular structure.

Introduction

The primary means for determining molecular structure remains high resolution X-ray crystallography and nuclear magnetic resonance (Blundell& Johnson, 1976; Markley & Opella, 1997). However, some molecules are difficult to study with these techniques, and so structural information must be gathered using a variety of experimental means. Distance information is obtained from chemical cross-linking (Harris et al, 1994), enzymatic and chemical protection experiments (Powers & Noller, 1995), and fluorescence energy transfer. Volume and shape information can be obtained from small angle scattering (Glatter, 1979), and surface/buried information can be obtained from solvent accessibility (Lane & Jardetzky, 1985) or sensitivity to chemical probes (Moazed et al, 1986).

Methods for handling distance information are the most mature, including distance geometry algorithms (Crippen & Havel, 1988), restrained molecular dynamics methods (Nilges et al, 1988), and our method of probabilistic least squares estimation (Altman, 1995; Chen et al, 1996). The representation of non-distance information is more problematic, however. Some experimental techniques yield information that a particular atom is near the surface of a molecule, or is buried and relatively inaccessible. The representation of surface/buried information cannot unambiguously be translated into a set of distances, and so distance-based algorithms need to have additional pre- or post-processing modules to handle these constraints. Our method of probabilistic least squares was designed specifically to allow the introduction of multiple, noisy, heterogeneous data sources. The method uses probability theory to combine sources of evidence with different degrees of reliability. The main requirements of the probabilistic least squares estimation technique are that (1) each constraint must be represented as a deterministic function of the coordinates of the structure being computed (plus an additive noise term as described in next section), and (2) the function should be differentiable, so that the algorithm can conduct a search of conformational space based on the gradient.

In this paper, we investigate the addition of a measure of surface proximity in our computation of molecular structure. While numerous surface measures have been defined, none seemed ideally suited for our purpose. The commonly used measure of solvent accessibility of individual atoms (Connolly 1985, Kabsch and Sanders 1983) provides a sensitive characterization of atoms on the molecular surface, but was not designed to differentiate between buried and deeply buried atoms. All buried atoms have a solvent accessibility of zero. In addition, since

infinitesimal movement of any of the atoms does not change the solvent inaccessibility of the buried atom, the derivative of such a function is zero as well, and a gradient based method would gain no information on how to move such atom towards the surface of the molecule. Other measures focus on the total surface area, a critical parameter for computing solvation energy; these include Arteca et al 1988, F. Eisenhaber, P. Argos 1993 G. Perrot et al 1992, Sridharan et al 1994 and Eisenhaber et al 1995, and an extensive overview of these by Connolly 1996, to mention just a few. These measures do not provide a surface measure for each individual atom.

We introduce a new measure of surface proximity called the *maximal conic view* (MCV, illustrated in Figure 1). The measure, as we will show, is efficiently computable, continuous and differentiable in the interval [-1,1], and well suited for our computations. The surface proximity of a point is measured by the degree to which a point has an unobstructed view beyond the borders of the molecule, looking out into the surrounding media. Formally, for any atom X, we determine the widest circular cone with apex at X (and axis extending to infinity in one direction) which does not include any other points in the structure. The entire "view" as determined by the cone is hence unobstructed by any other point. Our measure of the view is defined as the cosine of the angle between the axis of the cone and a ray originating at the apex along the surface of the cone. A limiting value of 1 hence denotes a point that is completely buried, for which the widest cone of view reduces to a line. A value of 0 represents a cone that delimits a half plane and corresponds to a "wide" and "unobstructed" view (all points in the structure lie on one side of the half plane). We also extend our measure to cover a cone that is even larger and "flips over". In particular, a value approaching -1 corresponds to a point situated at the tip of a thin peninsula and having virtually a 360-degree view of the surrounding media. We can also capture the occurrence of pockets on the surface by allowing a cutoff radius which limits the relevant environment of a point X. MCV bears some similarity to the solid angle measure defined by Connolly (Connolly 1985), but differs in several important ways as it was designed for a different purpose. Connolly places a (very small) sphere of fixed radius around an atom on the surface, and measures the solid angle corresponding to the portion of the sphere that lies inside the protein. His method is sensitive to detecting changes in surface shape, but would assign a solid angle of 360 degrees to any buried atom. In addition, his measure is also considerably more expensive to compute, which is not surprising as it was not designed to be used in structure estimation. Connolly's solid angle measure has been applied to rigid body docking (Hendrix & Kuntz 1998).

Our measure does not treat surface atoms differently from internal atoms and uses a continuous measure to characterize their surface proximity. It is easily differentiable, and is thus well suited to gradient-based search methods, such as is employed by our probabilistic least squares algorithm. We present an initial validation of our measure showing that it correlates with the commonly used "solvent accessibility" measure (Lee & Richards, 1971). We have also performed experiments in estimating structure from a set of synthetic distance constraints with and without surface information. We show that surface measurements produce improved structures when added to a baseline set of distance information. The benefits of the surface information are not overly sensitive to the accuracy of the provided measurements, and the value of the surface information seems to be highest when 30—50% of short-range distances (simulating a typical NMR data set) are provided. We conclude that our measure is useful for both analysis of structure, as well as for computation of structure using constraints on the degree to which a subset of atoms are buried or on the surface.

Methods

We summarize the method for probabilistic least-squares estimation which has been described in detail previously (Altman, 1995; Chen et al, 1996). The atomic coordinates of the molecule are represented by an n-dimensional state vector

$$\mathbf{x} = (x_1, y_1, z_1, \ldots, x_p, y_p, z_p)^t, \qquad [1]$$

where $p = n/3$ is the number of atoms. In addition, we maintain the state covariance matrix, $\mathbf{C}$, which contains the variances of each atomic coordinate along the diagonal and covariances between atomic coordinates in the off-diagonal entries. The covariances are a linearized summary of the manner in which changes in one coordinate affect another. An observation or constraint on the molecule is modeled by a deterministic function $\mathbf{h}$ of the state vector and a Gaussian noise term $\mathbf{v}$, with zero mean and a variance which reflects the uncertainty in the value of the constraint $\mathbf{z}$:

$$\mathbf{z} = \mathbf{h}(\mathbf{x}) + \mathbf{v}. \qquad [2]$$

A common type of experimental data is near-neighbor distances from NMR measurements. Such a constraint may be modeled in the following form,

$$z_{ij} = h_{ij}(\mathbf{x}) + v = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} + v \,. \qquad [3]$$

The left-hand side, z_{ij}, is the observed value of the distance between atoms i and j; h_{ij} is a function which computes the distance between two points and therefore relates the observation to the underlying state of the system; and $v \sim N(0,R)$ denotes the uncertainty in the measurement. The uncertainty usually depends on the source of the distance information. For NMR measurements, the variance of v may be 1—2 Å^2, while for cross-linking

measurements, the variance may be 4—20 Å^2. In general, an observation **z** may be vector-valued, and the functional dependency of **h** on the state vector **x** may be sparse (as in a distance constraint, which involves 6 coordinates) or dense (such as in the overall volume of a molecule, which involves the coordinates of all the atoms).

The probabilistic least squares algorithm used in these experiments transforms a collection of noisy observations into optimal (in the least squares sense) mean atomic positions and associated uncertainties. The least squares criteria measures the difference between the target mean value (**z** in equation 2) and the computed value, **h(x)**, and divides this distance by the standard deviation of the noise term, thus producing an error measure that is sensitive to the variance in the measurement. Starting with an initial guess $\mathbf{x}_0$ and uncertainty estimates $\mathbf{C}_0$ (usually an uncorrelated covariance matrix), this algorithm sequentially processes the observations on the molecule, producing a series of improved structures and uncertainty estimates until the constraint set is exhausted. For each constraint or vector of constraints, both the state-vector, **x**, and the covariance matrix **C(x)** are updated using a Bayesian update formula borrowed from the optimal filtering literature (e.g. the Kalman Filter, Gelb, 1984). Because of the nonlinearities in **h**, the cycle of constraint application needs to be iterated. In particular, the updated value of **x** and **C(x)**, based on a new measurement, **z**, with variance, **R**, is given by:

$$\mathbf{x}_{new} = \mathbf{x}_{old} + \mathbf{K}(\mathbf{z}-\mathbf{h}(\mathbf{x}_{old})) \quad [4]$$

$$\mathbf{C}(\mathbf{x}_{new}) = \mathbf{C}(\mathbf{x}_{old}) - \mathbf{KHC} \quad [5]$$

Where

$$\mathbf{K} = \mathbf{C}(\mathbf{x}_{old})\mathbf{H}^t(\mathbf{HC}\,\mathbf{H}^t + \mathbf{R})^{-1} \quad [6]$$

is the equivalent of the Kalman gain matrix, and balances the uncertainty in the measurement with the uncertainty in the current state-vector in order to provide a weight to the update term. The matrix, **H**, is the matrix of partial derivatives of **h(x)**, with respect to the state-vector, and is given by:

$$\mathbf{H} = \delta\mathbf{h}(\mathbf{x})/\delta\mathbf{x}. \quad [7]$$

Thus, the **H** vector (or matrix, if **h(x)** is vector-valued) allows the method to search along the gradient of the function describing the constraint. In general, **h(x)** is nonlinear in **x**, and so **H** is evaluated at the current **x** to provide a linearized summary of the gradient. The update equations given above are optimal when **h(x)** is linear, but when it is not, they produce an improved, but imperfect updated value of **x**. In practice, we iterate the application of constraints to the improved estimates of **x** until it converges to a value that satisfies all (or most) of the constraints by the least squares criterion. Detailed discussions of the theoretical basis of the algorithm have been published (Gelb, 1984; Stengel,1994) as have empirical demonstrations of its performance (Altman, 1996; Liu et al 1992; Pachter et al, 1990). The method is not absolutely immune to local minima, but the use of the covariance matrix during search ensures that atoms move in a concerted fashion (based on their covariance), and some strategies analogous to simulated annealing have been used to allow the search space to be explored more robustly (Altman, 1996). The update equations in equations 4 through 7 are quite general, and can be used for any constraint represented as a function of the vector of coordinates, **x**. The task of adding a new type of constraint, such as the maximal conic view, therefore, entails encoding the new constraint as a new function $\mathbf{h}_{mcv}(\mathbf{x})$, and providing its partial derivative for the computation of **H**. The task of using a new constraint requires that the program be provided target mean values and variances.

Maximal Conic View

Our measure for quantifying the degree to which a particular point is on the surface of a collection of points is based on the intuition that a point that is on the surface should have a wide view of the surrounding media, while a point that is buried will have only small angles of view towards the surrounding (the outside of the structure). Figure 1 illustrates in two-dimensions the notion of measuring the size of the maximal "cone" of view to assess the degree to which an object is on the surface. In two dimensions, if we think of a collection of points as houses on an island, we are looking for the widest, unobstructed "ocean view" from each house. In some cases, there may be an unobstructed "bay view", and so we include a cutoff radius that defines the maximum distance for a point to be considered obstructing another.

The computation of the MCV function is quite simple and differs somewhat for cones with positive and cones with negative MCV values. We observe that the largest cone for atom *X* remains unchanged when all neighboring points are projected onto a unit sphere centered at *X*. At this point, it suffices to consider finite cones whose apex is *X* and whose limiting circles lie on the unit sphere. A cut along this limiting circle would remove a section of the sphere which does not contain any points. Computing the largest such cone is therefore tantamount to determining the largest section that can be removed from the sphere in a single cut without removing any of the points. The computation for a point of interest *X* proceeds as follows:

1. All points within the cutoff radius are projected onto the unit sphere centered at *X*.
2. The convex hull of the projected points is computed.

If *X* is inside the convex hull (the removable section is less than half the sphere and the cone is a "real cone")
then
The cut along the limiting circle of the cone is always along a facet of the convex hull.

The distance from X to each facet of the convex hull is computed and the facet with minimum distance reported. This distance corresponds to $\cos(\alpha)$, (where α is the angle between the axis of the cone and a segment from the apex to the limiting circle of the cone.)

Else if X is not inside the convex hull (the removable section is greater than half the sphere and the cone is an "inverted cone")

then

The best cut does not necessarily go along a facet of the convex hull but might go along a circle determined by only two points on the same facet. In either case the limiting circle is readily determined by examining all the hull facets, and our measure is again $\cos(\alpha)$, corresponding to the negative distance from X to the cut surface.

Note that although the computation of the MCV function examines all points within the cutoff range of X, the actual value of the function is eventually determined by four points; the point X and the three points which determine the closest facet of the convex hull (with the following three exceptions: the closest facet contains more than 3 points; there are equidistant closest facets to X; the limiting circle of the cone is supported by two projected points only). The partial derivatives of the MCV are readily expressed analytically and their computation very efficient. The derivative is zero for any atom other than X that is not in the closest facet(s). For the central atom X, and for the atoms that form the closest facet on the convex hull, the derivative can be computed analytically, and is basically the derivative of the cosine of the conic angle in the three Cartesian directions. Due to symmetry considerations these derivatives can be computed by a single procedure. These derivatives guide the procedure for updating the positions of a set of points using constraints on the surface or buried status of some or all of the points.

Preliminary Evaluation of Maximal Conic View

The goal of our first set of experiments was to ensure that our measure correlated roughly with known measures of surface. We computed the MCV for a set of atoms from a myoglobin crystal structure (PDB entry 1mba), and compared these with a well established solvent accessibility surface algorithm, used by the DSSP program (Kabsch & Sander, 1983). Surface accessibility is defined for each amino acid in the structure, while our measure is defined for each atom. To obtain a meaningful comparison we computed, for each amino acid in the structure, both the average value as well as the minimum value of our MCV function over the atoms within the amino acid. (Note that the minimum MCV value corresponds to the widest cone.) We compared both values to the solvent accessibility measures provided by DSSP.

Using Maximal Conic View for Structure Computations

The primary aim of our second set of experiments was to gauge the utility of our surface measure as a constraint on the position of atoms during a structure computation. We start with a series of small sets of short-range distances, not enough to uniquely determine a 3D structure. The goal of the experiments was to compare the quality of structures produced from these distances alone, with the quality of structures produced using these distances augmented with information about the surface proximity of some of the atoms. For our test case, we used the Cα atoms backbone of a myoglobin (PDB entry 1mbd, 153 residues), and extracted sets of between 20% and 55% of all short-range (up to 10 Å) distances within the Cα skeleton. We examined the effectiveness of adding surface constraints along two independent directions: abundance and accuracy. We chose 38 atoms with the most extreme DSSP surface accessibilities (≤ 10 or ≥ 170) for the small data set. The large data set contained the most extreme 75 DSSP ACC values (≤ 40 or ≥ 140). For each set of data, we created two versions: one set with exact MCV values of the targeted atoms, and another set with an estimation of the MCV values based on the linear correlation with the DSSP measures. The data sets with the exact values were designed to evaluate the ability of our algorithm to use the MCV constraints to produce improved structures (as compared to structures computed based on distances alone). For an unknown structure, of course, exact surface values would not be known, but would be estimated by experimental methods. The data sets with predicted values were therefore designed to evaluate the sensitivity of the method to noisy data. We first computed structures based on the distance data sets alone and then computed structures based on these distances with each of the four surface constraint data sets.

Results

Figure 2 shows the relationship between solvent accessibility as used in the DSSP program and both the average as well as the minimum MCV value from atoms within each amino acid. For the comparison with solvent accessibility, the correlation coefficient using the average MCV value was -0.83, and the correlation when using the minimum MCV value was -0.91. Solvent accessibility correlates better with the minimum value of MCV among the atoms within an amino acid. This is not surprising: for an amino acid near the surface of the molecule, several atoms in a large side chain might not have a very large unobstructed view of the surrounding media, due to the interspersion of its side chain atoms. Some of the side chain atoms, however, will have such views. Conversely, amino acids in the hydrophobic core of a protein will generally all be packed into the molecule, and none of them will have an unobstructed view.

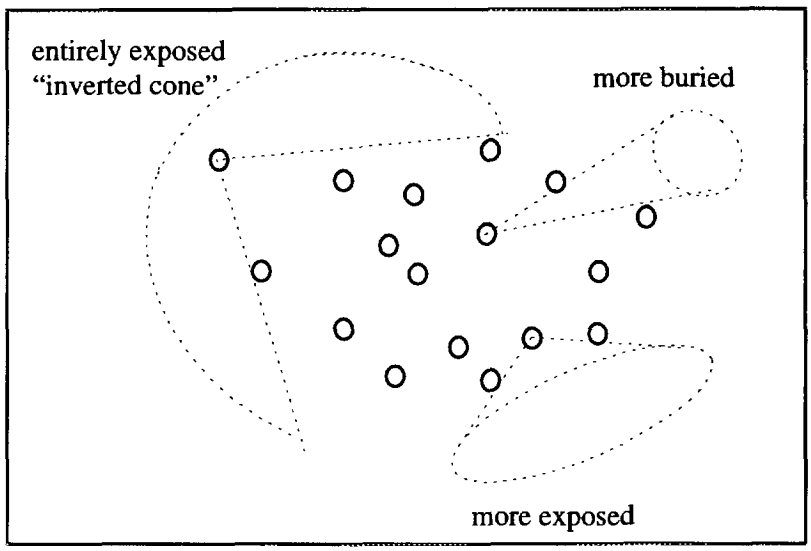


FIGURE 1
The Maximal Conic View (MCV) measure of surface. A set of points is shown, along with the maximum cone that can be drawn representing a "clear view" of the surrounding media. Buried points have a small maximal cone for viewing the media. Points on the surface have a large cone, or even an inverted cone.

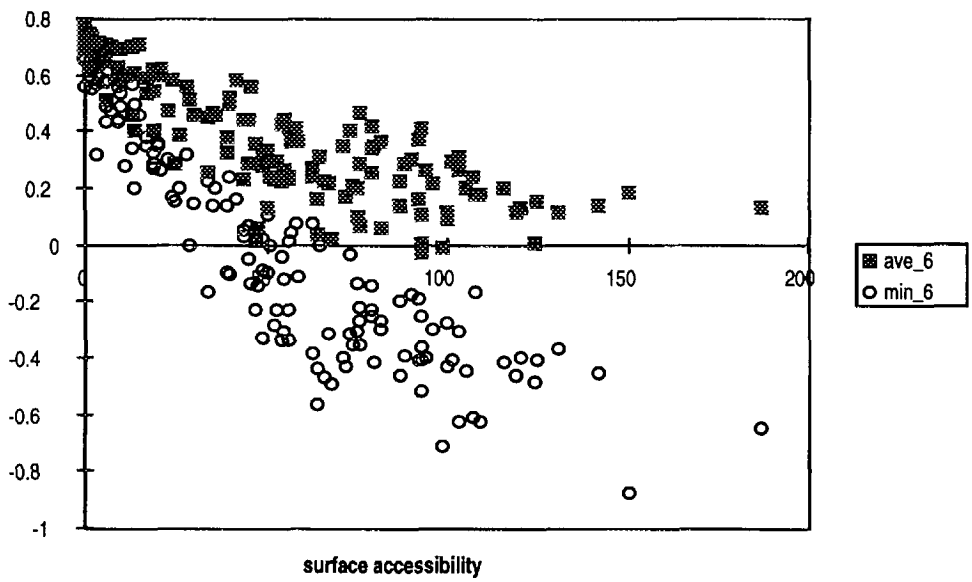
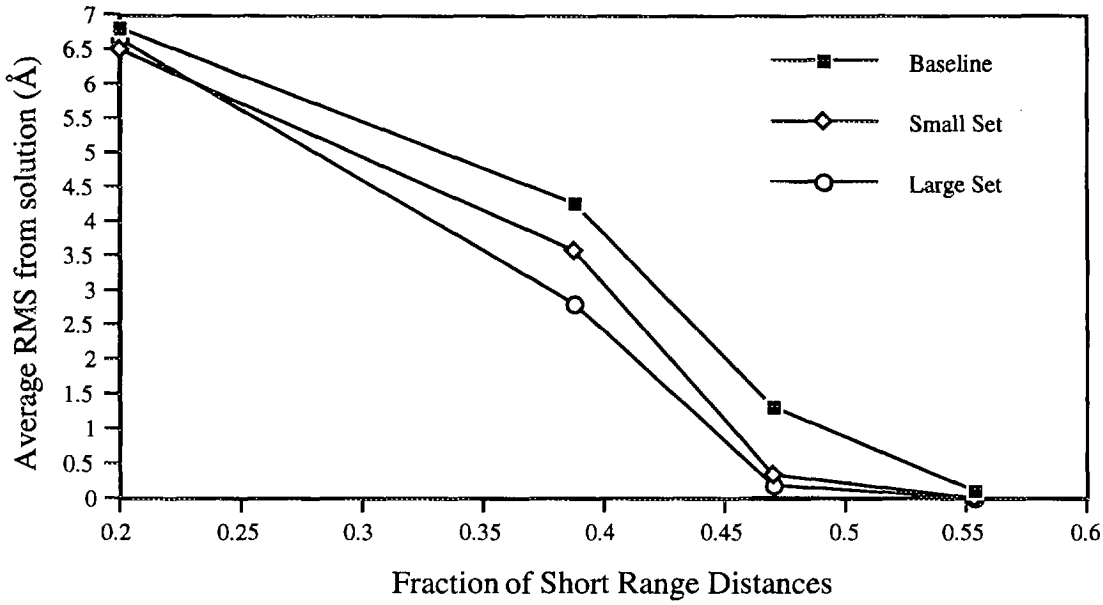
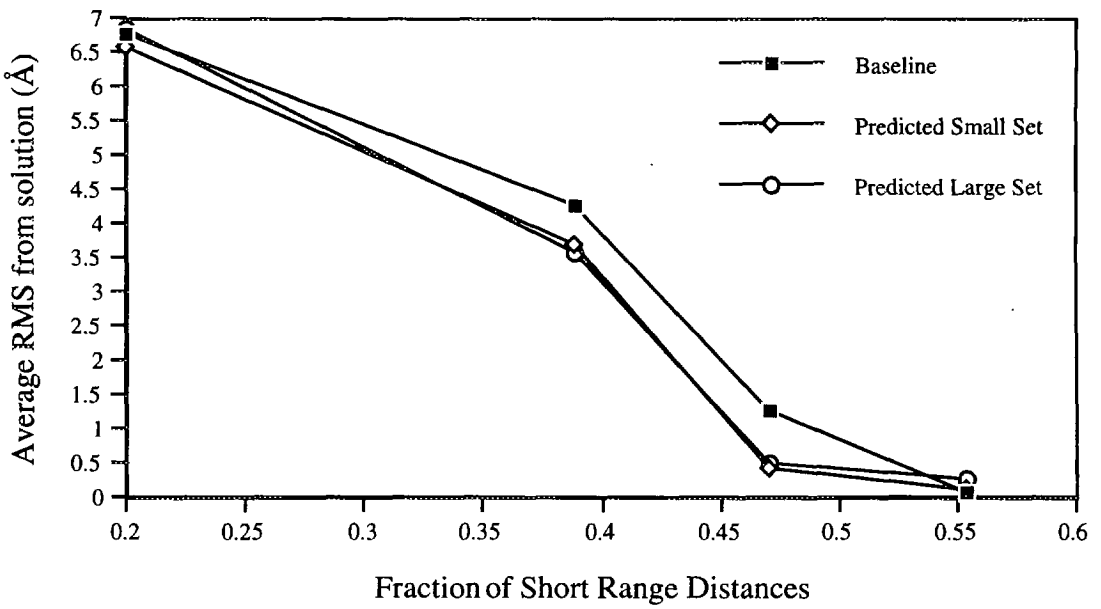


FIGURE 2
A scatter plot of our conic view function versus surface accessibility for the amino acids in myoglobin. Our MCV function for each amino acid was computed as the minimum value of all MCV values and as the average value of the MCV values for the atoms in the amino acid. The cutoff parameter around each atom was set to 6 Å.



A



B

FIGURE 3

(A) The RMSD of computed structures from the gold standard myoglobin structure is shown as a function of the number of distances provided in a series of synethetic data sets. The baseline calculation includes only the distances. The Small Set and Large Set plots refer to sets of surface and buried constraints provided to the probabilistic least squares algorithm with exact values. (B) The RMSD from gold standard results are shown for baseline and surface data sets that are predicted based on linear regression from the observed solvent accessibility values (in order to simulate noisy data).

The results of our structure computations are summarized in Figure 3. Baseline performance refers to the mean deviation from the gold standard of the structure computed from the distance subsets only. SmallSet refers to the distances augmented with the 38 most extreme surface values, and LargeSet refers to the distances augmented with the 75 most extreme values. The results for providing both exact surface constraints (ideal conditions) and inexact surface constraints (simulating data that could be obtained experimentally) are shown. For low data abundance (less than 30% of distances), the surface constraints do not have sufficient information to assist in the computation of the structure. However from 30% to 50% of distances, the surface constraints show significant improvement. In the case of 40% of data, 75 exact surface measurements improve the mean deviation from the gold standard by almost 2 Å (compared with baseline computation, from 4.5 to 2.7 Å), while the smaller data sets or inexact data sets have a 1 Å improvement. At 50% of the possible short-range distances, all the surface data sets produce approximately a 1 Å improvement, almost getting the structure exactly right. With 55% or more of the possible distances, all data sets have sufficient information to get the structure right, and so the surface measurements add little information.

Discussion

The computational complexity of our conic view measure is mostly dependent on the convex hull computation. The average complexity of the randomized algorithm that we used for computing the convex hull of n atoms is $O(n \log n)$. Interestingly, decreasing the cutoff radius allows for the detection of pockets on the outer surface and also decreases the computational complexity dramatically. Pockets can be detected with the MCV function by looking for sharp decreases in the value of the MCV as the cutoff is gradually decreased. For truly buried points, varying the cutoff has almost no effect on the MCV function. The gain in computational complexity is achieved because the number of atoms within a small sphere around a point X is constant. The hull is hence computed for only $O(1)$ points and the complexity of the computation reduced to $O(1)$. This assumes that the points within cutoff range can be identified in $O(1)$ time. This is true in an amortized sense. The points can be preprocessed in $O(n)$ time to enable the quick determination of the neighbor points around each atom X; the convex hull computations (one per atom in a chosen data set of size $O(n)$) will hence take a total of $O(n)$ time.

The computational complexity of the entire structure estimation process depends on several parameters. The evaluation of a distance constraint is achieved in constant time, while the conic view function for $O(N)$ atoms is computable in $O(1)$ (amortized) time per atom, as explained above. Therefore, provided that the cutoff radius is sufficiently small, for a molecule of size $O(N)$, one cycle of application of $O(D)$ distance constraints and $O(S)$ surface proximity constraints takes $O(D+S+N)=O(N)$ time, in the evaluation of the constraint functions alone. Associated with each constraint application is a covariance matrix update, which takes $O(N^2)$ time and is the bottleneck of the computation. One cycle of $O(C)$ constraint applications is thus $O(CN^2)$, (or $O(N^3)$ for $C=O(N)$). The number of cycles of constraint application until convergence is usually small, typically 30—200.

The comparison with solvent accessibility shows that MCV correlates roughly with other commonly used measures of surface proximity (Figure 2). Previous methods for determining the degree to which an atom is on the surface have relied on different insights. The *accessible surface* is computed by tracing the center of a sphere probe as it rolls over the atoms of the molecule (Connolly, 1996). *Molecular surface* is similar to accessible surface, but removes sharp discontinuities that occur with accessible surfaces. A comprehensive review of such methods was recently given by Connolly, (Connolly, 1996). As pointed out earlier surface area measures do not provide a continuous scale of buried to exposed status, as does MCV, and have difficulty distinguishing between somewhat buried and deeply buried. They are also computationally more expensive. In the case of surface area measures, if an atom is below the surface, the functional dependency of the surface area on that atom vanishes, and the partial derivatives with respect to that atom is zero. A gradient-based optimization procedure will be unable to use the surface area measure to nudge a buried atom to the surface, although movement in the opposite direction is possible. Surface information has nevertheless been used successfully in several instances. In particular, the MC-SYM program for modeling RNA structure does include a constraint-type called "accessibility", but this constraint is not based on a surface measure but simply on the number of neighbors (Major, 1991). In addition, MC-SYM uses a "generate and test" approach to satisfying surface constraints, and this can be very expensive. We are quite encouraged that our measure of surface can be used as a direct constraint within the objective function of our structure estimation method.

The real test of our measure is its utility in computing structure. In the case of perfect surface/buried information, the results are as expected. The quality of the structure when a set of exact surface measurements is provided is markedly better (for all abundances of distance data tested) than when no surface information is provided. The results show that a large set of exact measures provides the most information, and leads to improvement in the mean deviation of the computed structure from the gold standard

of almost 2Å in the case of 40% of distances, a 40% increase in accuracy. It leads to a 1Å improvement in the setting of 47% of distances, which is an 80% increase in accuracy compared to baseline. The smaller set of exact surface measures produces an increase in the quality to the solution that is as high as 1 Å. In the (more realistic) case of inexact measures, the large and the small data sets also produce clear improvements over the distances alone, but there is no marked difference between using a large or a small data set of surface measures. The small data set was chosen to represent the most extreme surface/buried values in the molecule, while the large data set includes these measures plus some more moderate values. It appears, therefore, that inexact measures for moderate conic view values do not provide a significant amount of information and thus the large and small inexact data sets basically contain the same information. The errors introduced in the inexact small data set however do not seem to affect the information content of this set. This is encouraging since surface constraints are most relevant for atoms that are either clearly buried or clearly exposed.

If we have access to experimental measurements of surface proximity, then it is clear that an accurate estimate of the uncertainty in these measurements will be useful. The probabilistic least squares method that we use requires that every constraint be represented as a mean value and a variance. When the variance is low, the algorithm will try to match the value with greater precision than when the variance is high. If the variance is over-estimated, then the information content of the measures can be lost. On the other hand, if the variance is under-estimated, then the information contained in the measures can be over-emphasized by the program and lead to incorrect structures. This work is motivated by our ongoing efforts to model the structure of large macromolecule ensembles such as the 30S subunit of the ribosome (Fink et al, 1996). In this biological structure, many measurements of surface accessibility are provided experimentally (e.g., Moazed et al, 1986). The published data are very similar to the simulated data produced here: there is a range of values associated with different parts of the molecule, and these correspond to different degrees of surface accessibility. Our results indicate that MCV will allow us to represent these experimental data and use them to complement the distance information that is also available.

Acknowledgements

This work was carried out while JPS was visiting Stanford University through VPW NSF HRD-9627109. JPS was also partially supported by NSF grant CCR-9305873. RBA is supported by NIH LM-05652, LM-06422, NSF DBI-9600637 and a grant from IBM. We would like to thank John Hennessy for his support through ARPA contract DABT63-94-C-0054, and Jaswinder Pal Singh for useful discussions and access to supercomputing facilities. We would also like to thank Michael Connolly and John Tillinghast for discussions, and Kenneth Clarkson for discussion and for making his convex hull code readily available (http://cm.bell-labs.com/netlib/voronoi/hull.html; Clarkson, 1993).

References

Altman, R.B. 1995. A probabilistic approach to determining biological structure: integrating uncertain data sources." *Intl. J. of Human-Computer Studies 42,* 593-616.

Arteca, A. G., and Mezey, P.G., Topological Characterization for simple Molecular Surfaces. Journal of Molecular Structure, 166, 11-16, 1988.

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

Chen, C. C., Chen, R.O., and Altman, R.B. 1996. Constraining volume by matching the moments of a distance distribution. *Comp. Appl. Biosciences 12(4),* 319-326.

K. L.Clarkson, K.Mehlhorn, and R.Seidel. Four results on randomized incremental constructions. Comp. Geom.: Theory and Applications, 185-121, 1993.

Connolly, M. Molecular Surfaces: A Review, Network Science, April, 1996. http://www.awod.com/netsci/Science/Compchem/feature14.html

Connolly, M. Measurement of protein surface shape by solid angles: Journal of Molecular graphics, 3-6, 1985.

Crippen, G. and Havel, T. 1988. *Distance Geometry and Molecular Conformation*, Research Studies Press, Taunton, Somerset, England.

Eisenhaber, F., and Argos, P., Improved Strategy in Analytic Surface Calculation for Molecular Systems: Handling Singularities and Computational Efficiency. Journal of Computational Chemistry, Vol. 14, No 11, 1272-1280, 1993.

Eisenhaber, F., Lijnzaad P., Argos P., Sander, C. And Scharf, M., The Double Cubic Lattice Method:. Journal of Computational Chemistry, Vol. 16, 3, 273-284, 1995.

Fink, D.L., Chen, R.O., Noller, H.F. and Altman, R.B. 1996. Computational methods for defining the allowed conformational space of 16S rRNA based on chemical footprinting data. *RNA* 2(9), 851-866.

Gelb, A. (editor) 1984. *Applied Optimal Estimation*, the MIT Press, Cambridge, MA.

Glatter, O., The interpretation of real-space information from small-angle scattering experiments. J. Appl. Cryst. 12: 166-175, 1979.

Harris, M.E., Nolan, J.M., Malhotra, A., Brown, J.W., Harvey, S.C., and Pace, N.R., Use of photoaffinity crosslinking and molecular modeling to analyze the global architecture of ribosnuclease P RNA. EMBO J. 13, 3953-3963, 1994.

Hendrix, D. K., and Kuntz, I. D., "Surface Solid Angle-Based Site Points for Molecular Docking", pp. 317—326,

in Proceedings of the Pacific Symposium on Biocomputing '98, Hawaii, 1998.
Kabsch, W. and Sander, C. Dictionary of protein secondary structures: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577-2637, 1983.
Lane, A., and Jardetzky, O., Identificatio of surface residues in the trp repressor of E.Coli, Eur. J. Biochem., 152: 411-418, 1985.
Lee, B. and F. M. Richards (1971). "The interpretation of protein structures: Estimation of static accessibility." J. Mol. Biol. 55 : 379-400.
Liu, Y., Zhao, D., Altman,R.B., and Jardetzky, O. 1992. A systematic comparison of three structure determination methods from NMR data: dependence upon quality and quantity of data. *J. Biomolecular NMR* 2, 373-388.
Major F. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. Science 253(5025), 1255-1260 (1991).
Markley, J.L. and Opella S.J. (editors). *Biological NMR Spectroscopy*, Oxford University Press, New York, 1997.
Moazed, D., Stern, S., Noller, H. F. Rapid chemical probing of conformation in 16 S ribosomal RNA and 30 S ribosomal subunits using primer extension, J Mol Biol 187(3) 399-416, 1986.
Nilges, M., Clore, G.M., and Gronenborn, A.M. 1988. "Determination of three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms", *FEBS Lett.* 239, 129—136.
Pachter, R., Altman, R.B., and Jardetzky, O. 1990. The dependence of protein solution structure on the quality of the input NMR data: Application of the double-iterated Kalman filter technique to oxytocin. *J. Mag. Res.* 89, 578-594.
Powers, T. and Noller, H.N. 1995. Hydroxyl radical footprinting of ribosomal proteins on 16S rRNA. *RNA 1(2),* 194-209.
Stengel, R.F. 1994. *Optimal Control and Estimation*, Dover Publications, New York.