Optimal Superpositioning of Flexible Molecule Ensembles

Vytautas Gapsys and Bert L. de Groot*
Computational Biomolecular Dynamics Group, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

ABSTRACT Analysis of the internal dynamics of a biological molecule requires the successful removal of overall translation and rotation. Particularly for flexible or intrinsically disordered peptides, this is a challenging task due to the absence of a well-defined reference structure that could be used for superpositioning. In this work, we started the analysis with a widely known formulation of an objective for the problem of superimposing a set of multiple molecules as variance minimization over an ensemble. A negative effect of this superpositioning method is the introduction of ambiguous rotations, where different rotation matrices may be applied to structurally similar molecules. We developed two algorithms to resolve the suboptimal rotations. The first approach minimizes the variance together with the distance of a structure to a preceding molecule in the ensemble. The second algorithm seeks for minimal variance together with the distance to the nearest neighbors of each structure. The newly developed methods were applied to molecular-dynamics trajectories and normal-mode ensembles of the Aβ peptide, RS peptide, and lysozyme. These new (to our knowledge) superpositioning methods combine the benefits of variance and distance between nearest-neighbor(s) minimization, providing a solution for the analysis of intrinsic motions of flexible molecules and resolving ambiguous rotations.

INTRODUCTION

The identification of internal motions in large biomolecules is of key interest in the field of structural biology. The unambiguous identification of intrinsic molecular movements requires the rigorous removal of the external degrees of freedom. Although one can trivially eliminate translational motion by fixing the center of mass of a molecule, rotational superpositioning makes it necessary to define an objective function whose minimization results in the generation of a rotation matrix that optimally superimposes the molecules. For structured molecules, robust superpositioning algorithms are available. However, it is challenging to unambiguously determine the internal dynamics of flexible, unfolded, or intrinsically disordered proteins. When the objective minimizable function is defined as a sum of squares between two or more matrices, the problem of finding an orthogonal rotation matrix (or a set of matrices for an ensemble of structures) to reach the least-squares condition is termed an orthogonal Procrustes problem (1). A number of solutions for the orthogonal superpositioning of two matrices are available (see the review by Flower (2)) and can roughly be classified into iterative approaches (3–5), singular value decomposition (SVD)-based methods (6,7) and quaternion-based methods (8–11).

Shapiro and Botha (12) attributed the first SVD-based solution for the superpositioning of two vectors to von Neumann (13). In 1966, the solution was also found by Schönemann (14). In the field of molecular modeling, the SVD approach for structural fitting is usually referred to as Kabsch's algorithm (6,7). Multiple solutions for the superpositioning of two vectors based on the quaternion notation have been presented (8–11). It was demonstrated that the SVD- and quaternion-based methods are equivalent and, when correctly applied, yield identical results (15,16). The pairwise SVD-based solutions to the superpositioning problem were extended to fit multiple vector sets (17–21). At the heart of these approaches lies an iterative procedure for calculating an average structure over an ensemble and subsequently superpositioning onto it. Kearsley (22) and Diamond (23) proposed algorithms for multiple structure superpositioning based on quaternions.

Another branch of superposition methods uses least-squares superpositioning in combination with an adaptive selection of the atoms that participate in the calculation of the rotation matrix. These algorithms iteratively fit subsets of atoms to distinguish rigid core regions from the flexible parts of a protein (24–26). Damm and Carlson (27) used an iterative pairwise least-squares fit, applying different weights to the atoms, to identify and anchor the more-rigid parts of a structure. Theobald and Wuttke (28–30) proposed a maximum-likelihood-based fitting approach for superpositioning ensembles of structures. The method, termed Theseus, tries to find a combination of translational and rotational motions that will maximize the likelihood of observing a given ensemble, assuming that the structures in the ensemble follow Gaussian distribution in the absence of external degrees of freedom.

The analysis of trajectories of flexible molecules (as occurs, for example, in folding/unfolding transitions or natively disordered peptides) remains challenging due to two issues: 1), incomplete separation of internal and external degrees of freedom; and 2), ambiguous rotations of intrinsically similar structures.

The difficulty of superpositioning flexible peptides lies in the diversity of the conformations that the molecules can

Submitted August 3, 2012, and accepted for publication November 1, 2012.

*Correspondence: bgroot@gwdg.de

Editor: Michael Feig.

0006-3495/13/01/0196/12 $2.00

http://dx.doi.org/10.1016/j.bpj.2012.11.003

adopt. Although the removal of translational motions requires one to simply reset the center of mass of each molecule, rotational superposition is more complex. Having a diverse pool of structures renders it nontrivial to find an optimal reference frame on which to fit all the other molecules. As a consequence, highly similar structures may be rotated differently, thereby introducing artificial internal motion. Such artificial rotations manifest as large steps between two subsequent steps in a trajectory, and can be tracked in a projection on the principal components or in an analysis of the root mean-square deviation (RMSD) of structures that deviate from each other significantly more when superimposed on a common reference than when superimposed on each other, as illustrated in Fig. 1. Throughout this work, we refer to these rotations as suboptimal or ambiguous. One can resolve this drawback by progressively fitting the trajectories, i.e., superpositioning on a previous frame throughout the trajectory. Yet, the progressive superpositioning suffers from another issue: due to the rotation toward a constantly changing reference structure and accumulation of the numerical rounding errors, variance over the ensemble increases and a nonoptimal superpositioning occurs, again leading to artificial internal motions that in turn result in inaccuracies in subsequent analyses, such as estimations of conformational entropies.

In this work, we demonstrate the problems that occur during the superpositioning of flexible intrinsically disordered peptide ensembles. We then formulate an approach for optimal superposition based on the following two principles: 1), minimal variance (all motion that can be considered external is removed); and 2), RMSDs between pairs of structures in the superimposed ensemble should deviate minimally from an optimal, direct pairwise superposition of these structures. The atoms are weighted according to their masses, and no additional structure-dependent bias is introduced. Note that this results in a minimization of the overall rotation and does not guarantee that rigid subsections of a molecule will perfectly align (other methods have been designed for this purpose (24–27)). For the variance minimization, we rewrite Kabsch's derivation for the SVD-based least-squares optimization problem and arrive at an iterative algorithm that is identical to the solution of least-squares minimization over all pairs of members in an ensemble proposed by Ten Berge (19). To fulfill the second requirement of minimal deviation from an optimal pairwise superposition, we modify the previously obtained minimal variance solution. For arbitrary structure ensembles, we propose two different approaches: 1), a traveling-salesman-type trajectory rearrangement; and 2), nearest-neighbor-based variance minimization superpositioning. We compare our method to an essentially different maximum-likelihood-based superpositioning approach. The new algorithms are designed to be specifically suited for large numbers of structurally diverse molecules, such as in molecular dynamics (MD) and Monte Carlo (MC) trajectories of disordered peptides. The approaches presented here do not require the definition of a single reference structure or a superposition region within the molecule, and thus provide an unbiased way to remove rotational degrees of freedom

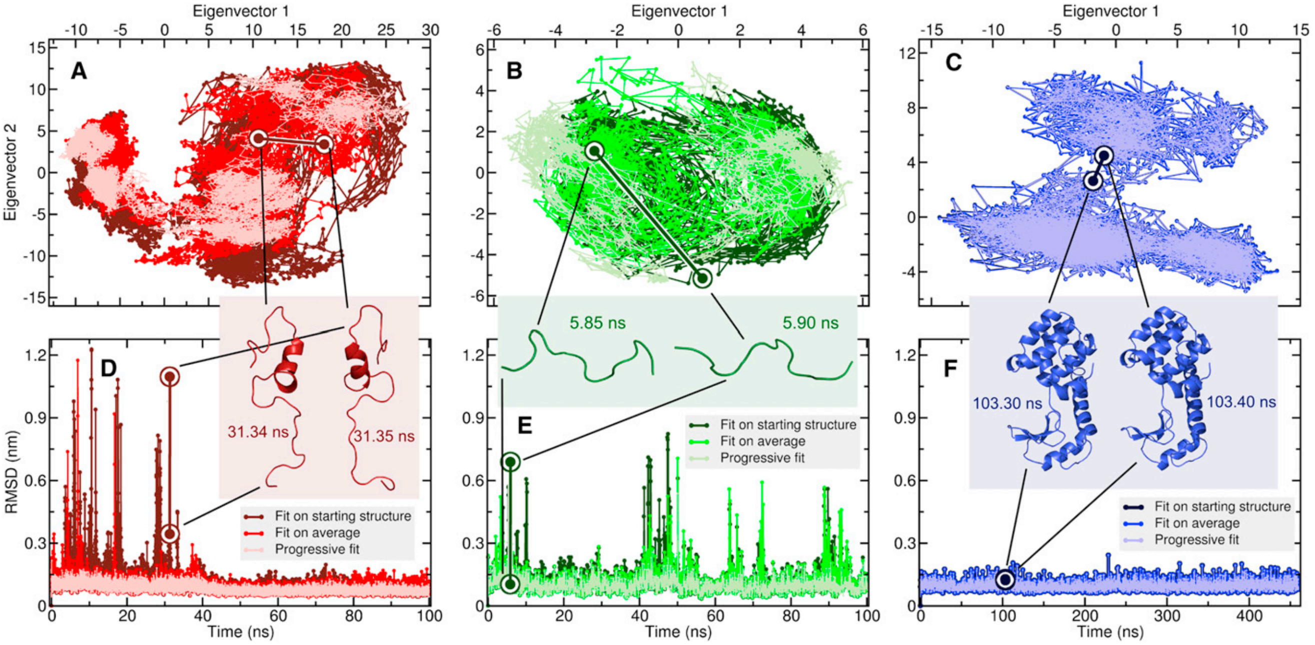

FIGURE 1 PCA and RMSD analyses illustrating ambiguous rotations during superpositioning. (*A–F*) Ensembles of the Aβ peptide (*A* and *D*), RS peptide (*B* and *E*), and lysozyme (*C* and *F*). (*A–C*) Projections of the MD trajectories on the eigenvectors with the largest eigenvalues. (*D–F*) RMSDs calculated between a structure at time τ and $\tau - 1$. Suboptimal superpositions result in large RMSD values despite structural similarity.

from a structural ensemble. The proposed methods are demonstrated to provide a robust framework for the quantitative analysis of the internal dynamics of flexible molecules.

THEORY

Variance minimization: min(Var)

A number of procedures for ensemble superpositioning based on least-squares minimization are well established (18–21). In this work, we first derive the solution for iterative ensemble superposition based on the convenient matrix decomposition approach introduced by Kabsch (6,7). This then allows us to modify the method for an optimal ensemble superposition, as discussed in the following sections.

We define our problem as a minimization of the function E:

$$E = \frac{1}{2T}\sum_{n}\sum_{\tau} w_n\left(U_\tau \mathbf{x}_{n\tau} - \frac{1}{T}\sum_{t} U_t \mathbf{x}_{nt}\right)^2 \tag{1}$$

where $\mathbf{x}_{nt}$ is a vector of coordinates of an atom n at time t, U_t is a rotation matrix applied to the system at time t, w_n is an atom-specific weight (which in our analysis corresponds to the atomic mass), and T is the total number of frames in an ensemble. The term $1/T\sum U_t\mathbf{x}_{nt}$ serves as a reference for superpositioning every structure in an ensemble. Following the derivation by Kabsch, we introduce a constraint in terms of a Lagrange multiplier to keep the rotation matrix orthogonal. Because we are dealing with an ensemble of structures, the constraint is defined for a rotation matrix at every frame τ:

$$F = \frac{1}{2T}\sum_{\tau}\sum_{i,j} l_{ij\tau}\left(\sum_{k} u_{ki\tau}u_{kj\tau} - \delta_{ij}\right) \tag{2}$$

where $u_{ki\tau}$ are the elements of a rotation matrix U_τ, l_{ij} are the elements of a Lagrange multiplier matrix L_τ and δ_{ij} are the elements of an identity matrix. Subsequently, we minimize $E+F$ by setting the derivative $\partial(E+F)/\partial u_{ij}=0$ and requiring the second derivative $\partial^2(E+F)/(\partial u_{mk}\partial u_{ij})$ to be positive definite, to obtain a solution in the form of

$$U_\tau\left(S_\tau - \frac{1}{T}S_\tau + L_\tau\right) = Q_\tau \tag{3}$$

where $q_{ij\tau} = 1/T\sum_n w_n x_{nj\tau}\sum_{t\neq\tau}\sum_k u_{ikt}x_{nkt}$ and $s_{ij\tau} = \sum_n w_n x_{ni\tau}x_{nj\tau}$.

When $t>2$, no closed-form solution to this problem exists (21); therefore, from this point, a numerical procedure is required. The SVD approach applied by Kabsch for the pairwise superpositioning can be used in an iterative algorithm, where at each iteration a Q_τ matrix is calculated using the rotation matrix ensemble U_τ as determined to that point. Various applications of similar algorithms have been used in the past. In the approach proposed by Kristof and Wingersky (17), an average matrix is calculated once every cycle and all the other matrices are fitted onto it, after which the average matrix is recalculated and the procedure is iterated. The same approach was also described by Gower (18) and Sutcliffe et al. (20).

Ten Berge (19) suggested an improved version of the Kristof and Wingersky algorithm wherein the reference frame (Q_τ) is updated after every update of a rotation matrix. More importantly, by minimizing the sum of squares over all pairs of matrices, Ten Berge showed that the reference matrix Q_τ must not contain the contribution from the $U_\tau\mathbf{x}_t$ structure. In our derivation, we arrived at the same result by minimizing variance over the ensemble. Further, we employed Ten Berge's iterative procedure to minimize the variance over an ensemble.

Progressive fitting and variance minimization: min(Var+Prev)

When variance minimization alone is used, deviations from optimal, direct pairwise superposition solutions can occur. Hence, we propose a modification of the variance minimization algorithm by combining it with the progressive fit approach. The function to be minimized is written as follows:

$$E_\tau = \frac{1}{2T}\sum_{n} w_n\left[\omega_\tau\left(U_\tau\mathbf{x}_{n\tau} - U_{\tau-1}\mathbf{x}_{n\tau-1}\right)^2 + \sum_{\tau}\left(U_\tau\mathbf{x}_{n\tau} - \frac{1}{T}\sum_{t}U_t\mathbf{x}_{nt}\right)^2\right] \quad \forall\tau>1 \tag{4}$$

where $\omega_\tau = e^{-RMSD^{optimal}_{\tau-1}}$. $RMSD^{optimal}_{\tau-1}$ is an RMSD value between a structure at time τ and a preceding frame $\tau-1$ after the τ and $\tau-1$ structures are superpositioned directly onto each other. This value is precalculated before the iterative minimization procedure is started. The weight ensures that only similar structures may contribute to the determination of the rotation of a structure τ. This is required, because superpositioning of a structure onto a very different frame may result in an ambiguous rotation. The weight for the ω_τ for the first structure is set to zero.

The construction of the function E_τ enables each frame, except for the first one, to be rotated such that the variance and the RMSD to a previous structure are minimized. By imposing the constraints that keep the rotation matrices U_τ orthogonal (Eq. 2), and calculating the derivative of the E_τ with respect to the U_τ, we get an expression similar to Eq. 3:

$$U_\tau\left(S_\tau\left(1-\frac{1}{T}+\omega_\tau\right)+L_\tau\right) = Q_\tau+\omega_\tau R_\tau \qquad (5)$$

where $r_{ij\tau}=\sum_n w_n x_{nj\tau}\sum_k u_{ik\tau-1}x_{nk\tau-1}$ is an outer product of a structure at time τ with a preceding in time structure that has already been rotated in a previous step.

This formulation of the minimization problem aims to reduce the variance over the ensemble by removing as much of rotational motion as possible, and at the same time resolves ambiguous superpositions with the previous frames in a trajectory, i.e., it ensures that similar structures are superimposed similarly, deviating minimally from a direct, pairwise superposition. We apply an iterative procedure to minimize the function in Eq. 4 similarly to the min(Var) algorithm: at every iteration, a reference frame for superpositioning is defined as a combination of the average over an ensemble and a previous frame, weighted by a distance to it. Following Ten Berge's method, the average is updated after every rotation.

An important feature of the min(Var+Prev) approach is that it resolves suboptimal rotations for the ensembles in which the structures are ordered such that two subsequent frames are sufficiently similar to each other to provide an unambiguous reference for an optimal pairwise superposition. Uninterrupted MD trajectories fulfill this requirement, whereas ensembles retrieved by MC or another stochastic sampler (e.g., CONCOORD (31)) will not benefit from the min(Var+Prev) method without additional trajectory preprocessing. One way to restructure a trajectory such that two subsequent, directly superpositioned frames have a suitably low RMSD is presented in the Supporting Material as a Hamiltonian path minimization problem, for which the solution to the traveling-salesman problem (TSP) was applied (32,33).

Nearest Neighbors: min(Var + NN)

Trajectory rearrangement by minimizing the Hamiltonian path together with the min(Var+Prev) algorithm is one way to solve the problem of suboptimal rotations of similar structures. Here we also propose another method, termed min(Var+NN), that does not require a specific ordering of an ensemble. Instead of using only a previous frame in a trajectory as a modification to the min(Var) algorithm, we consider a subensemble of the nearest neighbors for each member of an ensemble. The function to be minimized is written as

$$E_\tau=\frac{1}{2T}\sum_n w_n\left[\sum_{NN_\tau}\omega_{NN_\tau}\left(U_\tau\mathbf{x}_{n\tau}-U_{NN_\tau}\mathbf{x}_{nNN_\tau}\right)^2+\sum_\tau\left(U_\tau\mathbf{x}_{n\tau}-\frac{1}{T}\sum_t U_t\mathbf{x}_{nt}\right)^2\right]\ \forall\tau\in[1;T] \qquad (6)$$

where $\Sigma_{NN\tau}$ denotes summing over the nearest neighbors of the structure τ. The weight is defined as $\omega_{NN_\tau}=e^{-RMSD^{optimal}_{NN_\tau}}$. $RMSD^{optimal}_{NN_\tau}$ is the RMSD value obtained after superpositioning the structure τ and a member NN_τ of the ensemble of its nearest neighbors. These values are precalculated before the minimization is started. In analogy to the construction of the weight for the min(Var+Prev) algorithm, ω_{NN_τ} ensures that similar structures contribute to the superpositioning most, because they allow for the most unambiguous superposition.

Minimization of the function E_τ requires an iterative solution of the equation similar to Eq. 5:

$$U_\tau\left(S_\tau\left(1-\frac{1}{T}+\sum_{NN_\tau}\omega_{NN_\tau}\right)+L_\tau\right) = Q_\tau+\sum_{NN_\tau}\omega_{NN_\tau}P_{NN_\tau} \qquad (7)$$

where $p_{ij\tau}=\sum_n w_n x_{nj\tau}\sum_k u_{ikNN_\tau}x_{nkNN_\tau}$. Equation 7 is derived in a manner similar to that used for Eqs. 3 and 5, by constraining rotation matrices to be orthogonal and calculating a derivative of the function E_τ.

We applied an iterative procedure based on Ten Berge's approach to minimize the function min(Var+NN). We observed a dependence of the iterative algorithm for the function in Eq. 7 on the initial rotations of the structures. The effect becomes more pronounced for larger numbers of nearest neighbors considered. To obtain consistent results from the min(Var+NN), we implemented a step of progressive superpositioning of the trajectory before starting the iterative procedure, thus directing the algorithm to converge to a certain local minimum and ensuring reproducibility of the results.

The analysis of convergence of the min(Var), min(Var+Prev), and min(Var+NN) algorithms is presented in the Supporting Material.

The min(Var+NN) approach requires one to precalculate the RMSD values for the pairwise superimposed structures over the whole structural ensemble before starting the iterative procedure. To alleviate the computational complexity of this step, we employed a rapid RMSD calculation approach based on the quaternion notation (34). Additionally, we parallelized the pairwise RMSD calculation step to enable running it on a user-defined number of threads.

MATERIALS AND METHODS

Analyzed systems

To illustrate the difficulty of superpositioning multiple flexible, intrinsically disordered peptides, we used MD ensembles of the 40-amino-acids-long Aβ peptide and the 15-residues-long RS peptide. The trajectory of the Aβ peptide was 100 ns in length and consisted of 10,001 frames, whereas the RS peptide's trajectory was created by concatenating two independent MD runs with different starting conformations, each 50 ns long, and consisted of 2005 frames. Additionally, a T4 lysozyme ensemble (460.1 ns, 4601 frames) was used to allow the comparison of flexible peptide fitting with the superpositioning of an enzyme with a well-defined secondary and tertiary structure. Detailed information about the simulation parameters used to generate the ensembles of the Aβ and RS peptides is provided in the Supporting Material. The simulation parameters for the lysozyme trajectory have been described previously (35). For all of the superpositioning analyses described in this work, we removed the translational degrees of freedom by setting the center of mass of each structure in an ensemble to the origin of the coordinate axes.

Superpositioning of flexible peptides

Throughout the work for all of the superpositioning cases of the MD trajectories, we calculated the mass-weighted variance over each ensemble considering the Cα atoms of the structures. To avoid ambiguities that could arise due to rounding errors when using different formulas for the variance estimation, here and in all subsequent analyses we calculated the variance as a trace of a covariance matrix as implemented in the *g_covar* tool in the Gromacs (36) package. For all iterative schemes, 1000 iterations were carried out to ensure convergence.

Analysis of the local neighborhood

To quantify the orientation of a structure to the structurally similar members of an ensemble, we analyzed the local neighborhood for every frame in a trajectory. The local neighborhood of a structure is represented by its nearest neighbors that are identified by the pairwise superimposition of a molecule of interest onto all the structures in the ensemble. The procedure is performed for every member in an ensemble, and the sum of the RMSDs over the nearest neighbor subensemble is calculated for every structure. The RMSD sum yields an estimate of an optimal rotation of a structure with respect to its local neighborhood. After superpositioning of the whole ensemble, every structure is expected to be oriented such that the RMSD to its nearest neighbors is as close to the optimum value as possible. This consideration generalizes the progressive RMSD analysis for the local neighborhood of the ensemble members. We constructed two estimators, RMSD based and variance based, as described in detail in the Supporting Material.

Normal-mode ensembles: comparison with maximum-likelihood-based superpositioning

Structural ensembles that contain no translational and rotational motions by construction were generated as a reference case. For this purpose, we energy minimized the structures and performed normal-mode (NM) analysis for each of the studied proteins. Details of the procedure are provided in the Supporting Material.

To illustrate the importance of superpositioning for assessing the flexible structure ensembles, we calculated the absolute conformational entropies of the nonfitted NM ensembles for the Aβ peptide, RS peptide, and lysozyme using Schlitter's formula (37). These entropies were used as a reference for comparison with the superpositioned ensemble entropies calculated in the same way.

The principal motions of lysozyme were analyzed to gain a deeper insight into how the different superpositioning methods affect the perception of protein dynamics. The MD and NM ensembles of lysozyme were superimposed with both the min(Var) and maximum-likelihood-based methods, considering the Cα atoms only. The superimposed structures were subjected to a principal component analysis (PCA). We depicted the characteristic motions by projecting extreme conformations along the eigenvector with the largest eigenvalue and interpolating between the extremes.

RESULTS

Superpositioning of flexible peptides

Three cases of superpositioning are illustrated in Fig. 1 for three different systems: Aβ, RS peptide, and lysozyme. In the projections on the two principal components with the largest eigenvalues, some large steps between two subsequent frames are highlighted for the Aβ (Fig. 1 *A*) and RS (Fig. 1 *B*) peptides when superpositioning is performed on the initial structure. The suboptimal rotations can also be observed in the plots of the RMSDs calculated between subsequent frames that show occasional high RMSD values between adjacent frames (Fig. 1, *D* and *E*). A clear indication that such high RMSD values are based on a superpositioning artifact is the fact that the RMSD for such structure pairs is significantly lower when these pairs are directly superimposed onto each other. Once the peptides were superimposed onto the average structures of the respective ensembles, some of the spikes in the RMSD plots were removed or decreased, indicating that superpositioning artifacts were partially resolved. Consistent with this result, the progressively superimposed structures also exhibit much smoother RMSD profiles (*bright curves* in Fig. 1, *D* and *E*). The spike in the RS peptide RMSD plot at ~50 ns remains even for the progressively superimposed ensemble, and marks the position where two independent trajectories were concatenated. Fig. 1, *C* and *F*, show the PCA plots and RMSDs for the lysozyme trajectory. Apparently, no superpositioning artifacts affect the ensemble even when superpositioning is performed on a single starting structure. The RMSD traces match almost perfectly, independently of the superpositioning method used. For lysozyme, which has a well-defined secondary and tertiary structure, the reference frame shares a lot of similarity with any other frame in the trajectory, and therefore fitting is less ambiguous.

The variance values in Table 1 reveal that superpositioning on an initial structure and progressive fit yield larger variances over an ensemble than superpositioning onto an average structure. Hence, even though progressive fit trajectories removed large RMSD values for neighboring ensemble members (Fig. 1, *D* and *E*), the variance over the trajectory accumulated to a value significantly larger than that obtained with other superpositioning methods.

Progressive fitting and variance minimization: min(Var+Prev)

The iterative procedure of variance minimization, min(Var), reduces the variance over the ensembles of the Aβ and RS peptides (Table 1), whereas in the case of lysozyme, the variance could not be reduced significantly as compared with fitting onto an average structure. Interestingly, a minimal variance of a trajectory does not guarantee removal of the suboptimal rotations, as can be seen for both the Aβ and RS peptide ensembles in Fig. 2. The result of the

TABLE 1 Mass-weighted variances (nm^2u) over the MD ensembles

					min(Var+Prev)		
Structure	Fit on starting structure	Fit on average structure	Progressive fitting	min(Var)	MD	TSP	min(Var+NN)*
Aβ peptide	183.287	144.895	171.919	143.745	144.067	144.075	146.044
RS peptide	23.696	21.842	30.762	21.268	21.603	21.669	23.362
Lysozyme	65.231	65.176	65.956	65.172	65.175	65.174	65.186

*Fifty nearest neighbors were considered.

superpositioning method combining the progressive fit with the variance minimization, min(Var+Prev), is illustrated by the bright-colored curves in Fig. 2. The ambiguous rotations with respect to the previous frame in each trajectory were resolved, as can be inferred from the absence of spikes in the RMSD profiles. The removal of the rotational artifacts comes at the cost of a slight increase in the variance over the ensemble (Table 1). For the Aβ and RS peptide ensembles, the rise in the variance was marginal, and the lysozyme trajectory shows almost the same variance for the min(Var) and min(Var+Prev) algorithms.

Nearest neighbors: min(Var + NN)

In case trajectory rearrangement is not desired, we propose another approach to resolve ambiguities in the superpositioning. The min(Var+NN) algorithm is a generalization of the min(Var+Prev) method: instead of using a single previous frame to modify min(Var) fitting, we construct the reference for every member of an ensemble by taking its nearest neighbors into account. The influence of the number of nearest neighbors is analyzed in depth in the Supporting Material. As could be expected, larger NN numbers increase the variance over an ensemble and resolve ambiguous rotations with respect to a preceding structure. However, using large numbers of nearest neighbors forces the algorithm to behave similarly to the min(Var) approach: the variance starts decreasing and ambiguous rotations reoccur. Hence, our observations suggest that a smaller NN number (up to 100) is sufficient to remove suboptimal rotations in an ensemble.

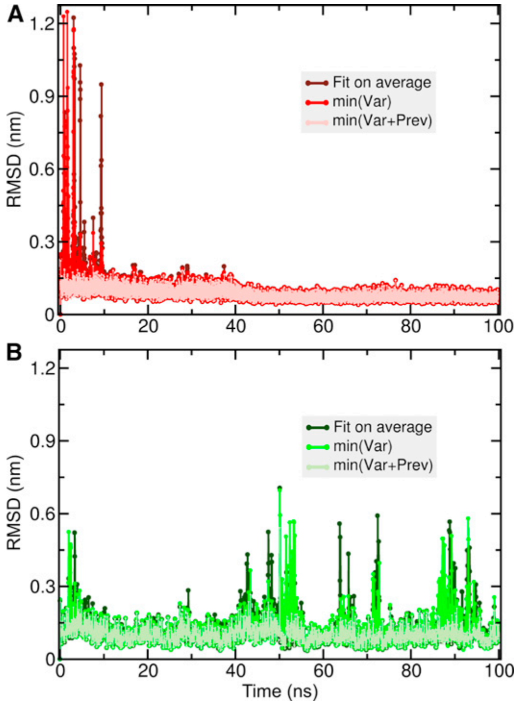

FIGURE 2 RMSD from a previous frame analysis after min(Var) and min(Var+Prev) superpositioning. (A and B) The original MD ensembles for the Aβ (A) and RS (B) peptides.

Including the nearest neighbors for the superpositioning makes it interesting to not only compare a structure with its preceding frame, as we have done so far, but to also consider the RMSDs of a local neighborhood. In fact, analysis of a local neighborhood of each member of an ensemble can be seen as a generalization of the RMSD to a preceding structure analysis. We calculated the RMSD matrices for the Aβ peptide (Fig. 3) and the RS peptide (Supporting Material) after applying different fitting methods to the trajectories. Superpositioning with the min(Var+Prev) and min(Var+NN) algorithms significantly smoothens the RMSD landscape in the local neighborhood of a structure. A smoother RMSD surface indicates that the suboptimal rotations, which we previously depicted as spikes in the RMSD plots (Figs. 1 and 2, and Supporting Material), are removed not only for the subsequent frames but also for the structures in the local neighborhood, i.e., subensembles of similar conformations. The min(Var+NN) algorithm used 50 nearest neighbors to construct a reference for the superpositioning, which allowed a strong reduction of the RMSDs among the similar structures (Fig. 3 *F*). However, the resolved ambiguous rotations in the local neighborhood come at the cost of an increased overall variance over the ensembles (last column in Table 1).

Analysis of the local neighborhood

We formulated the requirements for an optimal global ensemble superposition as 1), a minimal variance over an ensemble; and 2), a minimal deviation from an optimal pairwise superposition. Thus far, we have provided a quantitative estimate of variance for different fitting algorithms (Table 1). To quantify the second criterion, we analyze local neighborhoods over an ensemble, because the pairwise

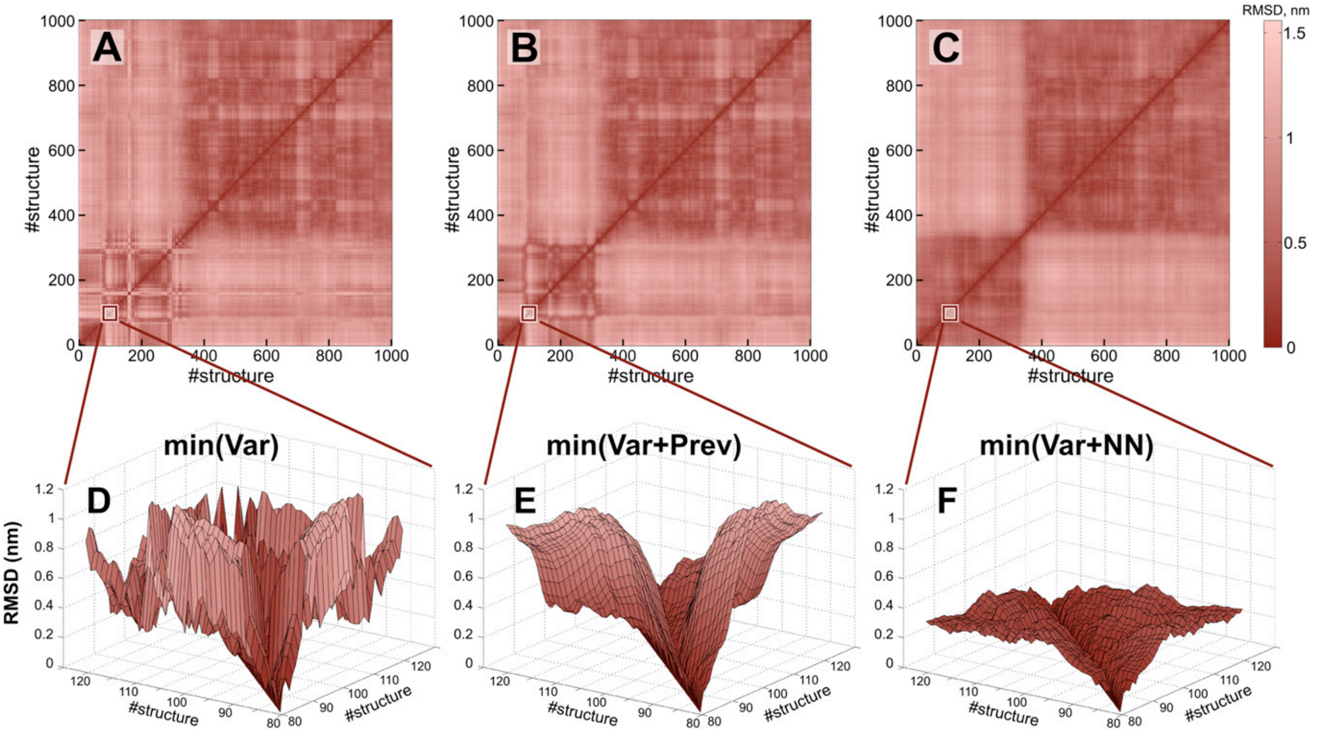

FIGURE 3 Pairwise RMSD matrices and surfaces of the Aβ peptide ensemble. (*A–C*) RMSD values after the min(Var) (*A*), min(Var+Prev) (*B*), and min(Var+NN) (*C*) superpositioning. (*D–F*) Excerpts from the matrices shown as surfaces illustrate the smoothening effect of the min(Var+Prev) and min(Var+NN) algorithms.

superposition is unambiguously defined for such pairs and hence serves as a suitable reference. As described in the Materials and Methods section, we constructed two estimators, RMSD based and variance based, defining the requirements of an optimal superposition method to quantify the quality of the fitting algorithms. This enabled us to compare the methods by mapping them in space defined by both estimators (Fig. 4).

The optimal solution, having low variance and small deviation from the optimal pairwise superposition in the local neighborhood, should be located in the lower-left quadrant in the graphs in Fig. 4. The min(Var+NN) (with NN = 50) and min(Var+Prev) algorithms fulfill both requirements, whereas the min(Var) approach has a higher RMSD value. An exception is the RS peptide case (Fig. 4 *B*), where min(Var) mildly outperforms min(Var+Prev). This effect comes from the fact that the RS peptide's trajectory is constructed by concatenating two independent MD ensembles, and thus taking into account the preceding structure for an ensemble superpositioning may not lead to the RMSD reduction in the local neighborhood. Superpositioning on an initial or average structure yields larger variance and deviations from an optimal pairwise superpositioning than could be achieved with the methods introduced here. It is interesting to note that progressive superpositioning not only results in large variances but, in the case of lysozyme, also exhibits a stronger deviation from the optimal RMSD in the local neighborhood than the other methods. A local neighborhood of 10 nearest neighbors was considered for the analysis presented in Fig. 4. The analysis of a local neighborhood consisting of different NN numbers is provided in the Supporting Material.

NM ensembles: comparison with maximum-likelihood-based superpositioning

An unbiased way to evaluate the performance of superpositioning algorithms is to apply the fitting procedure to an ensemble that by construction contains only internal fluctuations and therefore serves as reference for an optimal superposition procedure. Removal of the external degrees of freedom may introduce bias into the description of the internal dynamics. To generate a reference ensemble without external degrees of freedom, we performed an NM analysis for the Aβ peptide, RS peptide, and lysozyme. Ordering the eigenmodes of the Hessian matrices in an ascending manner by their corresponding eigenvalues revealed that the first six eigenvalues, corresponding to the translational and rotational motions (38), are equal to zero to machine precision, for all of the protein systems analyzed. Hence, the ensembles generated by sampling the seventh mode fulfill the requirement of having no external degrees of freedom. In Table 2, the variances over the superpositioned NM ensembles are summarized for the

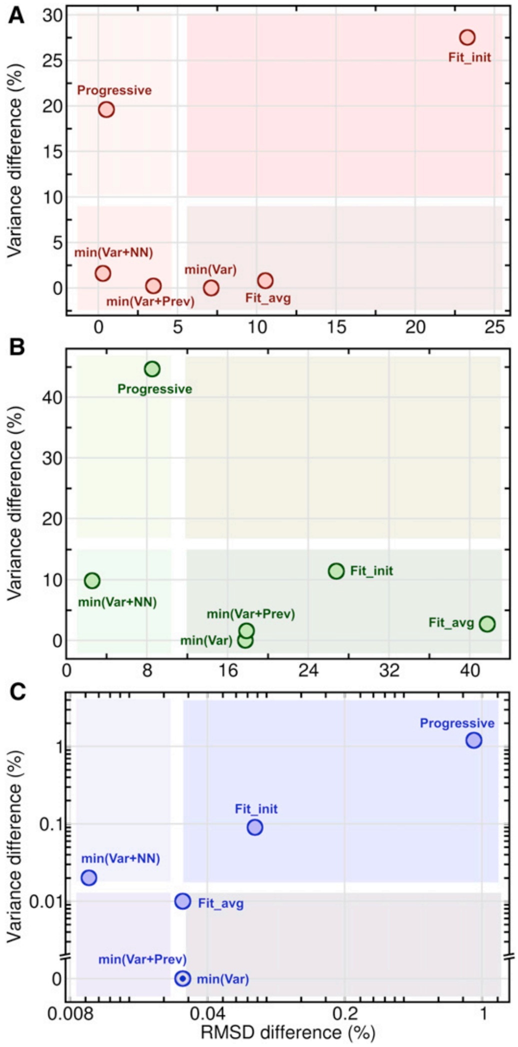

FIGURE 4 Mapping of the superpositioning algorithms into a common space defined by the ensemble variance and local neighborhood RMSD. (*A–C*) MD ensembles of the Aβ (*A*) and RS (*B*) peptides, and lysozyme (*C*) were used for the analysis. The local neighborhood of 10 nearest neighbors was considered for the RMSD calculation. The background colors of the quadrants are to guide the eye only and do not represent any calculated result.

various fitting algorithms. Fitting on the initial structure and progressive superpositioning left the variance over the Aβ peptide's ensemble unaltered in comparison with the nonfitted ensemble, whereas a slight decrease was observed for the RS peptide. The min(Var), min(Var+Prev), and min(Var+NN) approaches rotated the structures such that the variance for the ensembles of both peptides remained close to the nonfitted ensemble. The least-squares-based superpositioning practically had no influence on the variance of the lysozyme's NM ensemble. The maximum-likelihood approach Theseus introduced additional variance for all three ensembles.

Conformational entropy calculations based on the variances along the principal components of the ensembles (Schlitter's formula (37)) revealed a strong effect of the different fitting approaches (Table 3). For the flexible Aβ and RS peptides, superpositioning on the initial structure, as well as the iterative procedures, provided entropy values close to those of the reference ensembles. The progressive superpositioning significantly overestimated the conformational entropy of the RS peptide, but not that of the Aβ peptide. Interestingly, the min(Var+Prev) method returned the conformational entropy value for the RS peptide closer to the reference estimate. The estimation of the entropy in the case of lysozyme was only slightly affected by the least-squares-based superpositioning. Similarly to the variance estimation, the Theseus fitting increased the conformational entropy of all ensembles considered in the analysis. However, it should be noted that part of the increase in entropies may be attributed to the level of precision at which Theseus stores the trajectories. The algorithm is based on the PDB file format, allowing a precision of 10^{-13} m, whereas the other algorithms used a precision of 10^{-15} m. Conformational entropy estimates for the MD ensembles are provided in the Supporting Material.

The principal motions of lysozyme and therefore the interpretation of internal dynamics differ significantly depending on the superpositioning algorithm applied (Fig. 5). The motion along the first eigenvector of the nonsuperimposed NM ensemble (Fig. 5 *A*) matches the motion of a fitted min(Var) ensemble (Fig. 5 *B*). The maximum-likelihood approach represents the motion differently (Fig. 5 *C*): half of the protein is kept rigid, whereas the other part undergoes a large conformational transition. The same (albeit more pronounced) effect is observed for the characteristic motion extracted from the MD trajectory (Fig. 5, *D* and *E*). A more quantitative analysis of the principal motions of lysozyme in terms of the root mean-square fluctuations, as well as the eigenvalue spectra, is provided in the Supporting Material.

The NM ensembles were also generated for the stromal-cell-derived factor-1 (PDB ID: 2SDF (39); model 1), which was used by Theobald and Wuttke (30). Following their approach, we calculated the correlation matrices between the atoms in the ensemble after applying min(Var) and Theseus superpositioning (Fig. 6, *A–C*). The nonfitted NM ensemble served as a reference, because by definition it contains no external degrees of freedom. It is evident that the variance-minimizing fitting has a smaller influence on the correlation matrix pattern than the maximum-likelihood based approach. Additionally, we evaluated the variance of the Cα atoms in the ensemble (Fig. 6, *D–F*), which shows a significant change in the atom flexibility when the Theseus algorithm is applied.

DISCUSSION

An iterative approach for finding a set of rotation matrices to minimize the pairwise sum of squares over an ensemble

TABLE 2 Mass-weighted variances (nm^2u) over the NM ensembles

Structure	No fit	Fit on first structure	Progressive fit	min(Var)	min(Var+Prev)	min(Var+NN)*	Theseus
Aβ peptide	5.187	5.187	5.187	5.187	5.187	5.187	5.422
RS peptide	8.657	8.656	8.657	8.656	8.656	8.656	9.883
Lysozyme	6.033	6.033	6.033	6.033	6.033	6.033	6.879

*Fifty nearest neighbors were considered.

of vectors is well established (17–20). We reformulated the problem into the variance minimization over a trajectory and arrived at the known SVD-based solution by following Kabsch's derivation (6). Ten Berge's iterative procedure (19) was found to be suitable for the variance minimization, and provided rapid convergence even for large structural ensembles of highly flexible molecules. Although the results achieved with the iterative min(Var) algorithm meet the requirement of minimal variance over an ensemble, some rotations may position similar molecules differently with respect to each other, as illustrated in Fig. 1. This effect is not a malfunction of the algorithm or a product of an improper rotation. Rather, it is a result of ambiguous superpositioning of structures onto an unsuitable reference frame, which for the case of variance minimization is represented by an average structure. Removal of all the rotational motion from an ensemble by means of variance minimization is suitable for a semirigid body (e.g., structured enzyme lysozyme), whereas in the case of highly flexible peptides, some rotational motion bears the character of an intrinsic motion. Removal of such motions results in ambiguous rotations, whereas the ambiguities may be resolved by reintroducing the rotations and returning some variance to an ensemble. The superpositioning method may have an effect on the free-energy landscape of a structural ensemble. Although the rotation of a molecule does not change its conformation, different schemes for separating internal from overall motion may affect the analysis of internal dynamics. For example, when PCA is used to analyze the energy landscape of a trajectory, the superpositioning method will influence not only the projection itself (and therefore the projected phase space densities) but also the eigenvectors of the covariance matrix, thereby changing the whole coordinate system for the projected ensemble.

A reformulation of the problem of optimal removal of the external degrees of freedom resulted in the min(Var+Prev) and min(Var+NN) algorithms, yielding solutions that both minimize the variance and resolve the ambiguous rotations. The min(Var+Prev) is suited for ensembles ordered such that the neighboring frames share structural similarity. Although MD trajectories fulfill this requirement, for other types of trajectories we suggest an ordering approach based on the solution to the TSP. An interesting side observation from the application of the TSP algorithm to the MD trajectories was the reordering of the sequential-in-time ensembles into the new trajectories with even shorter Hamiltonian paths than the original time ordering. This possibly has implications for ensemble clustering and convergence checks. Another method we presented, min(Var+NN), alleviates the need for ensemble reordering by precalculation of the local neighborhood for each member of an ensemble. In this approach, the size of the neighborhood to be smoothened, in terms of a pairwise RMSD between the structures, can be chosen freely. We assessed the performance of the new superposition approaches using large structural ensembles, and thus demonstrated the ability of our methods to remove external degrees of freedom efficiently when an average structure is smeared out over a long trajectory.

Mapping different superpositioning methods in the space defined by two estimators, the variance over an ensemble and the RMSD of the local neighborhood, revealed the inherent complexity of the removal of external degrees of freedom from the structural ensembles. Satisfying one of the constraints results in a strong violation of the other. For example, the progressive superposition closely matches the optimal pairwise superposition, but it leads to a significant increase of the variance over an ensemble. Also, minimizing the variance does not guarantee that similar structures will be rotated in a similar way. The approaches introduced here optimize both criteria, removing as much external rotation as possible and ensuring a minimal deviation from an optimal pairwise superposition.

In contrast to the superposition algorithms that assign different weights to the atoms of structures being superimposed (24–27), our methods weigh the atoms according to their mass, and assign a unique reference frame for each

TABLE 3 Conformational entropies (J/mol K) of the NM ensembles

Structure	No fit	Fit on first structure	Progressive fit	min(Var)	min(Var+Prev)	min(Var+NN)*	Theseus
Aβ peptide	61.925	61.828	61.820	61.827	61.836	61.840	65.968
RS peptide	52.589	52.914	57.794	52.924	53.007	52.910	76.682
Lysozyme	123.217	123.065	123.124	123.055	123.077	123.147	127.166

*Fifty nearest neighbors were considered.

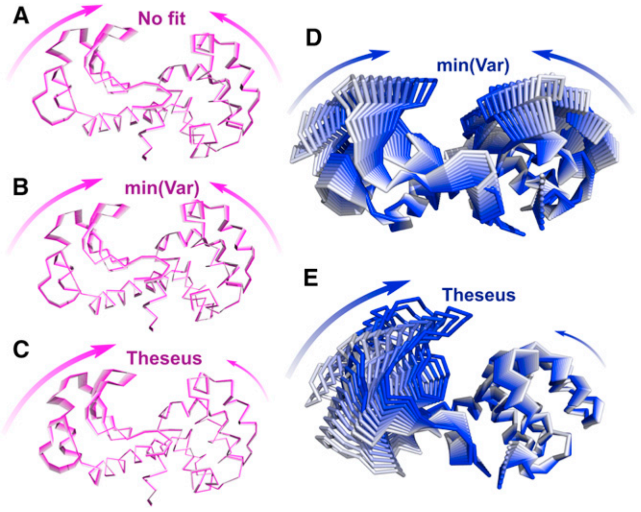

FIGURE 5 Principal motions of lysozyme. (*A–C*) Principal motions extracted from the NM ensembles. (*D* and *E*) Principal motions from the MD ensembles. (*A*) Nonsuperimposed NM ensemble. (*B* and *D*) min(Var) superpositioned ensembles. (*C* and *E*) Theseus superpositioned ensembles.

structure. The former approach allows one to predict the dynamics from the superposition of several structures by elucidating the rigid and flexible parts of a protein. In contrast to that approach, our methods do not bias the rotations according to the assumed intrinsic motions of a molecule; rather, they remove the external degrees of freedom, leaving the interpretation of the dynamics to further analysis. This makes the min(Var+Prev) and min(Var+NN) particularly suitable for large structural ensembles and their analysis in terms of internal fluctuations, such as by PCA (38,40,41).

Ensembles of multiple structures can be superpositioned based on a maximum-likelihood approach such as Theseus (28–30). Using the NM ensembles of the lysozyme and stromal-cell-derived factor-1, as well as the lysozyme's MD trajectory, we demonstrated that the maximum-likelihood approach may lead to a significantly different interpretation of the internal protein motions. The nonsuperimposed NM ensembles in our analysis served as a reference without any external degrees of freedom. By comparing the motions extracted from the ensembles treated with the different fitting procedures, we observed that the least-squares-based methods yielded results in better agreement with the reference ensembles than the Theseus algorithm. From our analysis, we conclude that for large structural ensembles, such as molecular or stochastic dynamics, NM, CONCOORD trajectories, where the full removal of the external degrees of freedom is required, least-squares-based superpositioning provides a robust and minimal basis for the analysis of internal dynamics. Maximum-likelihood-based superpositioning finds its strength in fitting NMR-like ensembles, where the implicit prediction of rigid and flexible regions of a protein is desired.

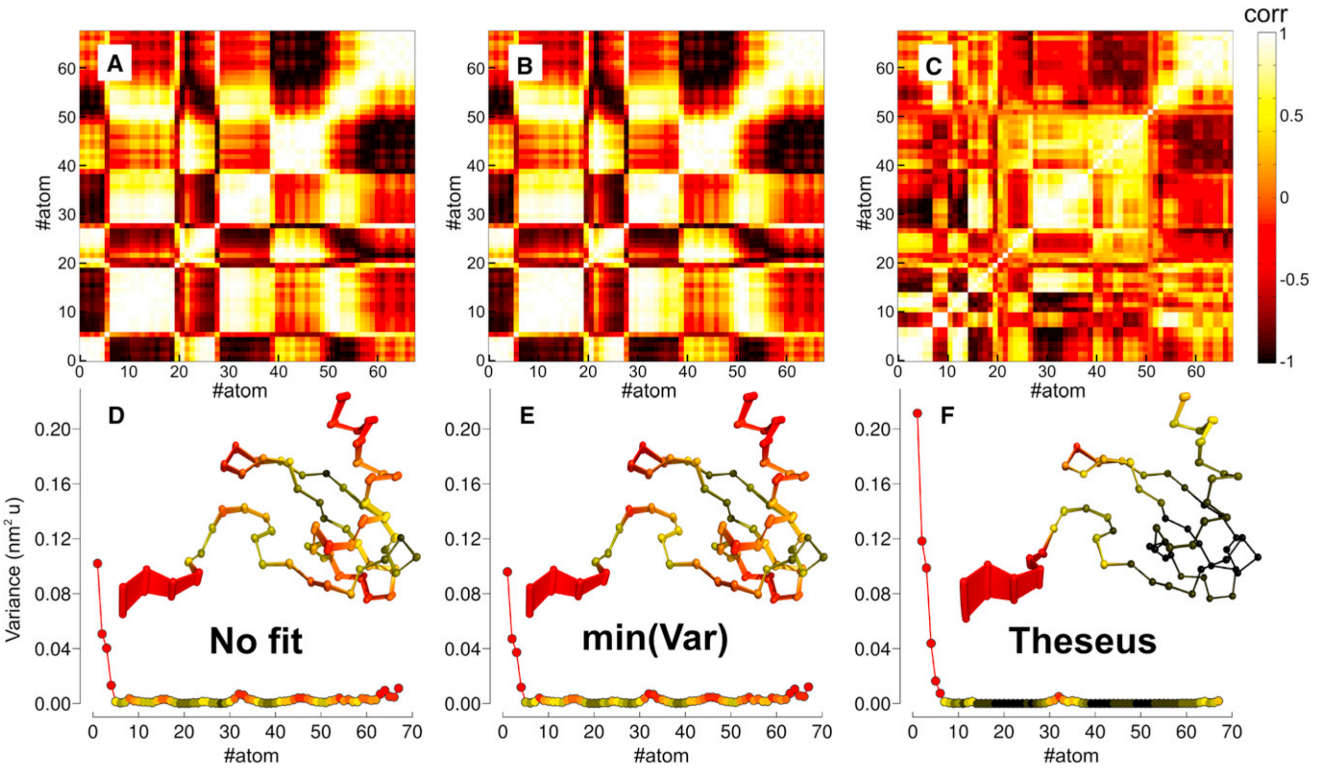

FIGURE 6 Correlation matrices and mass-weighted variances of the Cα atoms for the NM ensemble of stromal-cell-derived factor-1. (*A* and *D*) Nonsuperimposed ensemble. (*B* and *E*) min(Var) superpositioning. (*C* and *F*) Theseus superpositioned ensembles.

As revealed by the comparison of the ensembles treated with different superpositioning methods (see Supporting Material), the min(Var), min(Var+Prev), and min(Var+NN) approaches yield comparable sets of rotations. Hence, the selection of an algorithm for ensemble superpositioning by minimizing variance and removing ambiguous rotations depends only on the ordering of the trajectory and user preference.

CONCLUSIONS

We have introduced a new (to our knowledge) class of methods for multiple-structure superpositioning that minimize the variance over an ensemble. Ambiguous rotations, which often occur in the ensembles of the flexible intrinsically disordered peptides, are alleviated by applying the min(Var+Prev) algorithm, in case subsequent frames in a trajectory are structurally similar. Otherwise, a solution to the TSP may be used to reorder an ensemble before using the min(Var+Prev) superpositioning. The min(Var+NN) algorithm allows one to resolve suboptimal rotations by minimizing the variance and the RMSD to the structures in the local neighborhood of each member of the ensemble. We showed that both approaches are able to remove the external degrees of freedom from large structural ensembles while at the same time they eliminate ambiguous rotations and avoid bias toward any intrinsic molecular motion. Therefore, the algorithms introduced here provide a solid basis for the unbiased analysis of internal dynamics.

The implementation of the methods developed in this work is based on the Gromacs framework and is freely available from our group's website: http://www3.mpibpc.mpg.de/groups/de_groot/software.html.

SUPPORTING MATERIAL

Nine figures, four tables, additional analysis, simulation parameters, and references (42–66) are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(12)01195-2.

We are grateful to Dirk Matthes and Servaas Michielssens for carefully reading the manuscript and providing valuable comments.

Vytautas Gapsys was supported by a stipend from the International Max Planck Research School for Physics of Biological and Complex Systems.

REFERENCES

1. Gower, J. C., and G. B. Dijksterhuis. 2004. Procrustes Problems., Vol. 30. Oxford University Press, New York.
2. Flower, D. R. 1999. Rotational superposition: a review of methods. *J. Mol. Graph. Model.* 17:238–244.
3. McLachlan, A. D. 1972. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallogr. A.* 28:656–657.
4. Nyburg, S. C. 1974. Some uses of a best molecular fit routine. *Acta Crystallogr. B.* 30:251–253.
5. McLachlan, A. D. 1982. Rapid comparison of protein structures. *Acta Crystallogr. A.* 38:871–873.
6. Kabsch, W. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A.* 32:922–923.
7. Kabsch, W. 1978. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A.* 34:827–828.
8. Horn, B. K. P. 1987. Closed-form solution of absolute orientation using unit quaternions. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* 4:629–642.
9. Diamond, R. 1988. A note on the rotational superposition problem. *Acta Crystallogr. A.* 44:211–216.
10. Kearsley, S. K. 1989. On the orthogonal transformation used for structural comparisons. *Acta Crystallogr. A.* 45:208–210.
11. Kneller, G. R. 1991. Superposition of molecular structures using quaternions. *Mol. Simul.* 7:113–119.
12. Shapiro, A., and J. Botha. 1988. Dual algorithm for orthogonal procrustes rotations. *SIAM J. Matrix Anal. Appl.* 9:378–383.
13. von Neumann, J. 1937. Some matrix inequalities and metrization of matrix space. *Tomsk Univ. Rev.* 1:286–300.
14. Schönemann, P. H. 1966. A generalized solution of the orthogonal Procrustes problem. *Psychometrika.* 31:1–10.
15. Coutsias, E. A., C. Seok, and K. A. Dill. 2004. Using quaternions to calculate RMSD. *J. Comput. Chem.* 25:1849–1857.
16. Coutsias, E. A., C. Seok, and K. A. Dill. 2005. Rotational superposition and least squares: the SVD and quaternions approaches yield identical results. Reply to the preceding comment by G. Kneller. *J. Comput. Chem.* 26:1663–1665.
17. Kristof, W., and B. Wingersky. 1971. A generalization of the orthogonal Procrustes rotation procedure to more than two matrices. *Proc. Ann. Conv. Am. Psychol. Assoc.* 6:89–90.
18. Gower, J. C. 1975. Generalized procrustes analysis. *Psychometrika.* 40:33–51.
19. Ten Berge, J. M. F. 1977. Orthogonal Procrustes rotation for two or more matrices. *Psychometrika.* 42:267–276.
20. Sutcliffe, M. J., I. Haneef, …, T. L. Blundell. 1987. Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.* 1:377–384.
21. Shapiro, A., J. D. Botha, …, A. M. Lesk. 1992. A method for multiple superposition of structures. *Acta Crystallogr. A.* 48:11–14.
22. Kearsley, S. K. 1990. An algorithm for the simultaneous superposition of a structural series. *J. Comput. Chem.* 11:1187–1192.
23. Diamond, R. 1992. On the multiple simultaneous superposition of molecular structures by rigid body transformations. *Protein Sci.* 1:1279–1287.
24. Gerstein, M., and C. Chothia. 1991. Analysis of protein loop closure. Two types of hinges produce one motion in lactate dehydrogenase. *J. Mol. Biol.* 220:133–149.
25. Wriggers, W., and K. Schulten. 1997. Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. *Proteins.* 29:1–14.
26. Yu-Shen, L., F. Yi, and R. Karthik. 2009. Using least median of squares for structural superposition of flexible proteins. *BMC Bioinf.* 10:29.
27. Damm, K. L., and H. A. Carlson. 2006. Gaussian-weighted RMSD superposition of proteins: a structural comparison for flexible proteins and predicted protein structures. *Biophys. J.* 90:4558–4573.
28. Theobald, D. L., and D. S. Wuttke. 2006. Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem. *Proc. Natl. Acad. Sci. USA.* 103:18521–18527.
29. Theobald, D. L., and D. S. Wuttke. 2006. THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics.* 22:2171–2172.

30. Theobald, D. L., and D. S. Wuttke. 2008. Accurate structural correlations from maximum likelihood superpositions. *PLOS Comput. Biol.* 4:e43.

31. de Groot, B. L., D. M. F. van Aalten, …, H. J. Berendsen. 1997. Prediction of protein conformational freedom from distance constraints. *Proteins.* 29:240–251.

32. Hahsler, M., and K. Hornik. 2007. TSP—infrastructure for the traveling salesperson problem. *J. Stat. Softw.* 23:1–21.

33. Hahsler, M., and K. Hornik. 2011. Traveling salesperson problem (TSP). r package version 1.0–6. http://CRAN.R-project.org/. Accessed November 13, 2012.

34. Theobald, D. L. 2005. Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. *Acta Crystallogr. A.* 61:478–480.

35. Hub, J. S., and B. L. de Groot. 2009. Detection of functional modes in protein dynamics. *PLOS Comput. Biol.* 5:e1000480.

36. Hess, B., C. Kutzner, …, E. Lindahl. 2008. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* 4:435–447.

37. Schlitter, J. 1993. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem. Phys. Lett.* 215:617–621.

38. Hayward, S., and B. L. de Groot. 2008. Normal modes and essential dynamics. *In* Methods in Molecular Biology, Molecular Modelling of Proteins, *Vol. 443*. Humana Press, Totowa, NJ. 89–106.

39. Crump, M. P., J. H. Gong, …, I. Clark-Lewis. 1997. Solution structure and basis for functional activity of stromal cell-derived factor-1; dissociation of CXCR4 activation from binding and inhibition of HIV-1. *EMBO J.* 16:6996–7007.

40. Hayward, S., A. Kitao, and N. Gō. 1995. Harmonicity and anharmonicity in protein dynamics: a normal mode analysis and principal component analysis. *Proteins.* 23:177–186.

41. Kitao, A., and N. Go. 1999. Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.* 9:164–169.

42. Bollobás, B. 1998. Modern Graph Theory., Vol. 184. Springer Verlag, Berlin.

43. R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org. Accessed November 13, 2012.

44. Papadimitriou, C. H. 1977. The Euclidean travelling salesman problem is NP-complete. *Theor. Comput. Sci.* 4:237–244.

45. Applegate, D., W. Cook, and A. Rohe. 2003. Chained Lin-Kernighan for large traveling salesman problems. *INFORMS J. Comput.* 15: 82–92.

46. Applegate, D., R. Bixby, …, W. Cook. 2006. Concorde TSP solver. http://www.tsp.gatech.edu/concorde. Accessed November 13, 2012.

47. Lin, S., and B. W. Kernighan. 1973. An effective heuristic algorithm for the traveling-salesman problem. *Oper. Res.* 21:498–516.

48. Sticht, H., P. Bayer, …, P. Rösch. 1995. Structure of amyloid A4-(1-40)-peptide of Alzheimer's disease. *Eur. J. Biochem.* 233:293–298.

49. Kelley, L. A., and M. J. Sutcliffe. 1997. OLDERADO: on-line database of ensemble representatives and domains. On Line Database of Ensemble Representatives And DOmains. *Protein Sci.* 6:2628–2630.

50. Seeliger, D., J. Haas, and B. L. de Groot. 2007. Geometry-based sampling of conformational transitions in proteins. *Structure.* 15:1482–1492.

51. Jorgensen, W. L., D. S. Maxwell, and J. Tirado-Rives. 1996. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* 118:11225–11236.

52. Kaminski, G. A., R. A. Friesner, …, W. L. Jorgensen. 2001. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B.* 105:6474–6487.

53. Hornak, V., R. Abel, …, C. Simmerling. 2006. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins.* 65:712–725.

54. Best, R. B., and G. Hummer. 2009. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J. Phys. Chem. B.* 113:9004–9015.

55. Lindorff-Larsen, K., S. Piana, …, D. E. Shaw. 2010. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins.* 78:1950–1958.

56. Jorgensen, W. L., J. Chandrasekhar, …, M. L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926.

57. Berendsen, H. J. C., J. R. Grigera, and T. P. Straatsma. 1987. The missing term in effective pair potentials. *J. Phys. Chem.* 91:6269–6271.

58. Berendsen, H. J. C., J. P. M. Postma, …, J. R. Haak. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–3690.

59. Bussi, G., D. Donadio, and M. Parrinello. 2007. Canonical sampling through velocity rescaling. *J. Chem. Phys.* 126:014101.

60. Parrinello, M., and A. Rahman. 1981. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* 52:7182–7190.

61. Darden, T., D. York, and L. Pedersen. 1993. Particle mesh Ewald: an Nlog (N) method for Ewald sums in large systems. *J. Chem. Phys.* 98:10089.

62. Essmann, U., L. Perera, …, L. G. Pedersen. 1995. A smooth particle mesh Ewald method. *J. Chem. Phys.* 103:8577.

63. Byrd, R. H., P. Lu, …, C. Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* 16:1190–1208.

64. Zhu, C., R. H. Byrd, …, J. Nocedal. 1997. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.* 23:550–560.

65. Metropolis, N., A. W. Rosenbluth, …, E. Teller. 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21:1087.

66. Kirkpatrick, S., C. D. Gelatt, Jr., and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science.* 220:671–680.