SUPPLEMENTARY MATERIALS

THEORY

A Gaussian statistical model for the superposition problem

A more general treatment of the following theoretical likelihood analysis can be found elsewhere in detail (Theobald and Wuttke, 2006). Here we briefly recap a simplification of the main results that specifically apply to macromolecular structural superpositioning as implemented in THESEUS.

Consider the case of superpositioning N different structures ($\mathbf{X}_i$, $i = 1 \dots N$), each with 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.

We assume a perturbation model in which each macromolecular structure $\mathbf{X}_i$ is distributed according to a matrix Gaussian probability density (Goodall, 1991; Lele, 1993). In this likelihood model, each observed structure $\mathbf{X}_i$ is considered to be a randomly rotated and translated Gaussian perturbation of a mean structure $\mathbf{M}$:

$$\mathbf{X}_i = (\mathbf{M} + \mathbf{E}_i)\mathbf{R}_i' - \mathbf{1}_K \boldsymbol{t}_i \tag{1}$$

where $\boldsymbol{t}_i$ is a 1×3 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 zero, *i.e.*, $\mathbf{E}_i \sim N_{K,3}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_3)$. Here $\boldsymbol{\Sigma}$ 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 spherical.

The superposition likelihood equation

In general, the covariance matrix $\boldsymbol{\Sigma}$ is inestimable unless it is parametrically constrained. Therefore, to estimate the atomic covariance matrix we assume that its eigenvalues are distributed according to an inverse gamma distribution, which is physically reasonable for macromolecular structures. The joint log-likelihood for our likelihood superposition problem is thus the sum of the log-likelihood for the atomic covariance matrix eigenvalues and the log-likelihood for a multivariate matrix normal distribution (Arnold, 1981; Dutilleul, 1999) corresponding to the perturbation model described by Eq. **1**. The full superposition log-likelihood $\ell(\mathbf{R}, \boldsymbol{t}, \mathbf{M}, \boldsymbol{\Sigma}; \mathbf{X}) = \ell_S$ is given by

$$\begin{aligned} \ell_S \quad = \quad & -\frac{1}{2}\sum_i^N \|(\mathbf{X}_i + \mathbf{1}_K \boldsymbol{t}_i)\mathbf{R}_i - \mathbf{M}\|^2_{\boldsymbol{\Sigma}^{-1}} \\ & -\frac{3NK}{2}\ln(2\pi) - \frac{3N}{2}\ln|\boldsymbol{\Sigma}| \\ & -(1+\gamma)\ln|\boldsymbol{\Sigma}| - \alpha \operatorname{tr}\boldsymbol{\Sigma}^{-1} \\ & +K\gamma\ln\alpha - K\ln\Gamma(\gamma) \end{aligned} \tag{2}$$

where $|\mathbf{U}|$ denotes the determinant of the matrix $\mathbf{U}$, $\|\mathbf{U}\|^2_{\mathbf{V}} = \operatorname{tr}\{\mathbf{U}'\mathbf{V}\mathbf{U}\}$ denotes a squared Frobenius Mahalanobis matrix norm, and α and γ are the scale and shape parameters, respectively, of an inverse gamma distribution of the atomic covariance matrix's nonzero eigenvalues (λ_j):

$$\mathrm{P}(\lambda_j) = \frac{\alpha^\gamma}{\Gamma(\gamma)}\lambda_j^{-(1+\gamma)} e^{-\frac{\alpha}{\lambda_j}} \tag{3}$$

ML superposition solutions

We provide in the following the ML solutions for each of the unknown parameters of the above superposition log-likelihood equation.

Each uncentered structure $\mathbf{X}_i$ must be centered by translating it so that its row-weighted center is at the origin. Row-wise weighted centering is applied by

$$\check{\mathbf{X}}_i = \mathbf{X}_i + \mathbf{1}_K \hat{\boldsymbol{t}}_i \tag{4}$$

where $\hat{\boldsymbol{t}}_i$ is the ML estimate of the optimal translation:

$$\hat{\boldsymbol{t}}_i = -\frac{\mathbf{1}_K'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i}{\mathbf{1}_K'\boldsymbol{\Sigma}^{-1}\mathbf{1}_K}$$

The extended ML estimator of the inverse gamma distributed atomic covariance matrix $\hat{\boldsymbol{\Sigma}}_{I\gamma}$ is given by:

$$\hat{\boldsymbol{\Sigma}}_{I\gamma} = \frac{3N}{3N + 2(\gamma+1)}\left(\frac{2\alpha}{3N}\mathbf{I} + \hat{\boldsymbol{\Sigma}}_U\right) \tag{5}$$

where the unrestricted ML estimator of the covariance matrix $\hat{\boldsymbol{\Sigma}}_U$ is:

$$\hat{\boldsymbol{\Sigma}}_U = \frac{1}{3N}\sum_i^N (\check{\mathbf{X}}_i\mathbf{R}_i - \hat{\mathbf{M}})(\check{\mathbf{X}}_i\mathbf{R}_i - \hat{\mathbf{M}})' \tag{6}$$

Similarly, the extended ML estimator of the inverse gamma distributed eigenvalues $\hat{\boldsymbol{\Lambda}}_{I\gamma}$ is given by:

$$\hat{\boldsymbol{\Lambda}}_{I\gamma} = \frac{3N}{3N + 2(\gamma+1)}\left(\frac{2\alpha}{3N}\mathbf{I} + \hat{\boldsymbol{\Lambda}}_U\right) \tag{7}$$

where $\hat{\boldsymbol{\Lambda}}_U$ is the diagonal matrix of eigenvalues of the unrestricted sample covariance matrix $\hat{\boldsymbol{\Sigma}}_U$. The rotations are calculated via a singular value decomposition (SVD). Let the SVD of an arbitrary matrix $\mathbf{D}$ be $\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}'$. The optimal rotations $\hat{\mathbf{R}}_i$ are then estimated by

$$\begin{aligned} \hat{\mathbf{M}}'\hat{\boldsymbol{\Sigma}}^{-1}\check{\mathbf{X}}_i &= \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}' \\ \hat{\mathbf{R}}_i &= \mathbf{V}\mathbf{P}\mathbf{U}' \end{aligned} \tag{8}$$

Rotoinversions are prevented by ensuring that the rotation matrix $\hat{\mathbf{R}}_i$ has a positive determinant: $\mathbf{P} = \mathbf{I}$ if $|\mathbf{V}|\,|\mathbf{U}| = 1$ or $\mathbf{P} = \operatorname{diag}(1, 1, -1)$ if $|\mathbf{V}|\,|\mathbf{U}| = -1$. The mean form is estimated as the arithmetic average:

$$\hat{\mathbf{M}} = \bar{\mathbf{X}} = \frac{1}{N}\sum_i^N \check{\mathbf{X}}_i\mathbf{R}_i \tag{9}$$

Finally, the overall "tightness" of a maximum likelihood superposition can be assessed by a maximum likelihood analog of the conventional least-squares root mean squared deviation (RMSD_{LS}):

$$\mathrm{RMSD}_{ML} = \sqrt{\frac{K}{\operatorname{tr}\hat{\boldsymbol{\Sigma}}^{-1}}} \tag{10}$$

When all variances are equal and there are no correlations (*i.e.* the least-squares assumption), the two measures are equivalent.

ALGORITHM

The ML solutions given above must be solved simultaneously using a numerical maximization algorithm. For this purpose, we have developed the following iterative algorithm based on the Expectation-Maximization (EM) method (Dempster *et al.*, 1977; Pawitan, 2001). In brief:

1. **Initialize:** Set $\hat{\mathbf{\Sigma}} = \mathbf{I}$. Estimate the mean structure $\hat{\mathbf{M}}$ using the EDMA embedding method (Lele, 1993).
2. **Translate:** Center each $\mathbf{X}_i$ according to Eq. **4**.
3. **Rotate:** Calculate each $\hat{\mathbf{R}}_i$ according to Eq. **8**, and rotate each centered structure accordingly: $\mathbf{X}_i = \check{\mathbf{X}}_i\hat{\mathbf{R}}_i$.
4. **Estimate the mean:** Recalculate the average structure $\hat{\mathbf{M}}$ according to Eq. **9**. Return to step 3 and loop to convergence.
5. **Estimate the inverse gamma distributed eigenvalues $\hat{\mathbf{\Lambda}}$:** Estimate $\hat{\mathbf{\Sigma}}_U$ from Eq. **6**, and spectrally decompose it to find the sample eigenvalues $\hat{\mathbf{\Lambda}}_U$. Estimation of the inverse gamma distributed eigenvalues $\hat{\mathbf{\Lambda}}_{I\gamma}$ involves the simultaneous solution of two problems: (1) the modification of $\hat{\mathbf{\Lambda}}_U$ according to Eq. **7** and (2) ML estimation of the scale and shape parameters of the inverse gamma distributed eigenvalues. Details of this step are given below in *Estimation of the eigenvalues*.
6. **Estimate the atomic covariance matrix $\hat{\mathbf{\Sigma}}$:** Modify $\hat{\mathbf{\Sigma}}_U$ according to Eq. **5**.
7. **Loop**: Return to step 2 and loop until convergence.

Note that if one assumes that the variances are all equal (*i.e.*, that $\mathbf{\Sigma} \propto \mathbf{I}$), then the above algorithm is equivalent to the conventional least-squares algorithm for multiple simultaneous superpositioning (Diamond, 1992; Gerber and Müller, 1987; Kearsley, 1990; Shapiro *et al.*, 1992).

Estimation of the eigenvalues:

Estimation of the the diagonal matrix of inverse gamma distributed eigenvalues $\hat{\mathbf{\Lambda}}$ involves three steps: (1) calculate the unconstrained sample eigenvalue matrix $\hat{\mathbf{\Lambda}}_U$ by eigen decomposition of the sample covariance matrix given by Eq. **6**, (2) estimate the scale and shape parameters of an inverse gamma distribution based on these eigenvalues, and (3) modify the unconstrained sample eigenvalues $\hat{\mathbf{\Lambda}}_U$ according to Eq. **7**. Steps 2 and 3 must be performed simultaneously. An iterative EM algorithm that performs this simultaneous estimation is described below.

Before iterating, the EM algorithm must be initialized. In general, the sample covariance matrix (Eq. **6**) is first spectrally decomposed to determine the sample eigenvalues. However, this decomposition is unnecessary if one assumes that the covariance matrix is diagonal, since then the variances are the eigenvalues (with an important exception described below). In either case, the scale and shape parameters of the inverse gamma distribution are estimated based on the non-zero, positive eigenvalues. Note that the sample covariance matrix is always rank degenerate, *i.e.*, there are always multiple zero eigenvalues, regardless of the number of structures used in the calculation. Because of the nature of the superposition problem, and usually due to insufficient data, the sample covariance matrix is of maximum rank $K - 3$, and may even be less when there are few structures (rank = $\min(3N - 6, K - 3)$). We treat these missing eigenvalues as missing data, and in the EM algorithm the inverse gamma fit is based only on the $\min(3N - 6, K - 3)$ full rank, positive eigenvalues. When assuming that the covariance matrix is diagonal (no correlations), it is then necessary to omit the smallest three variances from the inverse gamma fit, as they are known *a priori* to be a zero-valued eigenvalues.

The algorithm cycles until convergence between the following two steps:

1. **Fit the inverse gamma parameters:** Find the ML estimates of the inverse gamma scale and shape parameters for the current eigenvalues. A maximum likelihood fit to an inverse gamma distribution can be accomplished by taking the inverse of each data point and fitting the transformed data to a gamma distribution (Evans *et al.*, 2000). For example, if $\mu_i = \lambda_i^{-1}$ for all positive eigenvalues, then the corresponding gamma probability distribution is:

$$\mathrm{P}\left(\mu_j\right) = \frac{\alpha^\gamma}{\Gamma(\gamma)}{\mu_j}^{(\gamma-1)}e^{-\alpha\mu_j}$$

   The ML estimates of the parameters α and γ are then solutions of the simultaneous equations:

$$\hat{\alpha} = \frac{\gamma}{\bar{\mu}}$$

$$y = \ln\hat{\gamma} - \psi_{(0)}(\hat{\gamma}) - \ln\bar{\mu} + \frac{1}{K'}\sum_j^{K'}\ln\mu_j = 0$$

   where K' is the number of nonzero eigenvalues, $\bar{\mu} = \sum_j^{K'}\mu_j/K'$, and $\psi_{(0)}$ is the digamma function. Newton's method can be used readily to solve the last equation above using its first derivative:

$$\frac{\partial y}{\partial\hat{\gamma}} = \frac{1}{\hat{\gamma}} - \psi_{(1)}(\hat{\gamma})$$

   where $\psi_{(1)}$ is the trigamma function (the first derivative of the digamma function). For the first iteration, useful starting values for inverse gamma parameters in the Newton method fit are given by the method of moments estimators, $\hat{\alpha} = \bar{\mu}/\phi$ and $\hat{\gamma} = \bar{\mu}^2/\phi$, where ϕ is the variance of the nonzero eigenvalues. In subsequent iterations, starting values in the Newton method are given by the parameter values from the previous iteration.
2. **Modify the sample eigenvalues:** Modify the sample eigenvalues according to Eq. **7**.

REFERENCES

Arnold, S. F. (1981) *The Theory of Linear Models and Multivariate Analysis*. Wiley, New York.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc B Met*, **39**, 1–38.

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

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

Evans, M., Hastings, N. and Peacock, J. B. (2000) *Statistical Distributions*. Wiley series in probability and statistics. Probability and statistics. John Wiley and Sons, New York, 3rd edition.

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

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

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

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

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

Shapiro, A., Botha, J. D., Pastore, A. and Lesk, A. M. (1992) A method for multiple superposition of structures. *Acta Crystallogr A*, **48**, 11–14.

Theobald, D. L. and Wuttke, D. S. (2006) Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem. *PNAS*, **in press**.