Supporting Information: Optimal superpositioning of flexible molecule ensembles

Vytautas Gapsys
Computational biomolecular dynamics group,
Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

Bert L. de Groot[1]
Computational biomolecular dynamics group,
Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

[1]Corresponding author. Address: Computational biomolecular dynamics group, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, Göttingen 37077, Germany Tel.: ++49-551-2012308 , Fax: ++49-551-2012302

S1

Trajectory rearrangement for the min(Var+Prev) approach

Method

Restructuring a trajectory such that subsequent frames after direct superpositioning have a suitably low RMSD is equivalent to a problem of finding the shortest Hamiltonian path over an ensemble. In graph theory, a Hamiltonian path is defined as a route through a graph that visits every vertex once (1). The solution of this problem guarantees that the overall sum of the RMSDs to a previous frame over the trajectory will be minimal.

The shortest Hamiltonian path can be found by modifying the solution to the Traveling Salesman Problem (TSP), which originally aims at finding the shortest Hamiltonian cycle. To solve the TSP we used the infrastructure by Hahsler and Hornik (2, 3) as implemented in the statistical software package R (4). The TSP is proven to be an N-complete problem (5), hence, to obtain paths with the minimal overall RMSD we employed Concorde (6) TSP solver's chained Lin-Kernighan heuristic algorithm (7), which is a modification of the original Lin-Kernighan (8) heuristic. To reformulate the problem of finding the shortest Hamiltonian cycle into finding the shortest Hamiltonian path, a dummy structure, having zero RMSD to every other structure in the ensemble, was included. The dummy structure defined a termination point for the cycle.

The TSP based trajectory rearrangement was applied for the Aβ and RS peptides, as well as the lysozyme ensemble. For every ensemble, structures were pairwise superimposed and the Cα atoms based RMSD between them was calculated resulting in an RMSD matrix. The pairwise RMSDs between the structures were used as a distance measure that could be supplied as an input to the TSP solver. Prior to starting the Hamiltonian path minimization, we randomly shuffled the frames in each trajectory. After solving the TSP, the RMSD matrices were generated for the newly constructed trajectories. The traces of the RMSD matrices for the initial, shuffled and rearranged trajectories were calculated.

Results

The combination of the progressive fitting approach with the variance minimization requires an ensemble to be ordered such that two subsequent frames are similar enough to each other to be unambiguously superimposed. If this requirement is fulfilled, the superimposed subsequent frames should have a low RMSD value. Here we show how any ensemble can be rearranged to minimize the Hamiltonian path over a trajectory. We used MD trajectories of the Aβ peptide, RS peptide and lysozyme as the starting ensembles. For each protein ensemble the RMSD matrices considering the Cα atoms were calculated after pairwise superpositioning of the structures (Figures S1B,F,J). Afterwards, frames of each trajectory were randomly shuffled. The effect of this procedure is obvious from the RMSD matrices in Figures S1C,G,K: the sequence of the frames in each ensemble is random, any pattern that was present in the RMSD matrices is lost after the shuffling. The randomized trajectories were subjected to the Traveling Salesman Problem (TSP) solver. The RMSD matrices for the ensembles constructed from the solution of the TSP are shown in Figures S1D,H,L. It appears that the characteristic patterns in the matrices reappear, however, with more (RS peptide, lysozyme) or less (Aβ) prominent differences to the original MD trajectories. The effects of the ensemble shuffling and reconstruction

can also be analyzed by plotting the ordering of the shuffled (and reconstructed) trajectories against the original MD ensemble ordering (Figures S1A,E,I). The randomized frames almost uniformly cover the graph's space for all three structures, showing no correlation with the MD ensembles. The TSP trajectories follow distinct paths: in case of the Aβ peptide, the almost straight diagonal line indicates that the TSP ensemble was reconstructed in a very similar order as the original MD trajectory. For the other two cases, the reconstructed ensembles resulted in a different ordering from their original trajectories.

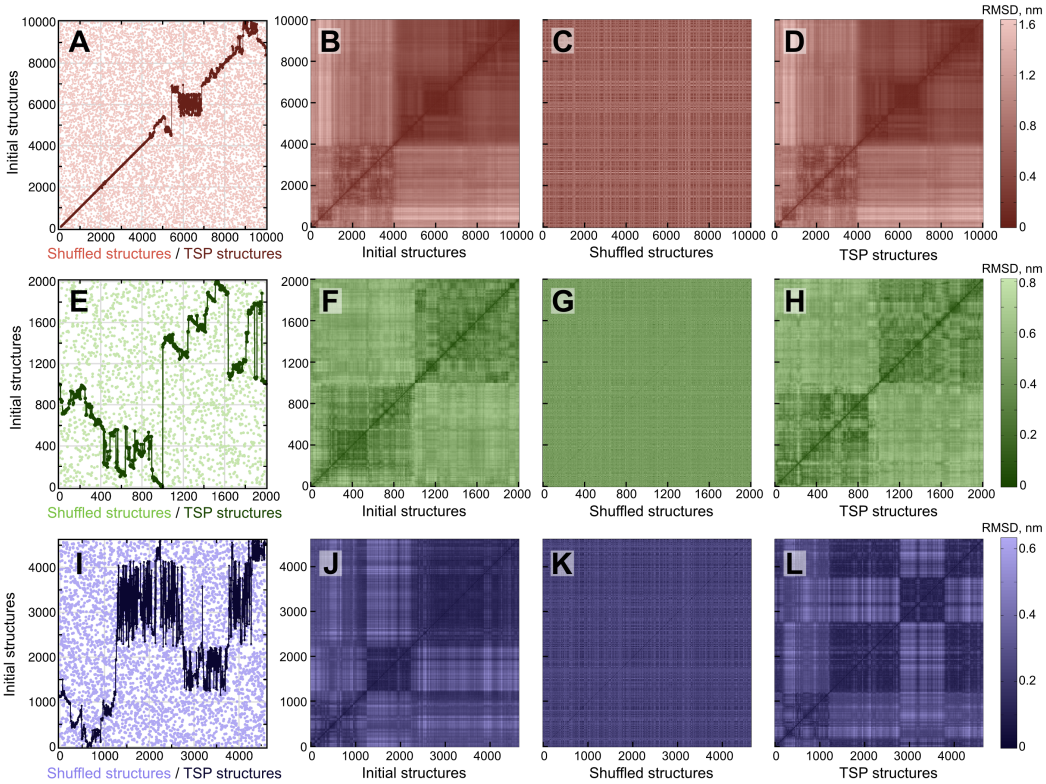

Figure S1: **Solution of the Traveling Salesman Problem as a method for the trajectory reconstruction.** Trajectories of the Aβ peptide (A-D), RS peptide (E-H) and lysozyme (I-L). The frame indices of the shuffled and TSP reordered frames against the initial trajectory frame indices (A,E,I). The RMSD matrices were calculated after the pairwise superpositioning of the initial MD trajectories (B,F,J), shuffled trajectories (C,G,K) and the trajectories reordered according to the TSP solution (D,H,L).

For all three systems, the reconstructed trajectories have a shorter path over the ensemble than the respective MD trajectories (Table S1). Hence, the TSP trajectories fulfill the requirement of low RMSD between two subsequent frames and, thus, can be processed with the min(Var+Prev) algorithm. The application of the min(Var+Prev) method for the TSP reordered trajectories is illustrated in the Figure S2.

Table S1: **Sum of RMSDs (nm) after progressive superposition.**

Structure	Initial trajectory	Shuffled frames	TSP trajectory
Aβ peptide	830.53	6884.42	794.79
RS peptide	199.78	906.42	175.13
Lysozyme	446.50	1114.73	362.00

Table S2: **Mass weighted variances (nm^2u) after progressive superposition.**

Structure	Initial trajectory	Shuffled frames	TSP trajectory
Aβ peptide	171.92	669.89	168.55
RS peptide	30.76	74.92	27.89
Lysozyme	65.96	105.93	67.57

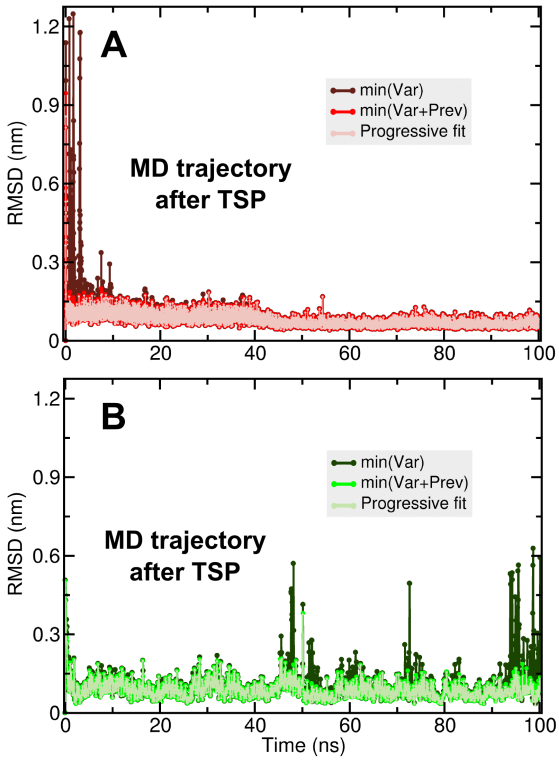

Figure S2: **RMSD analysis after the min(Var) and min(Var+Prev) superpositioning.** MD trajectories reordered according to the solution to the TSP for the Aβ (A) and RS (B) peptides. The RMSD was calculated between structures at time τ and $\tau - 1$.

S2

Parameters for the molecular dynamics simulations of the Aβ and RS peptides.

The simulations were performed using the Gromacs 4.5 (9) molecular dynamics package. The starting structure for the Aβ peptide was selected from the pdb ensemble 1AML (10). Model #6 was proposed by the OLDERADO (11) clustering procedure as one of the representative structures of the ensemble. The RS peptide's sequence was constructed as an alternating sequence of arginine-serine residues (15 amino acids in total). The starting structure for the peptide was generated with tCONCOORD (12). The OPLS (13, 14) force field was used for the simulations of the Aβ peptide. The RS peptide was simulated with the Amber99sb-ILDN* (15–17) force field. The Aβ and RS structures were solvated in TIP4P (18) and SPC/E (19) water, respectively. Sodium and chloride ions were added to neutralize the simulation box and match the physiological salt concentration of 0.15 M. Prior to the production runs, the structures were subjected to the steepest descent energy minimization. Virtual sites were used to replace hydrogen atoms, thus removing the fastest degrees of freedom. This allowed increasing the integration timestep for the molecular dynamics simulations to 4 fs. For the Aβ peptide the Berendsen scheme for temperature and pressure coupling (20) was used, keeping the temperature at 300 K and pressure at 1 bar ($\tau_t = 0.1$ ps, $\tau_p = 1.0$ ps). Temperature coupling for the RS peptide was performed using the velocity rescaling algorithm (21) with a reference temperature of 298 K and time constant of 0.1 ps. The Parrinello-Rahman barostat (22) with a time constant of 5 ps was used to maintain a pressure of 1 bar. The neighbour list radius for the Aβ peptide was set to 1 nm, whereas for the RS peptide it was set to 1.2 nm. The van der Waals cut-off for the Aβ peptide was set to 1 nm, and for the RS peptide the cut-off was set to 1.1 nm. For both systems electrostatic interactions were treated with the particle-mesh Ewald (PME) (23, 24) method. In the case of the Aβ peptide, the real space cut-off was set to 1 nm, and the spacing of the fourier grid was 0.12 nm, a PME order of 4 was used. For the RS peptide a real space cut-off of 1.2 nm was used, the fourier grid spacing was 0.14 nm and the PME order was set to 4.

S3

Parameters for the energy minimization and hessian matrix calculation for the normal mode analysis of the Aβ and RS peptides, lysozyme and stromal cell derived factor-1.

The assumptions for the Least Squares fitting require equal variances of the variables and absence of correlation between them (25). As noted by Theobald and Wuttke (26), atoms in protein ensembles may violate these requirements, in turn possibly causing artifacts in the least squares based superpositioning. To elucidate and quantify the effects of the potential artifacts that could be introduced by the least squares based variance minimization algorithms, we generated Aβ peptide, RS peptide and lysozyme structural ensembles that contain no translational and rotational motions by construction. For that purpose, we energy minimized the structures and performed normal mode (NM) analysis for each protein.

All the systems were energy minimized in vacuum using the Amber99sb-ILDN* force field. The structures for the Aβ and RS peptides were selected as described in S1. For lysozyme, the first structure from the ensemble simulated by Hub and de Groot (27) was used. For the stromal cell derived factor-1, the first structure from the 2SDF (28) NMR ensemble was selected. The L-BFGS energy minimization algorithm (29, 30) was employed running Gromacs 4.5 in double precision which was required to reach 10^{-8} kJ mol^{-1}nm^{-1} force or to converge at machine precision using a step size of 0.001. The non-bonded interactions were treated with simple cut-off schemes setting the cut-off values to include the whole protein. The minimized structures were subjected to the Hessian diagonalization procedure using the same parameter set as for the energy minimization.

The normal mode ensembles were generated by sampling from a Gaussian distribution over the slowest eigenmode, where the variance of the distribution is temperature dependent. To enhance the motion in a local minimum we used a temperature of 1000K for the ensemble generation. These ensembles by construction consist of internal motions only and are therefore devoid of any external motions. Thus, they serve as a suitable reference case for the evaluation of superpositioning schemes, as the introduction of any external contribution to the dynamics of the ensemble would be directly evident as an increase in the ensemble variance. The variances over the normal mode ensembles were estimated after superpositioning the structures using the newly introduced methods. Since the absence of the external degrees of freedom in normal mode ensembles is ensured when considering all the atoms of the structures, the superpositioning was performed on all atoms.

Various superpositioning procedures were applied to the normal mode ensembles including the maximum likelihood superpositioning method Theseus (31). The Theseus procedure was performed with the default parameters, except for setting on the option of full covariance and correlation matrix estimation as used by Theobald and Wuttke (26). The non-fitted normal mode ensembles provided the reference variance for the evaluation of the superpositioning methods.

S4

Analysis of the performance of the min(Var+NN) algorithm with different number of nearest neighbours considered.

In Figure S3 the effect of applying the min(Var+NN) algorithm on the Aβ and RS peptide trajectories is demonstrated for different numbers of nearest neighbours. The spikes in the RMSD to the previous frame (that may or may not have been among the nearest neighbours) plots (Figures S3 B,D), decrease and eventually disappear with an increasing number of the nearest neighbours considered. It is important to note, that with a large number of nearest neighbours (NN=1000) some spikes reappear in the case of the RS peptide (Figure S3 D): instead of resolving the sub-optimal rotations between similar structures, the large list of nearest neighbours puts a weight onto the structures that are less similar, which may cause re-introduction of ambiguous superpositions. In such a situation, the superposition becomes more similar to the variance minimization, min(Var) algorithm. This can be observed from the variance values with an increasing number of nearest neighbours (Figures S3 A,C): including large sets of nearest neighbours causes a decrease in the variance. Additionally, a large number of nearest neighbours requires more iterations for the algorithm to converge, hence, our suggestion is to limit this parameter to ≤ 100, which is sufficient to resolve ambiguous rotations in the local environment over an ensemble.

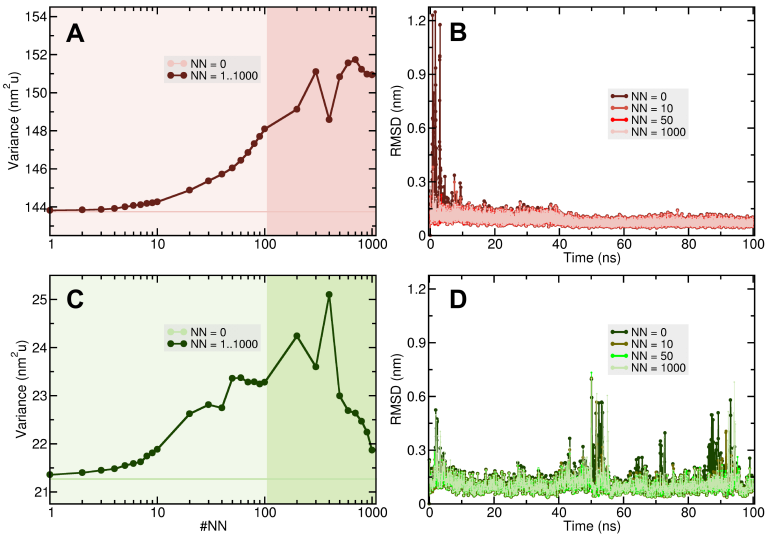

Figure S3: **The effect of different number of nearest neighbours used by the min(Var+NN) algorithm.** The mass weighted variance (A,C) and RMSD plots (B,D) after the min(Var+NN) superpositioning with varying number of nearest neighbours for the Aβ (A,B) and RS peptides (C,D).

S5

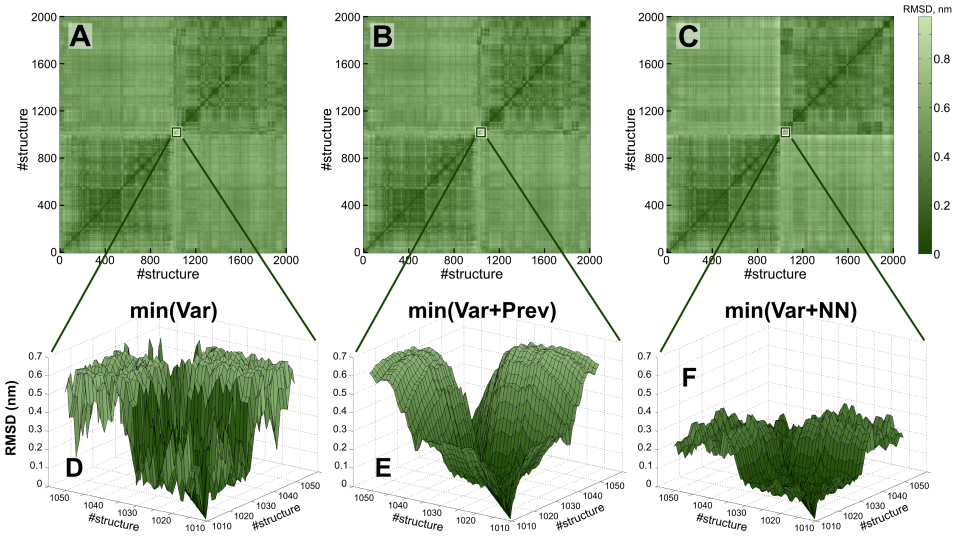

Figure S4: **Pairwise RMSD matrices and surfaces of the RS peptide ensemble.** RMSD values after the min(Var) (A), min(Var+Prev) (B) and min(Var+NN) (C) superpositioning. Excerpts from the matrices shown as surfaces (D, E and F) illustrate the smoothening effect of the min(Var+Prev) and min(Var+NN) algorithms, as well as the significant reduction of the RMSD in the local neighbourhood of each structure resulting from the latter method.

S6

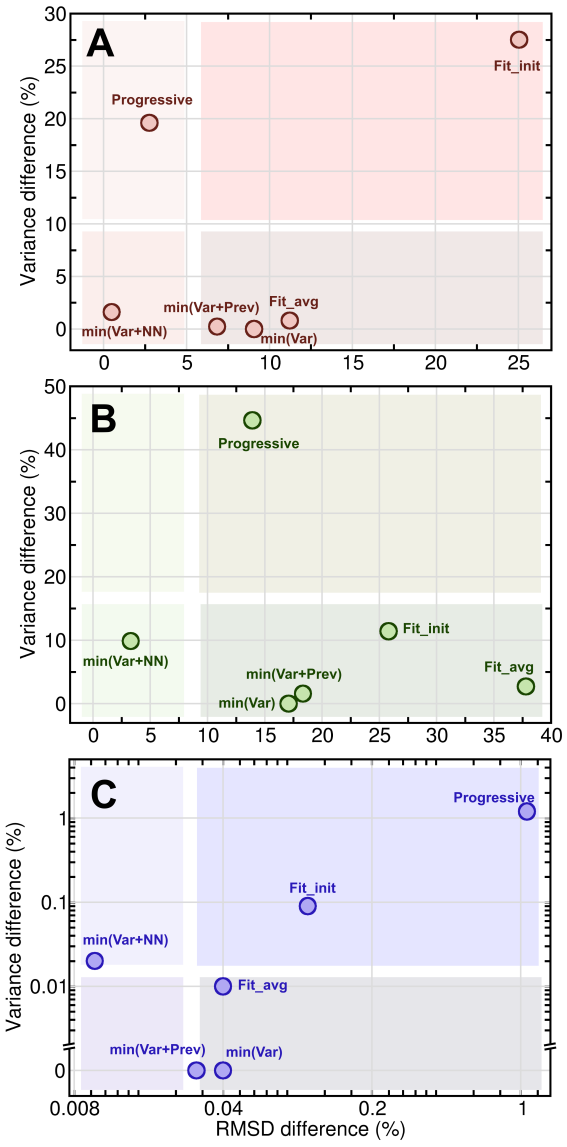

Figure S5: **Mapping of the superpositioning algorithms into a common space defined by the ensemble variance and local neighbourhood RMSD.** MD ensembles of the Aβ (A) and RS (B) peptides, as well as lysozyme (C) were used for the analysis. The local neighbourhood of 50 nearest neighbours was considered for the RMSD calculation, as opposed to 10 nearest neighbours considered in Figure 4 in the main text. The background colors of the quadrants are to guide the eye only and do not represent any calculated result.

S7

Table S3: **Local neighbourhood RMSD difference from the optimal value (%) of the superimposed ensembles.**

Structure	Fit on starting str.	Fit on average str.	Progressive fitting	min(Var)	min(Var +Prev)	min(Var +NN)
Aβ all[a]	13.85	6.13	10.14	1.32	1.40	1.91
Aβ NN=1[b]	21.19	09.63	0.18	6.38	1.30	0.27
Aβ NN=10[c]	23.30	10.55	0.52	7.14	3.49	0.31
Aβ NN=50[d]	25.05	11.22	2.76	9.07	6.85	0.49
Aβ prev[e]	22.85	10.83	0.00	6.56	0.88	0.28
RS all[a]	10.82	14.32	25.45	5.29	6.05	9.32
RS NN=1[b]	25.81	43.95	5.07	16.63	13.46	2.55
RS NN=10[c]	26.79	41.75	8.54	17.77	17.89	2.59
RS NN=50[d]	25.81	37.80	13.92	17.08	18.34	3.30
RS prev[e]	28.57	43.82	0.00	19.64	2.13	2.82
Lysozyme all[a]	0.08	0.03	0.75	0.03	0.03	0.04
Lysozyme NN=1[b]	0.07	0.03	0.69	0.03	0.03	0.01
Lysozyme NN=10[c]	0.07	0.03	0.91	0.03	0.03	0.01
Lysozyme NN=50[d]	0.07	0.04	1.07	0.04	0.03	0.01
Lysozyme prev[e]	0.10	0.05	0.01[f]	0.05	0.01	0.01

[a] All structures in the ensemble were considered for the RMSD calculation.

[b] One nearest neighbour was considered for the RMSD calculation.

[c] Ten nearest neighbours were considered for the RMSD calculation.

[d] Fifty nearest neighbours were considered for the RMSD calculation.

[e] A preceding structure in a trajectory was considered for the RMSD calculation.

[f] The value deviates from zero due to the accumulated numerical errors during the progressive superpositioning.

S8

Convergence analysis: description of the Monte Carlo simulated annealing procedure and graphical illustration of the min(Var), min(Var+Prev) and min(Var+NN) convergence.

Method
Iterative variance minimization algorithms in practice are known to converge rapidly (32, 33). Convergence analysis and the verification of the stationary solutions has previously been performed by Shapiro et al. (34). To investigate the dependence of the algorithms described above on the initial conditions and to estimate their convergence, we employed a Monte Carlo based simulated annealing procedure. A random structure from a trajectory was selected and rotated by applying a random rotation matrix. The acceptance of the new rotation was guided by the Metropolis criterion (35), where the minimizable functions were defined by the equations (1), (4) and (6) in the main manuscript for the min(Var), min(Var+Prev) and min(Var+NN) algorithms, respectively. The temperature schedule was applied in a simulated annealing manner (36) by gradually cooling the system down and heating it up again after no acceptance was encountered for a certain number of iterations. The procedure was repeated for 10^5 iterations over a trajectory. The search for the min(Var) and min(Var+Prev) algorithms was started with various initial conditions, also considering the best result achieved by the iterative procedures. As a dependence of the min(Var+NN) algorithm on the starting rotations of the structures was observed, the Monte Carlo search for this method was started from the progressively fit trajectory, as well as from the solution reached by the iterative procedure. The values obtained by the Monte Carlo routines can be considered as deep local minima, possibly close to the global minimum. They served as reference points for the iterative procedures. The min(Var) and min(Var+Prev) runs were initialized from an MD trajectory of the RS peptide superimposed onto a randomly selected structure. We generated 100 such starting ensembles to estimate the dependence of the algorithms on the initial conditions. For the min(Var+NN) approach, we pre-superimpose an ensemble using progressive fitting. Therefore, testing the dependence on the initial conditions was not necessary in this case. Instead, we chose various numbers of the nearest neighbours, namely 3, 10 and 50, and evaluated the convergence of the algorithm. The iterative runs and Monte Carlo search for the convergence analysis were performed using the RS peptide's MD trajectory, considering the Cα atoms for the superpositioning and variance calculation.

Results The min(Var) and min(Var+Prev) algorithms were found to converge rapidly to a value that was identified to be a deep minimum by the Monte Carlo search. For both methods, independent on the starting conditions, the solution for the RS peptide's MD trajectory was found within the first 10 iterations. The performance of the min(Var+NN) was evaluated for the three different numbers of nearest neighbours (3, 10, 50) considered during the superpositioning. Taking into account a larger local neighbourhood comes at a cost of slower convergence. A weak fluctuation of the variance was observed for larger numbers of nearest neighbours even after the variance stops decreasing. During the testing phase of the min(Var+NN) algorithm, it was observed that with larger numbers of nearest neighbours multiple deep local minima may occur

(data not shown). To avoid a dependence on the initial conditions, we implemented a progressive pre-superpositioning step for the algorithm.

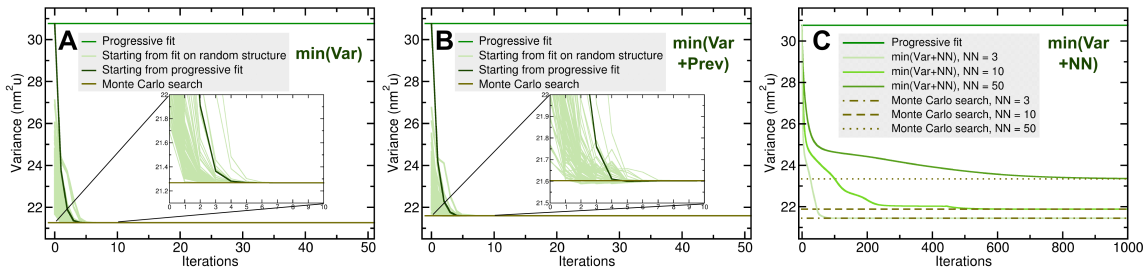

Figure S6: **Convergence analysis of the min(Var), min(Var+Prev) and min(Var+NN) algorithms.** (A,B) Depictions of the mass weighted variance over the ensemble for the different number of iterations of the algorithms starting from various initial conditions. (C) The change of the mass weighted variance over the iterative runs is given for the 3, 10 and 50 nearest neighbours taken into consideration by the min(Var+NN) algorithm. Monte Carlo search results mark the reference variance values. MD trajectory of the RS peptide was used for the analysis.

S9

Comparison of the superpositioning methods

Method

To compare the effect of different superpositioning approaches, it is desired to superimpose structural ensembles treated with different superpositioning techniques such that the least squares difference between the corresponding members of the ensembles is minimal. The relative rotations within each ensemble must be retained, as they carry information of the effect of a superpositioning method applied. This can be achieved by finding a single rotation matrix optimally frame-by-frame superimposing two ensembles. For this purpose we define the minimizable function

$$E = \frac{1}{2} \sum_{\tau} \sum_{n} w_n \left(U \mathbf{x}_{n\tau} - \mathbf{y}_{n\tau} \right)^2 \tag{S1}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the members of the two superpositioned trajectories. Imposing the orthogonality constraints and following Kabsch's derivation, we find that

$$U \left(V + L \right) = W \tag{S2}$$

with L being the Lagrange multiplier matrix, $v_{ij} = \sum_{\tau} \sum_{n} w_n x_{ni\tau} x_{nj\tau}$ and $w_{ij} = \sum_{\tau} \sum_{n} w_n x_{nj\tau} y_{ni\tau}$. Eq. (S2) can be solved as shown by Kabsch (37).

This method was used to pairwise superimpose the normal mode and MD ensembles of the Aβ peptide, RS peptide and lysozyme processed with the different superpositioning methods. To quantify the effect of the superpositioning approaches, the distance between the ensembles was calculated by summing the pairwise RMSD values over all structures and normalizing by the number of frames in a trajectory.

Results

A comparison of the superpositioning methods was performed by pairwise superimposing structural ensembles, retaining the relative orientations of the molecules within the trajectories. The quantitative evaluation of the methods is summarized in Figure S7, where small RMSD values indicate a similar superposition. The upper right triangular of the matrices in Figure S7 corresponds to the normal mode ensembles. In this case, all atoms were considered for the superpositioning and RMSD calculation. The lower left triangles of the matrices were constructed from the MD ensembles by considering the Cα atoms for the analysis. For all the three proteins analysed, all the superpositioning methods rotated the normal mode ensembles differently in comparison to the non-fitted reference. However, the differences are very small: for the normal mode ensembles, RMSD values for a pair of structures in Figure S7 are multiplied by the factor 10^3 to provide meaningful comparison. The min(Var), min(Var+Prev) and min(Var+NN) with 50 nearest neighbours considered appear to yield similar results for the Aβ and RS peptide normal mode ensembles, whereas for the case of lysozyme min(Var+NN) produces a different set of rotations from the other methods. However, the difference is minor and the min(Var+NN) trajectory could not be distinguished from the other trajectories treated with the other techniques when comparing it to the intact reference ensemble. Interestingly,

the progressively superpositioned normal mode ensemble of the RS peptide significantly differs from the others, but combining it with the variance minimization in the method min(Var+Prev) brings its RMSD close to the other approaches.

Comparison of the superpositioning methods based on the MD ensembles reveals a high similarity between the min(Var) and min(Var+Prev) procedures. This finding complements the observation in Figure S5, indicating that the min(Var) and min(Var+Prev) algorithms produce similar ensembles. In contrast, the min(Var+NN) method is mapped further from the other variance minimizing algorithms in Figure S5, which is corroborated in Figure S7 which shows that the min(Var+NN) algorithm samples a unique solution, particularly for the RS peptide. The Theseus method was not included in the comparison, because structures superpositioned with the maximum likelihood approach may not have their centers of mass set to the origin. However, the rotation procedure for two ensembles using one rotation matrix for every structure as applied here is correct only when the translational motion is removed.

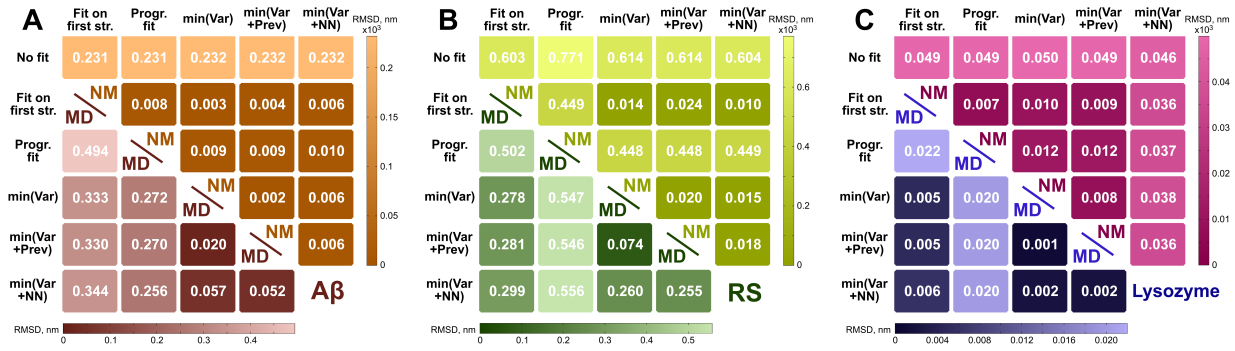

Figure S7: **Comparison of the superpositioning methods.** The mean of the pairwise RMSD values for every structure in the normal mode ensembles (upper right triangle) and MD trajectories (lower left triangle) for the Aβ peptide (A), RS peptide (B) and lysozyme (C) superimposed with different superpositioning methods.

S10

Analysis of the local neighbourhood

An estimator for the superposition quality is constructed as a normalized difference between the RMSD sum over the local neighbourhoods after the ensemble superpositioning and the optimal RMSD value:

$$\frac{\sum_t \sum_{NN} (RMSD_{NN_t} - RMSD^{opt}_{NN_t})}{\sum_t \sum_{NN} RMSD^{opt}_{NN_t}} \tag{S3}$$

where $RMSD_{NN_t}$ is an RMSD value calculated over the Cα atoms between a structure t and the molecules in it's local neighbourhood after applying one of the ensemble superpositioning algorithms. $RMSD^{opt}_{NN_t}$ is the RMSD value obtained from the pairwise superpositioning. Analogously, the difference between the superimposed ensemble variance and the minimal variance, for which we use the result of the min(Var) algorithm, is constructed

$$\frac{Var - Var_{opt}}{Var_{opt}} \tag{S4}$$

For convenience we express the measures (S3) and (S4) in percentage and use them to quantify the performance of the different superpositioning algorithms applied to the Aβ peptide, RS peptide and lysozyme.

S11

Analysis of the lysozyme's NM and MD ensembles

Effect of essentially different superpositioning approaches, min(Var) and Theseus, was analysed by looking into the principal motions of lysozyme. The normal mode and molecular dynamics ensembles were generated and superpositioned accordingly (see Figure 5 in the main manuscript). In addition to the variances and conformational entropies provided in the Tables 2 and 3 in the main manuscript, we analysed the Root Mean Square Fluctuations of the Cα atoms over the ensembles of lysozyme. Figure S8 provides a quantitative estimate of the effect that was visualised in the main manuscript's Figure 5. Theseus superpositioning has a strong effect on the residue flexibility in both, NM and MD, ensembles. NM ensemble without superposition serves as a reference, since by definition it contains no external degrees of freedom. min(Var) approach results in an RMSF profile almost identical to that of the intact NM trajectory. Similar difference, only of larger magnitude, between the min(Var) and Theseus superpositions can be observed in the MD ensembles.

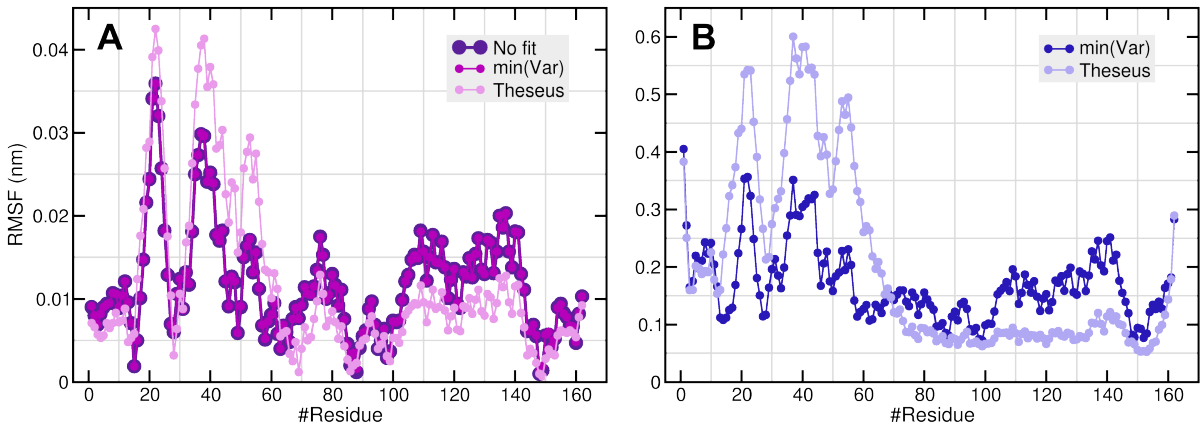

Figure S8: **Root Mean Square Fluctuations for the structural ensembles of lysozyme.** RMSF analysis was performed for the ensembles depicted in the Figure 5 in the main manuscript: (A) normal mode, (B) molecular dynamics ensembles.

Eigenvalue spectra of the convariance matrices (Figure S9) complement previous observation of significantly larger conformational entropy (Table 3 in the main manuscript) estimated for the Theseus superpositioned in comparison to the non-fitted NM ensemble.

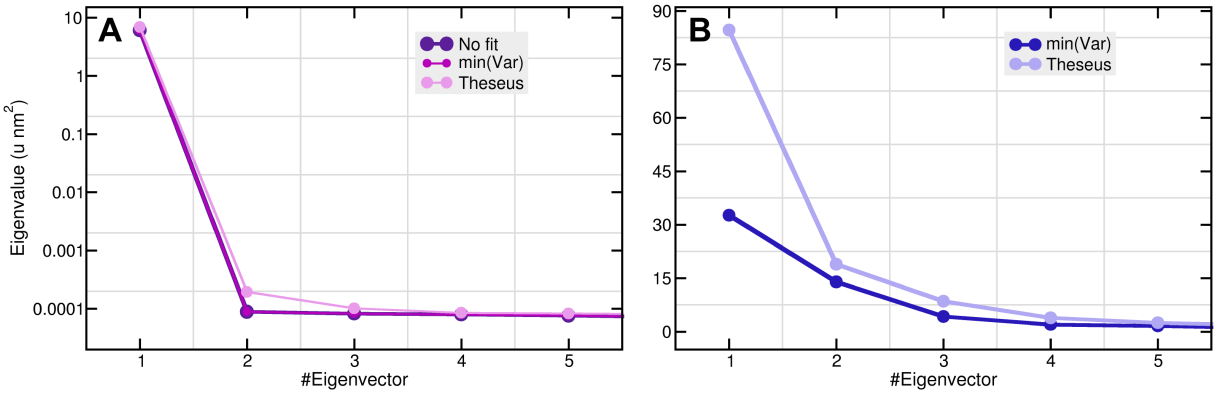

Figure S9: **Eigenvalue spectra of the covariance matrices for the structural ensembles of lysozyme.** Covariance matrix was constructed and diagonalized for the ensembles depicted in the Figure 5 in the main manuscript: (A) Normal Mode, (B) Molecular Dynamics ensembles.

S12

Conformational entropies of the molecular dynamics ensembles

Conformational entropies for the MD ensembles were calculated using Schlitter's formula (38) and considering Cα atoms only. The new superpositioning methods provide values for the conformational entropies that are in between of the variance minimized and progressively superimposed ensembles. The min(Var) approach consistently results in low conformational entropies, which is expected when as much as possible of external degrees of freedom are removed. Results of the maximum likelihood based superpositioning are divergent: for the Aβ peptide the estimated conformational entropy for the Theseus algorithm is smaller than for any other fitting approach. On the contrary, Theseus conformational entropy for lysozyme is comparable to the progressively superpositioned trajectory, whereas for the RS peptide, maximum likelihood based method yields a value similar to that of the min(Var) algorithm.

Table S4: **Conformational entropies (J/mol K) of the molecular dynamics ensembles.**

Structure	Fit on starting str.	Fit on average str.	Progressive fitting	min(Var)	min(Var+Prev) MD	min(Var+Prev) TSP	min(Var +NN)[a]	Theseus
Aβ peptide	2393.04	2356.21	2402.74	2349.76	2376.46	2376.88	2379.26	2259.85
RS peptide	989.40	978.97	1006.91	974.96	985.09	986.10	975.618	975.73
Lysozyme	5379.32	5380.25	5440.95	5377.70	5393.10	5388.96	5401.38	5439.57

[a] 50 nearest neighbours were considered.

S13

Computational time estimate for the min(Var+NN) approach

The computational time required by the min(Var+NN) algorithm was estimated on a workstation with 4 Intel Xeon 2.67 GHz CPUs. Performing 100 iterations for 10001 frame of the Aβ peptide considering Cα atoms, 50 NN and running 100 iteration cycles took about 14 minutes, where the pre-processing part was parallelized on 4 threads and the iteration part was running on 1 thread only. If the iteration part is also parallelized on four threads, the run time for this system decreases to approximately 5 minutes, however, the result in terms of variance deviates slightly from a serial run due to the specifics of parallelization.

References

1. Bollobás, B., 1998. Modern graph theory, volume 184. Springer Verlag.

2. Hahsler, M., and K. Hornik, 2007. TSP - Infrastructure for the traveling salesperson problem. *J Stat Softw* 23:1–21. `http://www.jstatsoft.org/v23/i02/`.

3. Hahsler, M., and K. Hornik, 2011. Traveling Salesperson Problem (TSP). `http://CRAN.R-project.org/`, r package version 1.0-6.

4. 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`, ISBN 3-900051-07-0.

5. Papadimitriou, C. H., 1977. The Euclidean travelling salesman problem is NP-complete. *Theoretical Computer Science* 4:237–244.

6. Applegate, D., R. Bixby, V. Chvatal, and W. Cook, 2006. Concorde tsp solver. `http://www.tsp.gatech.edu/concorde`.

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

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

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

10. Sticht, H., P. Bayer, D. Willbold, S. Dames, C. Hilbich, K. Beyreuther, R. W. Frank, and P. Rösch, 1995. Structure of Amyloid A4-(1–40)-Peptide of Alzheimer's Disease. *Eur. J. Biochem.* 233:293–298.

11. Kelley, L. A., and M. J. Sutcliffe, 1997. OLDERADO: On-line database of ensemble representatives and domains. *Protein Sci.* 6:2628–2630.

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

13. 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.

14. Kaminski, G. A., R. A. Friesner, J. Tirado-Rives, and 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.

15. Hornak, V., R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, 2006. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins Struct. Funct. Bioinf.* 65:712–725.

16. 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.

17. Lindorff-Larsen, K., S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, 2010. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinf.* 78:1950–1958.

18. Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926.

19. 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.

20. Berendsen, H. J. C., J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–3690.

21. Bussi, G., D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.* .

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

23. 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.

24. Essmann, U., L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, 1995. A smooth particle mesh Ewald method. *J. Chem. Phys.* 103:8577.

25. Theobald, D. L., and D. S. Wuttke, 2006. Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem. *PNAS* 103:18521–18527.

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

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

28. Crump, M. P., J. H. Gong, P. Loetscher, K. Rajarathnam, A. Amara, F. Arenzana-Seisdedos, J. L. Virelizier, M. Baggiolini, B. D. Sykes, and 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.

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

30. Zhu, C., R. H. Byrd, P. Lu, and J. Nocedal, 1997. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)* 23:550–560.

31. Theobald, D. L., and D. S. Wuttke, 2006. THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics* 22:2171–2172.

32. Gower, J. C., 1975. Generalized procrustes analysis. *Psychometrika* 40:33–51.

33. Ten Berge, J. M. F., 1977. Orthogonal Procrustes rotation for two or more matrices. *Psychometrika* 42:267–276.

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

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

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

37. Kabsch, W., 1976. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* 32:922–923.

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