BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 15 2012, pages 1972–1979
doi:10.1093/bioinformatics/bts243

Structural bioinformatics

Advance Access publication April 27, 2012

Optimal simultaneous superpositioning of multiple structures with missing data

Douglas L. Theobald* and Phillip A. Steindel

Department of Biochemistry, Brandeis University, MS009, 415 South Street, Waltham, MA 02454, USA

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Superpositioning is an essential technique in structural biology that facilitates the comparison and analysis of conformational differences among topologically similar structures. Performing a superposition requires a one-to-one correspondence, or alignment, of the point sets in the different structures. However, in practice, some points are usually 'missing' from several structures, for example, when the alignment contains gaps. Current superposition methods deal with missing data simply by superpositioning a subset of points that are shared among all the structures. This practice is inefficient, as it ignores important data, and it fails to satisfy the common least-squares criterion. In the extreme, disregarding missing positions prohibits the calculation of a superposition altogether.

Results: Here, we present a general solution for determining an optimal superposition when some of the data are missing. We use the expectation–maximization algorithm, a classic statistical technique for dealing with incomplete data, to find both maximum-likelihood solutions and the optimal least-squares solution as a special case.

Availability and implementation: The methods presented here are implemented in THESEUS 2.0, a program for superpositioning macromolecular structures. ANSI C source code and selected compiled binaries for various computing platforms are freely available under the GNU open source license from http://www.theseus3d.org.

Contact: dtheobald@brandeis.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 15, 2011; revised on April 13, 2012; accepted on April 21, 2012

1 INTRODUCTION

How should we properly compare and contrast the 3D conformations of similar structures? This fundamental problem in structural biology is commonly addressed by performing a superposition, which removes arbitrary differences in translation and rotation so that a set of structures is oriented in a common reference frame (Flower, 1999). For instance, the conventional solution to the superpositioning problem uses the least-squares optimality criterion, which orients the structures in space so as to minimize the sum of the squared distances between all corresponding points in the different structures. Superpositioning problems, also known as Procrustes problems, arise frequently in many scientific fields, including anthropology, archaeology, astronomy, computer vision, economics, evolutionary biology, geology, image analysis, medicine, morphometrics, paleontology, psychology and molecular biology (Dryden and Mardia, 1998; Gower and Dijksterhuis, 2004; Lele and Richtsmeier, 2001). A particular case we consider here is the superpositioning of multiple 3D macromolecular coordinate sets, where the points to be superpositioned correspond to atoms. Although our analysis specifically concerns the conformations of macromolecules, the methods developed herein are generally applicable to any entity that can be represented as a set of Cartesian points in a multidimensional space, whether the particular structures under study are proteins, skulls, MRI scans or geological strata.

We draw an important distinction here between a structural 'alignment' and a 'superposition.' An alignment is a discrete mapping between the residues of two or more structures. One of the most common ways to represent an alignment is using the familiar row and column matrix format of sequence alignments using the single letter abbreviations for residues (Fig. 1). An alignment may be based on sequence information or on structural information (or on both). A superposition, on the other hand, is a particular orientation of structures in 3D space.

Calculating an optimal superposition normally requires a one-to-one correspondence (a bijection) between the atoms in the different structures (Bourne and Weissig, 2003; Flower, 1999; Gower, 1975). For instance, a sequence alignment is necessary to superposition protein molecules. In many real cases, however, certain residues (and their atoms) are 'missing' in some of the structures. As a case in point, one crystal structure of a protein may omit loop regions that are present in another crystal structure of the same protein. Figure 2a shows a sequence alignment of four protein structures that we wish to superposition. In this example, the protein sequences are identical except that several residues are missing from some structures; only about half of the atoms to be superpositioned are shared among all four proteins (indicated by blue asterisks). An analogous situation exists when we wish to superposition a set of homologous proteins, which in general will have different sequences, various lengths, and gaps and insertions in the alignment (Fig. 1). In both cases, particular columns in the alignment will have gaps, and immediately the question emerges of how to properly incorporate these positions into a global superposition.

Current multiple superposition methods explicitly require complete data (Diamond, 1992; Flower, 1999; Gerber and Müller, 1987; Gower, 1975; Kearsley, 1990; Shapiro *et al.*, 1992; Sutcliffe *et al.*, 1987; Theobald and Wuttke, 2006a), and hence current superposition implementations deal with missing data crudely, usually by excluding many atoms from the calculation (Birzele *et al.*, 2007; Dror, 2003; Guda *et al.*, 2001; Hill and Reilly, 2006;

*To whom correspondence should be addressed.

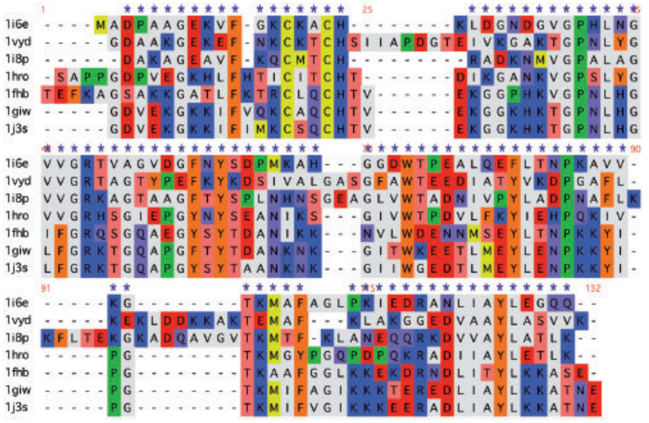


Fig. 1. Cytochrome *c* alignment. An alignment of cytochrome *c* proteins, each of known structure, from seven different species. The gaps can be considered as 'missing data' in a likelihood framework. As in Figure 2, the subset of residues completely shared among all four proteins is indicated by asterisks

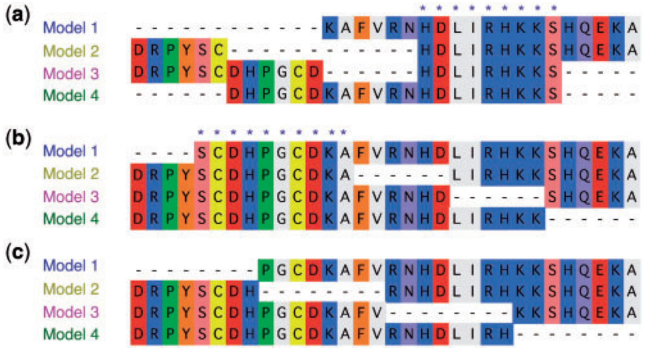


Fig. 2. Alignments with missing data. (**a**) A sequence alignment of four identical proteins, except that different residues are missing in each of the proteins. This could, for instance, correspond to the case of superpositioning different crystal structures where different regions of the protein are disordered in different crystal forms. The subset of residues completely shared among all four proteins is indicated by asterisks. (**b**) A second alignment of the same four proteins with different missing data. (**c**) A third alignment of the same four proteins with no 'common core,' in which no residues are completely shared among all four proteins

Konagurthu *et al.*, 2006; Maiti *et al.*, 2004; Menke *et al.*, 2008; Ortiz, 2002; Shatsky *et al.*, 2002; Ye and Godzik, 2005). For the proteins in Figures 1 and 2, standard practice would calculate the superposition based on only the small subset of fully shared residues, often referred to as the 'common core' (indicated above the alignment by blue asterisks). This method, which corresponds to superpositioning based only on columns in the alignment that contain no gaps, is therefore inefficient as it disregards much of the observed data. Atoms that are not shared among all structures are nevertheless informative, and ideally they should be considered in calculating an optimal superposition. In the most extreme case, no residues are completely shared among the macromolecules (Fig. 2c). Here, the practice of disregarding positions with missing data prohibits the calculation of a superposition altogether, because there is no common core and consequently nothing to include in the calculation. Ignoring some atoms in the superposition also clearly fails to satisfy the least-squares criterion, where the object is to minimize the sum of squared distances among *all* corresponding atoms, not just among a subset of the atoms.

Despite the popularity of ordinary least squares as an optimality criterion for determining the best superposition, other criteria are better justified both theoretically and empirically (Theobald and Wuttke, 2006a, b, 2008). According to the Gauss–Markov theorem, two basic assumptions must be met to justify the use of least squares. In terms of a superposition, these two assumptions require that all atoms have the same variance and that none of the atoms are spatially correlated. Both of these assumptions are strongly violated with biological macromolecules, as certain regions of a structure are more variable than others (due to a combination of experimental imprecision, dynamics and conformational heterogeneity) and because atoms physically communicate with each other (e.g. via electrostatic, Van der Waals and covalent interactions).

We previously presented a maximum-likelihood superposition method that addresses these problems with the least-squares criterion by explicitly allowing individual atoms to be correlated and to have different variances (Theobald and Wuttke, 2006a, b, 2008). As is often true with Gaussian distributed data, the conventional least-squares superposition solution falls out as a special case of the likelihood analysis (namely, when assuming uncorrelated data with equal variances). Most importantly for the present work, likelihood-based methods can elegantly handle cases of missing data via the expectation–maximization (EM) algorithm (Dempster *et al.*, 1977; McLachlan and Krishnan, 1997).

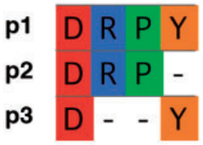
Here, we present solutions for finding the optimal superposition with incomplete structural data. We use the EM algorithm, a classic statistical technique for dealing with missing data, to find maximum-likelihood solutions, which include the conventional least-squares solution as a special case. For completeness, we present two different classes of solutions, listed in decreasing order of generality, complexity and computational requirements: (i) the 'non-isotropic' solution, which allows for heterogeneous atomic variances but assumes that there are no correlations; and (ii) the 'isotropic' solution, which corresponds to the least-squares solution and assumes that all atomic variances are equivalent and no atoms are correlated. This manner of presentation should allow for the straightforward modification of existing least-squares superpositioning routines to handle missing data.

2 APPROACH

2.1 A Gaussian statistical model for the superposition problem

To analyze the superposition problem within a likelihood-based framework and to use the EM algorithm, one must choose a statistical model for the observed data. We assume a perturbation model in which each structure is distributed according to a Gaussian (or Normal) probability density (Goodall, 1991; Lele, 1993; Theobald and Wuttke, 2006a).

The EM algorithm is frequently used in likelihood analyses to determine maximum-likelihood estimates of parameters of a model when some data are missing (Dempster *et al.*, 1977; Pawitan, 2001). The EM method involves cycling between two steps: (i) the



p1	D	R	P	Y
p2	D	R	P	-
p3	D	-	-	Y

Fig. 3. An alignment of three proteins with missing residues. Missing data in this alignment is indicated by the three 'observed' indicator matrices ν_1, ν_2 and ν_3 corresponding to proteins p1, p2 and p3, respectively, which are presented in equation (2)

'E-step' in which the expected likelihood function is calculated, conditional on the observed data and the current estimates of the model parameters; and (ii) the 'M-step' in which the expected likelihood function is maximized over an unknown parameter.

3 METHODS

3.1 Representation of structures with missing atoms

Consider the case of superpositioning r different structures ($\mathbf{X}_i$, $i=1,\ldots,r$), each with a total of k corresponding atoms. We represent each structure as a $k\times 3$ matrix of k rows of atoms, where each atom is a 3-vector. Some of the atoms in each structure may be missing (or unobserved). For each structure $\mathbf{X}_i$, there are m_i missing atoms and a complementary number of n_i observed atoms, so that $m_i+n_i=k$.

To represent structures with missing points, we imagine that the complete data are given in the matrix $\mathbf{X}_i$. A portion of this data is missing or unobserved. The complete data matrix can then be represented as the sum of the observed data and the unobserved data:

$$\mathbf{X}_i=\nu_i\mathbf{X}_i+\mu_i\mathbf{X}_i \tag{1}$$

where $\nu_i\mathbf{X}_i$ is the observed data and $\mu_i\mathbf{X}_i$ is the unobserved, missing data. The complementary indicator matrices ν_i and μ_i are square, symmetric, diagonal, $k\times k$ matrices. The 'observed' indicator matrix, ν_i, contains a one on the diagonal if the corresponding atom in the structure $\mathbf{X}_i$ is observed (i.e. not missing) and zeros otherwise. For example, the following three 'observed' ν indicator matrices correspond to the three small protein fragments in the alignment in Figure 3:

$$\nu_1\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}\qquad\nu_2\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&0\end{bmatrix}\qquad\nu_3\begin{bmatrix}1&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&1\end{bmatrix} \tag{2}$$

Conversely, the 'unobserved' indicator matrix, μ_i, contains a 1 on the diagonal if the corresponding atom in the structure $\mathbf{X}_i$ is missing, and 0s otherwise. Both indicator matrices, though redundant, are useful for simplifying the presentation of the solutions. If no data are missing for structure $\mathbf{X}_i$ then ν_i equals the identity matrix and $\mu_i=\mathbf{0}$, the square matrix of zeros.

The following useful properties hold for the indicator matrices:

$$\nu_i+\mu_i=\mathbf{I}_k \tag{3}$$

$$m_i=\mathrm{tr}\,\mu_i \tag{4}$$

$$n_i=\mathrm{tr}\,\nu_i \tag{5}$$

where $\mathbf{I}_k$ is the $k\times k$ identity matrix and $\mathrm{tr}\,\mathbf{A}$ is the trace of matrix $\mathbf{A}$ (i.e. the sum of the diagonal elements of $\mathbf{A}$).

3.2 The Gaussian perturbation model

In our probabilistic model, each structure $\mathbf{X}_i$ is considered to be a randomly rotated and translated Gaussian perturbation of a mean structure $\mathbf{M}$:

$$\mathbf{X}_i=\left(\mathbf{M}+\mathbf{E}_i\right)\mathbf{R}_i'-\mathbf{1}_k\mathbf{t}_i' \tag{6}$$

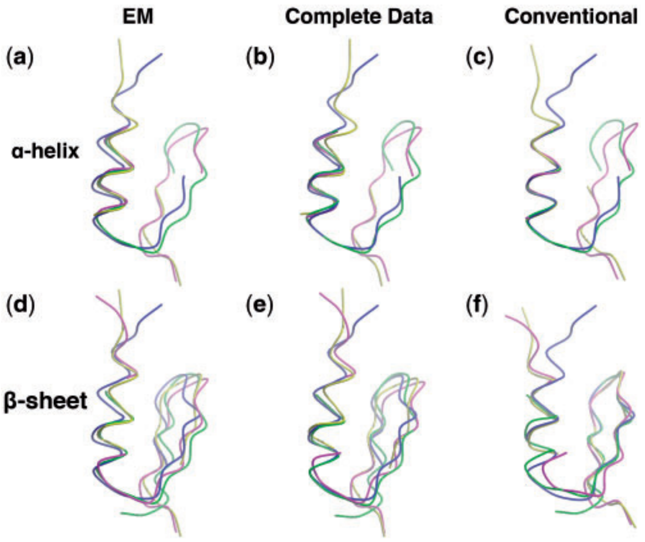


Fig. 4. Least-squares (isotropic) superpositions with missing data. In each pane, four protein structures are superpositioned, each with a different conformation. The top row, **(a)–(c)**, compares superpositions of proteins corresponding to the alignment in Figure 2**a**, where only residues in the α-helix are fully shared among the structures. Other regions of the structures, e.g. the two-stranded β-sheet in the right side of the images, are missing in some of the structures. The bottom row, **(d)–(f)**, compares superpositions corresponding to the alignment in Figure 2**b**, where only residues in the β-sheet are fully shared. The left-most column, **(a)** and **(d)**, shows superpositions found using the EM method described here. The middle column, **(b)** and **(e)**, shows the reference superposition using all of the data; this can be thought of as the 'true' superposition before regions of the structures were deleted. For ease of comparison, in these images, the missing residues are not displayed, even though all of the original data were included in the superposition calculation. The right-most column, **(c)** and **(f)**, shows conventional superpositions based on only the subset of fully shared residues. The structures used in these superpositions were derived from four NMR models of a zinc finger domain, PDB ID 1zfd

where $\mathbf{t}_i$ is a 3×1 translational row vector, $\mathbf{1}_k$ denotes the $k\times 1$ column vector of ones and $\mathbf{R}_i$ is a proper, orthogonal 3×3 rotation matrix. The $k\times 3$ matrix $\mathbf{E}_i$ is a matrix of Gaussian random errors with mean 0, being a random variate from a matrix Gaussian distribution i.e. $\mathbf{E}_i\sim N_{k,3}(\mathbf{0},\Sigma,\mathbf{I}_3)$. Here, Σ is a $k\times k$ covariance matrix for the atoms, which describes the variance of each atom and the covariances among the atoms. For simplicity, we assume that the variance about each atom is spatially spherical. Extensions to higher (and lower) dimensions are trivial (e.g. 4D data would be represented by a $k\times 4$ matrix and use 4×4 rotation matrices and 4×1 translation vectors).

For the non-isotropic solution, the covariance matrix is diagonal, with all offdiagonal, covariance elements constrained to 0. For the isotropic solution, which corresponds to least squares, the covariance matrix is constrained to be diagonal and to have identical diagonal elements (i.e. $\Sigma=\sigma^2\mathbf{I}$).

3.3 The superposition likelihood function

The full joint PDF for our likelihood superposition problem is thus obtained from a multivariate matrix normal distribution (Dutilleul, 1999; Gupta and Nagar, 2000) corresponding to the perturbation model described by equation (6)

$$\mathrm{p}\left(\mathbf{X}|\mathbf{R},\mathbf{t},\mathbf{M},\Sigma\right)\propto \tag{7}$$

$$|\Sigma|^{-dr/2}\exp\left[-\frac{1}{2}\sum_i^r\mathrm{tr}\left\{\left(\mathbf{Y}_i-\mathbf{M}\right)'\Sigma^{-1}\left(\mathbf{Y}_i-\mathbf{M}\right)\right\}\right]$$

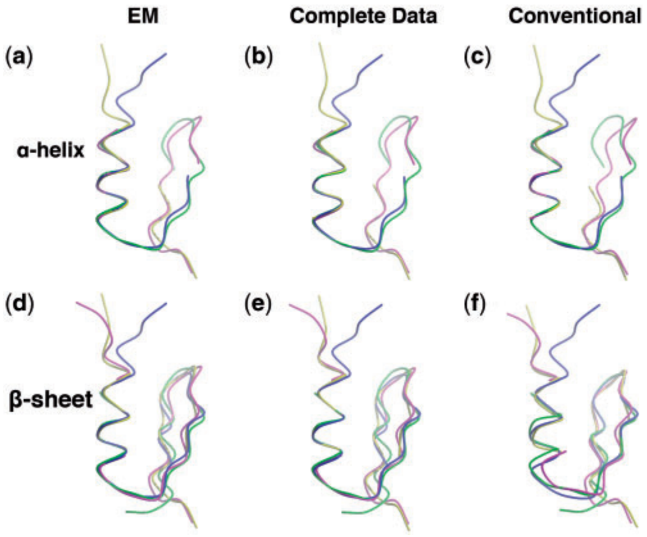


Fig. 5. Maximum-likelihood (non-isotropic) superpositions with missing data. Aside from the optimization criterion, all other details, structures and alignments are as in Figure 4

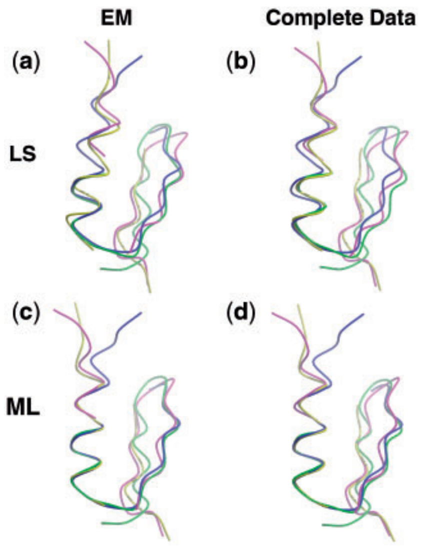


Fig. 6. Superpositions when there is no 'common core.' **(a)** and **(b)**. Least-squares superpositions when no residues are completely shared among the four proteins. Panel **(a)** shows the results of the EM missing data algorithm, based on the alignment shown in Figure 2**c**. **(b)** The original least-squares superposition when all missing data are included in the calculations **(c)** and **(d)**. Corresponding non-isotropic ML superpositions

where

$$\mathbf{Y}_i = \left(\mathbf{X}_i + \mathbf{1}_k \mathbf{t}'_i\right)\mathbf{R}_i \tag{8}$$

and $d = 3$ for 3D data. The Jacobian for the transformation from $\mathbf{Y}_i$ to $\mathbf{X}_i$ is the product of the Jacobians for the translation and rotation, which are each simply unity [see Chapter 1.3 of (Gupta and Nagar, 2000)]. Detailed background and justification of this likelihood treatment can be found elsewhere (Theobald and Wuttke, 2006a, b, 2008).

The full superposition log-likelihood $\ell(\mathbf{R}, \mathbf{t}, \mathbf{M}, \Sigma | \mathbf{X}) = \ell_S$ is therefore given by

$$\ell_S = -\frac{1}{2}\sum_i^r \mathrm{tr}\left\{\left(\mathbf{Y}_i - \mathbf{M}\right)'\Sigma^{-1}\left(\mathbf{Y}_i - \mathbf{M}\right)\right\} - \frac{dr}{2}\ln|\Sigma| \tag{9}$$

With the likelihood function in hand, the ML estimate of a parameter can be derived straightforwardly by taking the derivative of the log-likelihood with respect to the parameter (producing the 'score function'), setting the derivative to zero, and solving for the parameter. Note that columns of the alignment that contain all gaps except for one lone sequence have no influence on the likelihood and should be excluded from the maximization calculations. For the E-step of the EM algorithm, one first finds the expected log likelihood, where the expectation is over the missing data conditional on the observed data and current estimates of the other parameters. In practice, these conditional expectations can be cast in terms of the current parameter estimates, and hence the expectations can be combined with other terms in the log likelihood containing those parameters. For the M-step, the expected log likelihood is maximized over a given parameter by taking the derivative as explained above. In the following sections, the conditional ML estimates are provided for both the complete data and missing data cases. Detailed derivations are provided in the Supplementary Material.

3.4 The translations

ML estimates of the translation parameters are given below.

3.4.1 Complete data solution Where $\hat{\mathbf{t}}_i$ is the estimate of the translation:

$$\hat{\mathbf{t}}_{\mathrm{non,i}} = -\frac{\mathbf{X}'_i \Sigma^{-1}\mathbf{1}_k}{\mathrm{tr}\left(\Sigma^{-1}\right)} \tag{10}$$

$$\hat{\mathbf{t}}_{\mathrm{iso,i}} = -\frac{1}{k}\mathbf{X}'_i \mathbf{1}_k \tag{11}$$

3.4.2 Missing data solution

$$\hat{\mathbf{t}}_{\mathrm{non,i}} = -\frac{(\nu_i \mathbf{X}_i + \mu_i \mathbf{M}\mathbf{R}'_i)'\Sigma^{-1}\mathbf{1}_k}{\mathrm{tr}\left(\nu_i \Sigma^{-1}\right)} \tag{12}$$

$$\hat{\mathbf{t}}_{\mathrm{iso},i} = -\frac{1}{n_i}(\nu_i \mathbf{X}_i + \mu_i \mathbf{M}\mathbf{R}'_i)'\mathbf{1}_k \tag{13}$$

For the remaining solutions, it will be convenient to define a centred structure:

$$\check{\mathbf{X}}_i = \mathbf{X}_i + \mathbf{1}_k \hat{\mathbf{t}}'_i \tag{14}$$

3.5 The rotations

The optimal rotations are calculated using a singular value decomposition (SVD). Let the SVD of an arbitrary matrix $\mathbf{A}$ be $\mathbf{U}\Lambda\mathbf{V}'$. The ML rotations $\hat{\mathbf{R}}_i$ are estimated by

$$\hat{\mathbf{R}}_i = \mathbf{V}\mathbf{P}\mathbf{U}' \tag{15}$$

where rotoinversions can be avoided by constraining the determinant of $\hat{\mathbf{R}}_i$ to be 1 by using $\mathbf{P} = \mathbf{I}$ if $|\mathbf{V}||\mathbf{U}| = 1$ or $\mathbf{P} = \mathrm{diag}(1, \ldots, 1, -1)$ if $|\mathbf{V}||\mathbf{U}| = -1$.

3.5.1 Complete data solution In the non-isotropic case, the $\mathbf{U}$ and $\mathbf{V}$ matrices are determined from the SVD as follows:

$$\hat{\mathbf{M}}'\hat{\Sigma}^{-1}\check{\mathbf{X}}_i = \mathbf{U}\Lambda\mathbf{V}' \tag{16}$$

and in the isotropic case, from

$$\hat{\mathbf{M}}'\check{\mathbf{X}}_i = \mathbf{U}\Lambda\mathbf{V}' \tag{17}$$

3.5.2 Missing data solution For the non-isotropic case

$$\mathbf{M}'\Sigma^{-1}\nu_i\check{\mathbf{X}}_i = \mathbf{U}\Lambda\mathbf{V}' \tag{18}$$

and for the isotropic case

$$\mathbf{M}'\nu_i\check{\mathbf{X}}_i = \mathbf{U}\Lambda\mathbf{V}' \tag{19}$$

3.6 The mean structure

3.6.1 Complete data solution The mean structure is estimated as the arithmetic average of the optimally translated and rotated structures

$$\hat{\mathbf{M}} = \frac{1}{r}\sum_{i}^{r} \check{\mathbf{X}}_i \mathbf{R}_i \tag{20}$$

3.6.2 Missing data solution

$$\hat{\mathbf{M}} = \left(\sum_{i}^{r} \nu_i\right)^{-1} \sum_{i}^{r} \nu_i \check{\mathbf{X}}_i \mathbf{R}_i \tag{21}$$

Both of these solutions are independent of the covariance matrix.

3.7 Covariance matrix and superposition variance

The following equations for the covariance matrix estimates are only valid when the translations are known (Theobald and Wuttke, 2006a). In general, this is not the case, and thus a constrained (regularized) estimator of the covariance matrix is necessary. For instance, the estimates given below may be modified to give a hierarchical estimator as shown in equation (6) of (Theobald and Wuttke, 2006a) or equation (10) of (Theobald and Wuttke, 2008). The estimators of the isotropic variance are already adequately constrained and do not need to be adjusted.

To simplify the following formulae, we first define the matrix $\mathbf{D}_i$:

$$\mathbf{D}_i = \check{\mathbf{X}}_i \mathbf{R}_i - \mathbf{M} \tag{22}$$

3.7.1 Complete data solution The unconstrained estimate of the diagonal, non-isotropic covariance matrix is

$$\hat{\Sigma}_{U,\text{non}} = \mathbf{I}_k \odot \frac{1}{dr}\sum_{i}^{r} \mathbf{D}_i \mathbf{D}_i' \tag{23}$$

where $\mathbf{I}_k$ is the $k \times k$ identity matrix and '$\odot$' is the Hadamard operator for elementwise matrix multiplication. The Hadamard operation simply sets all offdiagonal elements of the covariance matrix to 0.

For a least-squares analysis, which is equivalent to assuming that $\Sigma = \sigma^2\mathbf{I}$, the ML estimate of the variance is

$$\hat{\sigma}^2 = \frac{1}{dkr}\,\mathrm{tr}\left\{\sum_{i}^{r} \mathbf{D}_i \mathbf{D}_i'\right\} \tag{24}$$

Note that in equation (24), one needs to only calculate the diagonal elements.

3.7.2 Missing data solution The ML estimate of the non-isotropic, diagonal covariance matrix is

$$\hat{\Sigma}_{U,non} = \mathbf{I}_k \odot \left(d\sum_{i}^{r} \nu_i\right)^{-1} \sum_{i}^{r} \nu_i \mathbf{D}_i \mathbf{D}_i' \nu_i \tag{25}$$

The estimate of the isotropic covariance matrix ($\Sigma = \sigma^2\mathbf{I}$) is given by

$$\hat{\sigma}^2 = \frac{1}{d\sum_{i}^{r} n_i}\,\mathrm{tr}\sum_{i}^{r} \nu_i \mathbf{D}_i \mathbf{D}_i' \nu_i \tag{26}$$

In both isotropic cases, calculation of the isotropic variance σ^2 can often be omitted, as it is not needed to estimate any of the other parameters. The summation terms may be calculated easily by noting that

$$\nu_i \mathbf{D}_i = \nu_i \check{\mathbf{X}}_i \mathbf{R}_i - \nu_i \hat{\mathbf{M}} \tag{27}$$

4 ALGORITHM

The simultaneous solution of the optimal parameters must be solved numerically, as each of the unknown parameters is a function of some of the others. Our iterative algorithm is an extension of similar algorithms proposed previously, and it is based on nested rounds of EM cycles and conditional maximization (Dempster *et al.*, 1977; Dutilleul, 1999; Goodall, 1991; Theobald and Wuttke, 2006a, b). As with all superposition algorithms, this algorithm requires *a priori* knowledge of the alignment (the one-to-one correspondence among atoms/points in the structures):

(1) *Initialize*: Set $\hat{\Sigma} = \mathbf{I}$ for all i. Estimate the mean structure $\hat{\mathbf{M}}$ by embedding the average of the distance matrices, including gaps, for each structure (Crippen and Havel, 1978; Lele, 1993; Lele and Richtsmeier, 2001). Rather than embedding, one may simply choose one of the structures (preferably with the fewest gaps) to serve as the mean for the first iteration, setting missing coordinates to zeros (convergence may be hindered in cases with a large fraction of gaps).

(2) *Translate*: Translate (i.e. centre) each $\mathbf{X}_i$. For the first iteration, the $\mu_i \mathbf{M}\mathbf{R}_i'$ term can be omitted from equations (12)–(13).

(3) *Rotate*: Calculate each rotation $\hat{\mathbf{R}}_i$ and rotate each translated structure: $\mathbf{X}_i = \check{\mathbf{X}}_i \hat{\mathbf{R}}_i$.

(4) *Estimate the mean*: Recalculate the average structure $\hat{\mathbf{M}}$. Return to Step 2 and loop until convergence.

(5) *Estimate the covariance matrix $\hat{\Sigma}$ or variance σ^2*: If the covariance matrix or the translations are unknown (the usual case), calculate $\hat{\Sigma}$ [equation (25)] and modify it to constrain the eigenvalues to all be >0. In the isotropic case, this step can be omitted if desired. Return to Step 2 and loop until convergence.

5 IMPLEMENTATION

The algorithm described above for calculating optimal superpositions with missing data is implemented in the command-line UNIX program THESEUS (Theobald and Wuttke, 2006a, b). THESEUS functions in two modes: (i) one mode for superpositioning structures with identical sequences [such as different (NMR) models] and (ii) an 'alignment mode,' which superpositions structures with different sequences (i.e. with missing data) conditional on a known alignment. THESEUS does not attempt to determine structure-based sequence alignments, which is a distinct bioinformatics problem (Bourne and Weissig, 2003). When superpositioning homologous proteins with different sequences or identical proteins with missing portions, a sequence alignment must be provided to THESEUS. We provide a wrapper script, *theseus_align*, to perform this latter procedure transparently for the user. The *theseus_align* wrapper script will automatically extract the proper sequences from the Protein Data Bank (PDB) files, align them with a sequence alignment program of the users choice and superposition the structures based on this alignment with THESEUS.

THESEUS will superposition any number of structures within the limits set by the operating system and memory capability. Via command line options, users can choose to superposition assuming an isotropic covariance matrix (i.e. the conventional LS (least squares)-method) or using the non-isotropic model. Specified alignment columns can also be excluded from the calculation. On modern personal desktop computers, convergence is usually very fast (within seconds for even very large problems).

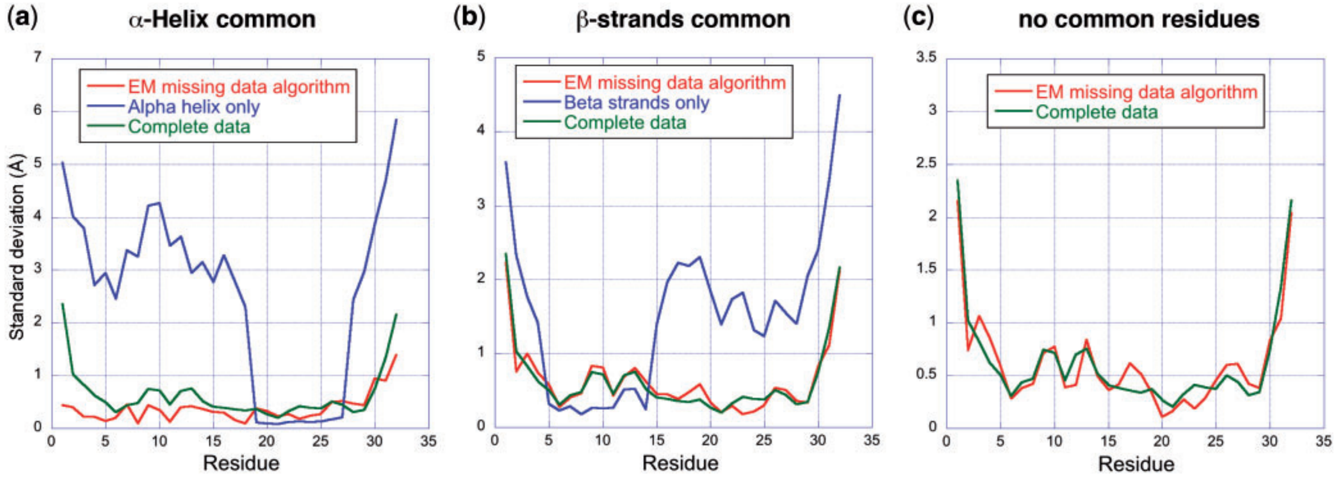


Fig. 7. Standard deviation of α-carbons for least-squares superpositions. For each residue in the structures, the α-carbon standard deviation is plotted. The reference superposition ('complete data') is plotted as the green line, a superposition using the conventional algorithm (based on only the subset of fully shared residues) is plotted as the blue line, and the superposition from our EM algorithm, which includes all data, is plotted as the red line. The reference data (green line) is in fact the same in all three panes; it is re-plotted in each for convenience. (**a**) Superpositions of proteins corresponding to the alignment in Figure 2**a**, where only residues in the α-helix are fully shared among the structures. (**b**) Superpositions corresponding to the alignment in Figure 2**b**, where only residues in the β-sheet are fully shared. (**c**) Superpositions when no residues are completely shared among the four proteins ('no common core')

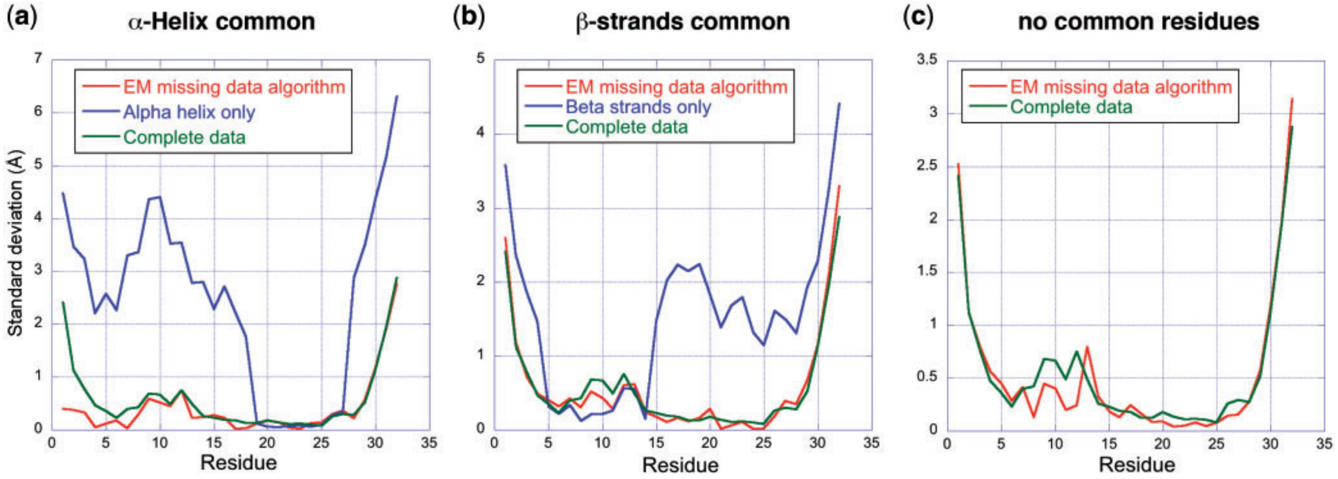


Fig. 8. Standard deviation of α-carbons for non-isotropic maximum-likelihood superpositions. For details see the legend to Figure 7

6 RESULTS AND DISCUSSION

To demonstrate the advantages of the method, we constructed three test sets of structures based on four NMR models of a zinc finger domain protein. In each of the three test sets, different portions of the the four structures were removed. In the first test set, the C-terminal helix of the zinc finger is the only region fully shared among all four partial structures (indicated by asterisks above the alignment columns in Fig. 2**a**). In the second test set, the N-terminal β-sheet is the only region fully shared (Fig. 2**b**). For the third test set, no regions of the protein are fully shared among all four structures—in this set, there is no 'common core' (Fig. 2**c**), and hence this test set is impossible to superposition using conventional algorithms.

The four original, unmodified zinc finger structures were superpositioned with THESEUS (using both the conventional least-squares criterion and non-isotropic maximum likelihood) to provide a 'complete data' superposition. For comparison, the modified structures (with various portions deleted) were then superpositioned using the EM algorithm and using the traditional method which omits columns with gaps. The superpositions with the modified structures can then be compared with the 'complete data' superposition, which serves as a reference.

The EM method produces a least-squares superposition much closer to the 'true' complete data superposition than the conventional method (Fig. 4). The EM superposition is also largely independent of which portions were fully aligned. The EM variances are much lower for the entire structure than the conventional method variances, and they are generally much closer to the 'true' variances (Fig. 7). Results for the non-isotropic ML superpositions are similar to those of the

LS superpositions (Figs. 5 and 8). The EM method can easily handle the 'impossible' situation seen in Figure 2c, with results similar to the true superposition (Figs. 6, 7c and 8c).

Because the conventional superpositions ignore regions of the structure that have missing data, they are biased to closely superposition only the regions that are fully shared. For instance, in Figures 4c and 5c, the superposition is biased to closely orient the α-helix at the expense of the β-sheet, as the α-helix is the only region of the protein fully shared among the three structures. Similarly, the conventional superposition is biased toward the β-sheet in panels 4f and 5f, since the β-sheet is the only region of the protein fully shared among this set of structures.

Our test sets represent somewhat extreme cases, and we expect the effect of accounting for missing data in real proteins to be more subtle. In practice, the effect of missing data will vary from case to case, depending on which regions are missing and on the patterns of structural variability in the proteins. We have found, however, that in many cases, properly handling missing data is an important concern with significant effects on the superposition. One salient example is given in Figure 9, which shows the effects of accounting for missing data in the superpositions of the NMR structure families of two different scorpion toxins. Using both ML and LS superposition methods, significant differences are seen when the missing data are accounted for properly. In all cases, our EM method results in a closer superposition for the unaligned regions (shown in cyan) that would conventionally be excluded from the analysis. The fully aligned regions (shown in dark blue) also exhibit changes in relative conformation, as seen clearly in the per residue RMSD plots.

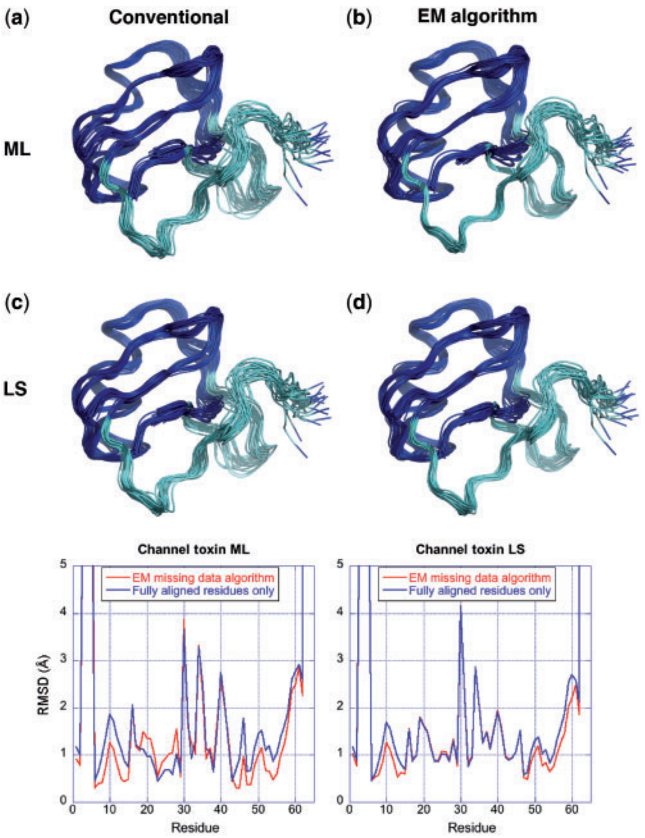


Fig. 9. Comparison of conventional superpositions versus our missing data method for NMR families of two scorpion neurotoxins (PDB IDs 1big and 1i6g). The *Centruroides* neurotoxin (1i6g) contains two insertions, shown in cyan, not found in the Chinese scorpion neurotoxin (1big). The cyan regions hence represent positions with gaps in the alignment, whereas the dark blue regions are fully aligned. The top row, **(a)** and **(b)**, compares non-isotropic ML superpositions of the two neurotoxins, with a conventional superposition using only fully aligned residues at left and a superposition accounting for all data using our EM algorithm at right. The middle row, **(c)** and **(d)**, compares analogous LS superpositions. The bottom row shows plots of α-carbon Root Mean Square Deviation (RMSD) for each residue position

7 CONCLUSION

We have developed a method for superpositioning multiple structures when some of the structures have residues that other structures lack. Our algorithm uses the EM algorithm to find the optimal superposition by treating the gaps in an alignment as 'missing data.' To our knowledge, this is the first superposition solution proposed for cases with incomplete or missing structural information. The use of indicator matrices allows our EM method to be incorporated easily in conventional LS superpositioning procedures. Programs need only be modified to keep track of these matrices and to adjust the calculations accordingly. Hence, our method should be widely applicable to the diverse superposition problems found throughout molecular biology.

ACKNOWLEDGEMENTS

The authors thank Catherine J.L. Theobald for comment and criticism.

Funding: National Institutes of Health, United States [GM094468 and GM096053].

Conflict of Interest: none declared.

REFERENCES

Birzele,F. *et al.* (2007) Vorolign—Fast structural alignment using voronoi contacts. *Bioinformatics*, **23**, e205–e211.

Bourne,P.E. and Weissig,H. (2003) *Structural Bioinformatics*, Vol. 44 of Methods of Biochemical Analysis. Wiley-Liss, Hoboken, NJ.

Crippen,G.M. and Havel,T.F. (1978) Stable calculation of coordinates from distance information. *Acta Crystallogr. A*, **34**, 282–284.

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B Metab.*, **39**, 1–38.

Diamond,R. (1992) On the multiple simultaneous superposition of molecular-structures by rigid body transformations. *Protein Sci.*, **1**, 1279–1287.

Dror,O. (2003) Multiple structural alignment by secondary structures: algorithm and applications. *Protein Sci.*, **12**, 2492–2507.

Dryden,I.L. and Mardia,K.V. (1998) *Statistical Shape Analysis*. Wiley series in probability and statistics. John Wiley & Sons, Chichester, New York.

Dutilleul,P. (1999) The MLE algorithm for the matrix normal distribution. *J. Stat. Comput. Simul.*, **64**, 105–123.

Flower,D.R. (1999) Rotational superposition: a review of methods. *J. Mol. Graph Model*, **17**, 238–244.

Gerber,P.R. and Müller,K. (1987) Superimposing several sets of atomic coordinates. *Acta Crystallogr A*, **43**, 426–428.

Goodall,C.R. (1991) Procrustes methods in the statistical analysis of shape. *J. Roy. Stat. Soc. B Metab.*, **53**, 285–321.

Gower,J.C. (1975) Generalized Procrustes analysis. *Psychometrika*, **40**, 33–51.

Gower,J.C. and Dijksterhuis,G.B. (2004) *Procrustes Problems*, Vol. 30 of Oxford Statistical Science Series, Oxford University Press, Oxford, New York.

Guda,C. *et al.* (2001) A new algorithm for the alignment of multiple protein structures using monte carlo optimization. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, **6**, 275–286.

Gupta,A.K. and Nagar,D.K. (2000) *Matrix Variate Distributions*, Vol. 104. Chapman and Hall.

Hill,A.D. and Reilly,P.J. (2006) Comparing programs for rigid-body multiple structural superposition of proteins. *Proteins*, **64**, 219–226.

Kearsley,S.K. (1990) An algorithm for the simultaneous superposition of a structural series. *J. Comput. Chem.*, **11**, 1187–1192.

Konagurthu,A.S. *et al.* (2006) Mustang: a multiple structural alignment algorithm. *Proteins*, **64**, 559–574.

Lele,S. (1993) Euclidean distance matrix analysis (EDMA)—estimation of mean form and mean form difference. *Math. Geol.*, **25**, 573–602.

Lele,S. and Richtsmeier,J.T. (2001) *An Invariant Approach to Statistical Analysis of Shapes*. Interdisciplinary statistics. Chapman and Hall/CRC, Boca Raton, FL.

Maiti,R. *et al.* (2004) SuperPose: a simple server for sophisticated structural superposition. *Nucleic Acids Res.*, **32**, W590–W594.

McLachlan,G.J. and Krishnan,T. (1997) *The EM Algorithm and Extensions*. Wiley series in probability, and statistics, Applied Probability and Statistics,. Wiley, New York.

Menke,M. *et al.* (2008) Matt: local flexibility aids protein multiple structure alignment. *PLoS Comput. Biol.*, **4**, e10.

Ortiz,A.R. (2002) Mammoth (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.

Pawitan,Y. (2001) *In All Likelihood: Statistical Modeling and Inference Using Likelihood*. Oxford Science Publications, Clarendon Press, Oxford.

Shapiro,A. *et al.* (1992) A method for multiple superposition of structures. *Acta Crystallogr. A*, **48**, 11–14.

Shatsky,M. *et al.* (2004) A method for simultaneous alignment of multiple protein structures. *Proteins: Structure, Function, and Bioinformatics*, 56:143–56.

Sutcliffe,M. *et al.* (1987) Knowledge based modelling of homologous proteins, part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Engi.*, **1**, 377–384.

Theobald,D.L. and Wuttke,D.S. (2006a) Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem. *Proc. Natl. Acad. Sci. USA*, **103**, 18521–18527.

Theobald,D.L. and Wuttke,D.S. (2006b) THESEUS: Maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics*, **22**, 2171–2172.

Theobald,D.L. and Wuttke,D.S. (2008) Accurate structural correlations from maximum likelihood superpositions. *PLoS Comput. Biol.*, **4**, e43.

Ye,Y. and Godzik,A. (2005) Multiple flexible structure alignment using partial order graphs. *Bioinformatics*, **21**, 2362–2369.