A rigorous and efficient method to reweight very large conformational ensembles using average experimental data and to determine their relative information content

Hoi Tik Alvin Leung[1], Olivier Bignucolo[2], Regula Aregger[3], Sonja Dames[4], Simon Bernèche[2], and Stephan Grzesiek[1,]*

[1]Focal Area Structural Biology and Biophysics, Biozentrum, University of Basel, CH-4056 Basel, Switzerland

[2]SIB Swiss Institute of Bioinformatics, University of Basel, Klingelbergstrasse 50/70, CH-4056 Basel, Switzerland

[3]Institut für Biochemie, University of Leipzig, Brüderstraße 34, D-04103 Leipzig, Germany

[4]Department of Chemistry, Technische Universität München, Lichtenbergstraße 4, D-85748 Garching, Germany

*Address correspondence to:
Stephan Grzesiek
Focal Area Structural Biology and Biophysics, Biozentrum
University of Basel, CH-4056 Basel, Switzerland
Phone: ++41 61 267 2100
FAX: ++41 61 267 2109
Email: Stephan.Grzesiek@unibas.ch

Keywords: maximum entropy, molecular dynamics, heteronuclear NMR, unfolded proteins, convex optimization, inner point method

Abstract

Flexible polypeptides such as unfolded proteins may access an astronomical number of conformations. The most advanced simulations of such states usually comprise tens of thousands of individual structures. In principle, a comparison of parameters predicted from such ensembles to experimental data provides a measure for their quality. In practice, analyses that go beyond the comparison of unbiased average data have been impossible to carry out on the entirety of such very large ensembles and have therefore been restricted to much smaller subensembles and/or non-deterministic algorithms. Here, we show that such very large ensembles on the order of 10^4 to 10^5 conformations can be analyzed in full by a Maximum Entropy fit to experimental average data. Maximizing the entropy of the population weights of individual conformations under experimental χ^2 constraints is a convex optimization problem, which can be solved in a very efficient and robust manner to a unique global solution even for very large ensembles. Since the population weights can be determined reliably, the reweighted full ensemble presents the best model of the combined information of simulation and experiment. Furthermore, since the reduction of entropy due to the experimental constraints is well defined, its value provides a robust measure of the information content of the experimental data relative to the simulated ensemble and an indication for the density of the sampling of conformational space. The method is applied to the reweighting of a 35000-frame molecular dynamics trajectory of the nonapeptide EGAAWAASS by extensive NMR ^3J-coupling and RDC data. The analysis shows that RDCs provide significantly more information than ^3J-couplings and that a discontinuity in the RDC pattern at the central tryptophan is caused by a cluster of helical conformations. Reweighting factors are moderate and consistent with errors in MD force fields of less than 3 kT. The required reweighting is larger for an ensemble derived from a statistical coil model, consistent with its coarser nature. We call the method COPER for Convex OPtimization for Ensemble Reweighting. Similar advantages of large-scale efficiency and robustness can be obtained for other ensemble analysis methods with convex targets and constraints such as constrained χ^2 minimization and the Maximum Occurrence method.

Introduction

Proteins exist as ensembles of interchanging conformations. Obviously, unfolded polypeptide chains, such as chemically or physically denatured proteins and intrinsically disordered proteins (IDPs), can access extremely large numbers of conformations.[1] A comprehensive description of their structural preferences is a prerequisite for understanding protein folding and the function of IDPs in health and disease.[2] However, also native, folded proteins usually adopt many conformations close to the global free energy minimum,[3] and their interchange is a hallmark of protein function such as catalysis[4] or signal transduction.[5]

A detailed experimental determination of individual structures in such protein ensembles becomes impossible as soon as their number exceeds a few, since the number of conformational degrees of freedom quickly outpaces the number of measurable parameters.[6] To make progress, often ensembles containing tens of thousands of conformers are simulated and compared to experimental data. Simulated ensembles can be obtained by many methods, e.g. the simulation of a random chain according to the coil model of the unfolded state,[7-9] coarse-grained simulations of protein domain motions[10,11] or all-atom molecular dynamics (MD) simulations with varying degrees of complexity.[12-16] The quantitative analysis of such very large ensembles presents a formidable challenge. An initial analysis needs to establish the accuracy and information content of the predicted ensemble relative to any experimental knowledge, and if necessary refine the ensemble to reproduce the experimental data. Only then, more detailed predictions of not observed parameters are warranted. Due to the very large size, so far analyses of entire large ensembles have been limited to the comparison of unbiased averages over the ensemble to measured experimental average values. Thus e.g. unbiased averages derived from even the most advanced MD force fields still fail to accurately predict the experimental data without further adjustments.[15]

Due to computational intractability, more detailed analyses of simulated ensembles have been restricted to much smaller size ensembles, i.e. typically on the order of at most several hundred conformers. Procedures such as Sample-And-Select (SAS),[17] the Ensemble Optimization Method (EOM),[18] ASTEROIDS,[19] and Sparse Ensemble Selection (SES)[20] select smaller subsets by various strategies from initially created large ensembles that satisfy the measured parameters. Similarly, Maximum Entropy (ME) reweighting of individual conformers,[11,21] Bayesian estimation of individual conformer weights,[22] and Maximum Occurrence (MO),[10] which estimates the maximal possible occurrence of a conformer within an ensemble, have been used only on smaller selected subsets or clusters, but not on entire very

large ensembles. Minimal-size ensembles compatible with experimental data may also be generated by constrained ensemble structure calculations.[6] Besides the recently proposed SES,[20] all proposed methods use stochastic mathematical procedures such as genetic algorithms or simulated annealing and their solutions are not guaranteed to be optimal and unique.

Here, we show that very large ensembles can be analyzed in full and very efficiently by a Maximum Entropy approach that reweights all individual populations in the ensemble such that the average over the ensemble reproduces the experimental data within the experimental error ($\chi^2 \leq 1$). This constrained search for the maximum entropy S^{max} falls into the class of convex optimization problems, which can be solved in a very efficient and deterministic manner even for very large data sets. As the population weights are calculated in a robust manner on the entire ensemble, the reweighted large ensemble represents the most accurate representation of the combination of simulation and experiment in an information-theoretical sense. Furthermore, since S^{max} is a well-defined parameter, its reduction relative to an unconstrained ensemble presents the true measure of the information content of experimental data relative to the simulated ensemble. We call the method COPER for Convex OPtimization for Ensemble Reweighting. As example, we analyze an ensemble of 35000 snapshots of a 700 ns MD trajectory of the nonapeptide EGAAWAASS in water, for which we had previously obtained extensive RDC, *J*-coupling and chemical shift data.[23,24] The results show that the unconstrained MD simulation overestimates the α-helical content. However, reweighting factors are moderate, corresponding to free energy changes of 2.6 *kT*, which are within the expected inaccuracy of MD force fields. A very strong discontinuity observed in the RDCs around the central tryptophan residue can be explained by a cluster of helical conformations of the central residues.[24] Not surprisingly, reweighting of a 35000-member ensemble generated from a random coil model of the unfolded state by the program Flexible-Meccano[8] requires a larger free energy change of 3.7 *kT*, consistent with its coarser nature of approximation. In contrast, a similar analysis carried out for the nonapeptide EGAAIAASS indicates largely extended conformations and much smaller necessary reweighting factors for its MD trajectory. As a corollary we show that also reweighting populations for χ^2 minimization and the Maximum Occurrence method[10] are convex optimization problems that can be solved in an equally efficient, deterministic manner.

Theory

Maximum entropy reweighting as a convex optimization problem

We consider an ensemble of N members with populations p_i ($0 \le p_i \le 1, \sum_i p_i = 1, i = 1,...,N$). Its entropy S in the sense of Shannon[25] is given as

$$S = -\sum_{i=1}^{N} p_i \ln p_i \tag{1}$$

Let $d_j^{\exp}$ be one of M measured experimental parameters ($1 \le j \le M$) and $d_{i,j}^{pred}$ its predicted value for the i-th member of the ensemble. Its predicted weighted average d_j^{pred} over the ensemble is then given as

$$d_j^{pred} = \sum_{i=1}^{N} p_i d_{i,j}^{pred} \tag{2}$$

Eq. 2 holds for cases where each conformation can be treated individually and the experimental parameter is a population-weighted average. Many NMR parameters such as chemical shifts, residual dipolar couplings, paramagnetic relaxation enhancements and J-couplings fulfill this condition. Assuming that the experimental system is ergodic, the ensemble average also equals the time average of an individual member. We apply this here to the prediction of NMR parameters from the average over the time frames from a MD trajectory, for which we assume that it is long enough for convergence.

The quality of the agreement between predicted average and the experimental data is judged by χ^2

$$\chi^2 = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{d_j^{pred} - d_j^{\exp}}{\sigma_j}\right)^2 \tag{3}$$

where $\chi^2 \le 1$ signifies agreement within the error limits σ_j. The error σ_j for the parameter j in Eq. 3 presents the total error, e.g. composed of the error of the measurement $\sigma_{j,\text{expt}}$ and the error of the model $\sigma_{j,\text{model}}$, i.e. $\sigma_j = \sqrt{\sigma_{j,\text{expt}}^2 + \sigma_{j,\text{model}}^2}$.

The maximum entropy search problem can now be formulated as the following optimization problem

Maximize $S(\boldsymbol{p})$ (4a)

Subject to $\chi^2(\boldsymbol{p}) \le 1$ (4b)

$$0 \le p_i \le 1, \quad i = 1,...,N \tag{4c}$$

$$\sum_{i=1}^{N} p_i = 1 \tag{4d}$$

Here the vector $\boldsymbol{p} = (p_1,...,p_N)$ is the optimization variable of the problem. An optimal solution $\boldsymbol{p}^{S-\max}$ is found when the object function $S(\boldsymbol{p})$ has its maximal value among all vectors $\boldsymbol{p}$ that satisfy the inequality constraints (4b and 4c) and the equality constraint (4d).

A convex optimization problem is one where both the objective function and the inequality constraint functions are convex, whereas the equality constraint functions are affine.[26] A function f is convex when its epigraph, epi f (the set of points above or on the graph of f, $\operatorname{epi} f = \{(x,t) \mid x \in \operatorname{dom} f, f(x) \le t\}$) is a convex set. A set is convex, if for any two points of the set, the connecting straight line segment between the two points is also in the set (Figure 1A). Thus the convex inequality constraints define convex sets of feasible points. Similarly, the affine equality constraints define affine sets. A set is affine, if for any two points of the set, their entire connecting straight line is also in the set. Since intersections of convex and affine sets are convex, the combined conditions imposed by convex inequality and affine equality constraints define a set of feasible points, which is also convex[26] (Figure 1B). Convex optimization problems can be solved very efficiently by interior point (IP) methods.[26,27] Figure 1B illustrates how the optimal solution can be reached from an interior point within the intersection of the feasible regions of all constraints. Starting from the interior point, the search follows the gradient of the objective function until the boundary of the set of feasible points is reached, from where the search continues along the boundary until the optimal solution is attained. If the optimum is located at an interior point, the problem reduces to an unconstrained optimization. The convex nature of the objective function ensures that the solution is unique in both cases.

It is easy to show that the negative of the entropy $-S(\boldsymbol{p})$ (Eq. 4a) and the constraining functions of the inequality constraints (Eq. 4b,c) are convex since their Hessians are positive semi-definite.

$$\frac{\partial^2}{\partial p_a \partial p_b} - S(\boldsymbol{p}) = \frac{\delta_{ab}}{p_a} \ge 0 \tag{5a}$$

$$\frac{\partial^2}{\partial p_a \partial p_b} \chi^2(\boldsymbol{p}) = \frac{2}{M} \sum_{j=1}^{M} \frac{d_{a,j}^{pred}}{\sigma_j} \frac{d_{b,j}^{pred}}{\sigma_j} \tag{5b}$$

where δ_{ab} is the Kronecker delta.

We also note that the constrained χ^2 minimization problem

$$\text{Minimize} \qquad \chi^2(\boldsymbol{p}) \tag{6a}$$

$$\text{Subject to} \qquad 0 \le p_i \le 1, \quad i = 1,...,N \tag{6b}$$

$$\sum_{i=1}^{N} p_i = 1 \tag{6c}$$

is a convex optimization problem, and hence can be solved very efficiently.

In order to find the maximum entropy according to Eq. 4 by interior point methods, we use the following two-step procedure:

1. search for the population vector $\boldsymbol{p}^{\chi^2\text{-min}}$ that minimizes χ^2 under the constraints of Eq. 6b-c starting from an interior point such as $p_i^{\text{equal}} = 1/N$. If $\chi^2(\boldsymbol{p}^{\chi^2\text{-min}}) \le 1$, $\boldsymbol{p}^{\chi^2\text{-min}}$ is an interior point for the constrained maximum entropy problem Eq. 4, otherwise it has no solution.

2. search for the population vector $\boldsymbol{p}^{S\text{-max}}$ that maximizes the entropy under the constraints of Eq. 4b-d starting from the interior point $\boldsymbol{p}^{\chi^2\text{-min}}$.

Change of entropy and free energy under reweighting

When no experimental information is present, i.e. the chi-square condition (4b) is dropped from the optimization problem of Eq. 4, the maximum entropy is achieved when all populations are equal and $p_i^{\text{equal}} = 1/N$. In this situation, the entropy takes the value $S(\boldsymbol{p}^{\text{equal}}) = \ln(N)$. The change in population weights due to the experimental information under maximum entropy principle leads to a decrease in entropy ΔS from this value

$$\Delta S = S(\boldsymbol{p}^{S\text{-max}}) - S(\boldsymbol{p}^{\text{equal}}) = -\sum_{i=1}^{N} p_i^{S\text{-max}} \ln p_i^{S\text{-max}} - \ln(N) \tag{7}$$

This decrease in entropy coincides with the definition of the relative entropy (Kullback-Leibler divergence)[28] ΔS_{AB} of two populations $\boldsymbol{p}^A$ and $\boldsymbol{p}^B$

$$\Delta S_{AB} = -\sum_{i=1}^{N} p_i^A \ln\left(\frac{p_i^A}{p_i^B}\right) \tag{8}$$

for the case $p_i^B = p_i^{\text{equal}} = 1/N$. The negative of the relative entropy ΔS_{AB} presents the mean information $I(A:B)$ for discrimination in favor of $\boldsymbol{p}^A$ against $\boldsymbol{p}^B$. Thus $-\Delta S$ in Eq. 7 is the information content of experimental data for discrimination against an equal population.

To quantify the reweighting of individual populations under the experimental constraints and the maximum entropy principle, we define the reweighting factor r_i and its associated free energy change ΔG_i

$$r_i = p_i^{S\text{-max}} / p_i^{equal} \tag{9a}$$

$$\Delta G_i = -kT \ln(r_i) \tag{9b}$$

where k is the Boltzmann constant and T the absolute temperature. Using Eqs. 7-9, it is obvious that $kT\Delta S$ presents the mean free energy change $\langle \Delta G \rangle$

$$\langle \Delta G \rangle = \sum_{i=1}^{N} p_i^{\text{S-max}} \Delta G_i = -kT \sum_{i=1}^{N} p_i^{\text{S-max}} \ln\left(\frac{p_i^{\text{S-max}}}{p_i^{\text{equal}}} \right) = kT\Delta S \tag{10}$$

Results and discussion

Experimental NMR data and MD simulations on the EGAAWAASS nonapeptide

Previously, we had systematically investigated the influence of single amino acid substitutions X on the conformation of unfolded model peptides EGAAXAASS as monitored by $^1D_{\text{NH}}$ and $^1D_{\text{C}\alpha\text{H}\alpha}$ RDCs, $^3J_{\text{HNH}\alpha}$ scalar couplings, and $^{13}\text{C}^\alpha$ secondary shifts.[23] Homogeneous RDC, chemical shift and *J*-coupling values along the peptide sequence indicated extended peptide conformations for most amino acid types X. However, substitutions by the aromatic amino acids tryptophan and tyrosine led to a kink in the center of the peptide as evident from a discontinuity in the NMR data. The original NMR data were obtained on peptides at natural abundance of ^{13}C and ^{15}N. To obtain access to further NMR parameters, the tryptophan-substituted EGAAWAASS peptide was ^{13}C- and ^{15}N-isotope labeled in a bacterial expression system.[29] Figure 2 shows sequential RDC ($^1D_{\text{NH}}$, $^1D_{\text{C}\alpha\text{H}\alpha}$, $^1D_{\text{C}\alpha\text{C}'}$) and *J*-coupling ($^3J_{\text{HNH}\alpha}$, $^3J_{\text{H}\alpha\text{N}}$) data acquired on this isotope-labeled EGAAWAASS peptide (a complete list of experimental data is provided in Supporting Information Table S1). The discontinuity is evident in the sequence profile of the $^1D_{\text{NH}}$ and $^1D_{\text{C}\alpha\text{H}\alpha}$ RDCs. Whereas they are negative and positive, respectively, for almost all amino acids consistent with an extended conformation of the peptide in horizontally compressed polyacrylamide gels,[30] they change sign at the central residues A6 ($^1D_{\text{NH}}$) and W5 ($^1D_{\text{C}\alpha\text{H}\alpha}$) indicative of a kink. Figure 2 also shows experimental statistical error estimates for the *J*-coupling and RDC values (Supporting Information Table S1). The error estimates for the $^3J_{\text{HNH}\alpha}$, $^3J_{\text{H}\alpha\text{N}}$ couplings are very close to RMSD values found previously between experimental data and data predicted from structural knowledge by the respective Karplus parameters.[31,32] For the RDC data, a true error estimate is much harder to

establish due to the lack of detailed knowledge on the interactions of the peptide with the alignment medium and possible induced conformational changes during this interaction. We have previously observed a similar discontinuity in the RDC pattern with a different alignment medium (Pf1 phages),[23] which indicates that the kink in the peptide is not induced by interactions with the medium. Nevertheless, the RDC model error is unknown for flexible peptides. Since better estimates are not available, the total RDC error was taken as the experimental error.

To identify the structural reason for the kink in the peptide, we have carried out a total of seven, 100-ns MD simulations on the EGAAWAASS peptide under full hydration. Theoretical RDCs and *J*-couplings were then calculated for every 20-ps frame using a steric exclusion model[6] and available Karplus parameters.[31,32] Figure 2 shows the equally weighted averages of the different observables over the total of 35000 conformations (green solid lines). Clearly, the unbiased averages do not reproduce the experimental data (blue) within the indicated error, which is particularly noticeable for the kinks observed in the experimental RDC data in the region around residues W5 and A6. These deviations lead to a total χ^2 value (Eq. 3) of 116.

COPER procedure: χ^2-minimization followed by entropy maximization

The COPER procedure was then applied to reweight the individual conformations. In practice, using a single total ($\chi^2 \leq 1$) constraint for all different data types in the Maximum Entropy search led to a very uneven distribution of deviations among the different RDC and *J*-coupling data types. Therefore we rather used individual $\chi^2_\alpha \leq 1$ constraints for each of the different data types α (RDC or *J*-coupling). To find a feasible inner point for this Maximum Entropy search, the initial χ^2 minimization was then carried out on the sum $\sum_\alpha \chi^2_\alpha$, which differs from the original χ^2 definition in Eq. 3 only by reweighting via the number of data points M_α in the individual data sets. Minimization of $\sum_\alpha \chi^2_\alpha$ within the usual constraints on population weights (Eq. 6b-c) of the 35000 conformations led to very good agreement of the average data predicted from the minimizing population vector $\boldsymbol{p}^{\chi^2\text{-min}}$ with the experimental data (Figure 2, dashed magenta lines). As compared to the equal population entropy $S(\boldsymbol{p}^{\text{equal}}) = \ln(35000) = 10.46$, the entropy for the $\boldsymbol{p}^{\chi^2\text{-min}}$ vector is significantly reduced to a value $S(\boldsymbol{p}^{\chi^2\text{-min}}) = 5.59$ ($\Delta S = -4.87$, Table 1).

The minimized $\sum_\alpha \chi^2_\alpha$ value of 0.37 (Table 1) guarantees that also the individual χ^2_α values are smaller than 1, and hence $\boldsymbol{p}^{\chi^2\text{-min}}$ presents a feasible starting point for the Maximum Entropy

search within the χ^2 and population constraints (Eq. 4b-d). The subsequent Maximum Entropy search then yielded average predicted data that agree less well to the experimental data than the minimal χ^2 prediction, but are still within the error limits (Figure 2, red lines). Consequently the entropy is again increased to a value $S(\boldsymbol{p}^{S\text{-max}}) = 7.95$, but still reduced relative to the equal population situation by $\Delta S = -2.51$. This reduction in entropy corresponds to the minimal restriction of the accessible conformational space needed to satisfy the experimental information. In the simplest case, it may be pictured as making certain conformations completely inaccessible, whereas the accessible conformations remain equally likely. In this situation, $S(\boldsymbol{p}^{S\text{-max}})$ would correspond to the logarithm of the number of accessible conformations, and hence $e^{-\Delta S}$ (12.3 for the current case) to the factor by which the number of accessible conformations is reduced due to the experimental information.

Due to the efficiency of the inner point method, the entire COPER procedure of constrained χ^2 minimization followed by constrained entropy maximization took only about 9 minutes to complete on a single core of a 2.6-GHz Intel Xeon CPU for the 35000-member EGAAWAASS peptide ensemble. Tests with different ensemble sizes showed that this time increased approximately linearly with ensemble size for ensembles of up to 70000 members.

Robustness test, effect of error uncertainty, and information content of individual data types

To estimate the information content of different types of NMR constraints on the MD ensemble, we systematically determined the entropy loss induced by these constraints via COPER fitting relative to the equilibrium population. Table 1 lists these losses for different combinations of scalar and dipolar couplings. In order to estimate the errors and robustness of the method, the set of 35000 conformations from the MD trajectory was further subdivided into two randomly chosen subsets of 17500 conformations, for which the COPER fit procedure was repeated and the entropy calculated. Table 1 lists the average and standard deviations of the resulting entropy reductions for the two subpopulations relative to their equal population entropy $S(\boldsymbol{p}^{\text{equal}}) = \ln(17500) = 9.77$. It is obvious that the entropy losses are highly reproducible, with relative standard deviations of less than 7 %, and very close to the losses calculated for the 35000 conformation data set. This indicates that the sampling of the conformational space is dense, since significant variations of the entropy reduction would be expected for a too low sampling of conformational space. In this manner, the comparison of the entropy reduction within different subsets of an ensemble provides both a test for the robustness of the reweighting and for the density of sampling.

To assess the effect of the used errors on the total entropy reduction, we have also varied the limits for χ^2_α from 0.25 to 4, corresponding to a scaling of the errors by factors between 0.5 and 2 (Supporting Information Figure S2A). Consistent with the expectation that weaker constraints allow larger conformational entropy and with previous findings by Hummer and colleagues,[11] the entropy reduction decreases monotonously with increasing χ^2_α limits. A limit of $\chi^2_\alpha \le 4$ decreases the entropy reduction to 1.75 from its value of 2.51 for $\chi^2_\alpha \le 1$. Thus a twofold increase of the error size has an effect of less than 1 *kT*-unit on the free energy change.

The entropy reduction induced by the experimental constraints may range from zero, for which the reweighted population is identical to the equally populated state, to $\ln(N)$, for which the constrained ensemble reduces to a single conformation (equivalent to a final entropy value of zero). The entropy losses due to ${}^3J_{\mathrm{HNH}\alpha}$ (ϕ-angle) or ${}^3J_{\mathrm{H}\alpha\mathrm{N}}$ (ϕ-angle) constraints are 0.05 and 0.22, respectively (Table 1). Thus the ${}^3J_{\mathrm{HNH}\alpha}$ data carry three times less information than the ${}^3J_{\mathrm{H}\alpha\mathrm{N}}$ data. This is in agreement with the fact that among the different conformations accessible to the polypeptide, i.e. helical vs. extended, variations in ϕ-angle are much smaller than in ψ-angle. The entropy losses for individual ${}^1D_{\mathrm{NH}}$, ${}^1D_{\mathrm{C}\alpha\mathrm{H}\alpha}$, and ${}^1D_{\mathrm{C}\alpha\mathrm{C}'}$ constraints are 0.67, 0.54, and 0.46 respectively, which indicates a significantly higher information content of the dipolar couplings relative to the scalar couplings. Combining all three dipolar couplings constraints increases the entropy loss to 2.30, which is a more than additive effect on the restriction of conformational space. Finally, when both dipolar and scalar coupling constraints are applied simultaneously, the total entropy loss of 2.51 approximately equals the sum of their individual contributions. This shows that the dipolar and scalar coupling constraints each contain information not captured by the other data type. Since the entropy loss times the thermal energy *kT* represents the mean free energy change (Eq. 10), an adjustment of the MD force field by 2.51 *kT*-units would be necessary to bring the ensemble in agreement with the experiment.

Comparison to Flexible-Meccano random coil ensemble

The entropy reduction presents a measure of the accuracy of the model ensemble. This can be used to quantitatively compare different types of ensembles. For this, we created a 35000-conformation ensemble based on a random coil model of the unfolded with residue-specific ϕ/ψ propensities using the program Flexible-Meccano.[8] Reweighting its populations by COPER using the same backbone *J*-coupling and RDC constraints as for the MD ensemble caused an entropy reduction by 3.61 (Table 1), which is more than the corresponding value of

2.51 for the MD ensemble and consistent with the considerably coarser nature of approximation used in Flexible-Meccano.

It is interesting to note that the $\sum\chi_\alpha^2$ difference between the experimental and back-calculated data for the MD ensemble before reweighting amounts to 968 and is therefore only slightly smaller than the corresponding $\sum\chi_\alpha^2$ value of 1158 for the Flexible-Meccano ensemble (Table 1). However, the minimized $\sum\chi_\alpha^2$ drops to 0.37 for the MD, but only to 0.98 for the Flexible-Meccano ensemble. Thus, the MD ensemble contains conformations, which as population-weighted combinations better represent the experimental data than the Flexible-Meccano ensemble. Since the $\sum\chi_\alpha^2$ minimum is lower for the MD ensemble, it is expected that the $\chi_\alpha^2 \leq 1$ conditions lead to a larger allowed space for the population weights and in consequence to less reduction in entropy.

Cross-validation of COPER ME populations

Cross-validation of ME-derived populations with additional, independent experimental data is problematic, since the ME solution is per definition underdetermined. If the predictions agree with the additional data, their information is redundant and they would not have constrained the original fit. In contrast, if the additional data deviate from the predictions of the original fit, they contain independent information. Using COPER, the information content of the additional data can be estimated from the entropy reduction that results from including these data in the fit.

Using this quantitative concept, populations of the 35000-frame MD data set obtained by the COPER ME fit of the ${}^1D_{\mathrm{NH}}$, ${}^1D_{\mathrm{C}\alpha\mathrm{H}\alpha}$, ${}^1D_{\mathrm{C}\alpha\mathrm{C}'}$, ${}^3J_{\mathrm{HNH}\alpha}$, ${}^3J_{\mathrm{H}\alpha\mathrm{N}}$ couplings (Figure 2) were cross-validated by experimentally determined χ_1-angle populations of the W5 side chain. Assuming staggered conformers, the χ_1-angle populations were derived by a simple linear transformation (linear least squares fit) from experimental ${}^3J_{\mathrm{NC}\gamma}$ and ${}^3J_{\mathrm{C'C}\gamma}$ couplings[33,34] (Supporting Information Table S1), which had not been used as constraints (Figure 3). These experimental populations of the χ_1 +60°, +180° and -60° rotamers are 22, 46, and 31 %, respectively. The χ_1 populations from the MD simulation (58, 34, 6 %) deviate strongly, but get closer (44, 45, 12 %) to the experimental values after COPER reweighting by the ${}^1D_{\mathrm{NH}}$, ${}^1D_{\mathrm{C}\alpha\mathrm{H}\alpha}$, ${}^1D_{\mathrm{C}\alpha\mathrm{C}'}$, ${}^3J_{\mathrm{HNH}\alpha}$, and ${}^3J_{\mathrm{H}\alpha\mathrm{N}}$ data, thereby confirming the correct trend of the independent fit. Obviously, including the ${}^3J_{\mathrm{NC}\gamma}$ and ${}^3J_{\mathrm{C'C}\gamma}$ scalar couplings in the COPER procedure leads to the best

agreeing χ_1 populations (31, 40, 28 %) at a cost of reducing the entropy by 2.63 relative to the equal population situation (Table 1). However, this reduction is only 0.12 larger than for the fit without the side chain $^3J_{NC\gamma}$ and $^3J_{C'C\gamma}$ scalar couplings. Therefore, their additional information content is rather small and reduces the conformational space only by an additional 11 %.

Structural interpretation by ϕ/ψ cluster analysis

To obtain structural insights into the effects of the ME reweighting, the 35000 conformations were clustered into 20 clusters based on the similarity of the ϕ and ψ torsion angles of the central five residues (A3-A7) using a hierarchical clustering algorithm. The clusters were ordered according to the size of their populations in the original MD trajectory. Figure 4A shows the ϕ/ψ angle distributions of the four most highly populated clusters accounting for 66 % of all conformations. The largest cluster 1 has α-helical conformations for residues A4 to A6 and partially α-helical conformations for residue A3 and A7, whereas the other clusters contain more extended conformations. A representative set of conformations of cluster 1 is shown in Figure 4B. It is obvious that residues A3 to A7 form a turn with backbone hydrogen bond contacts. These contacts are protected from the water by the bulky aromatic side chain of residue W5 as shown in a recent analysis[24] of the full MD trajectory, which explains the tendency of the aromatic groups to induce kinks in the unfolded peptide chain.

Figure 5A shows the 20 cluster populations before and after reweighting by the $^1D_{NH}$, $^1D_{C\alpha H\alpha}$, $^1D_{C\alpha C'}$, $^3J_{HNH\alpha}$, $^3J_{H\alpha N}$ COPER fit. Before reweighting, cluster 1 has a population of about 33 %, whereas the other clusters have populations of less than 12 %. After reweighting, the population of cluster 1 decreases significantly to about 15 %, the population of cluster 2 decreases and those of 3 and 4 increase. The rest of the cluster populations remain below 10 %. To test the statistical significance of this result, the cluster populations after reweighting were also determined for the two randomly selected subsets of 17500 conformations. Figure 5A also shows their averages and standard deviations. The maximal standard deviation of populations in the 17500-conformation sets is only 3 %, and their averages agree within this limit to the results from 35000-conformation set. Thus the reproducibility of the COPER-derived populations is very high.

We have also assessed the effect of the used errors on the cluster populations by varying the limits for χ^2_α from 0.25 to 4 (Supporting Information Figure S2B). Again as for the induced entropy changes, the cluster populations vary monotonously with the χ^2_α limits. This is a

further indication of the robustness of the results. For a change of the χ^2_α limits from 1 to 4, the populations for most clusters vary by less than twofold, with cluster 1 always remaining the dominant cluster.

The reduction of the population of cluster 1 caused by the COPER reweighting with experimental data indicates an overestimation of helical content by the AMBER03 force field, which is in agreement with findings by Best, Lindorff-Larssen and colleagues.[15,35] It is noted that this reduction of the helical cluster 1 is stronger than reported previously.[24] This is caused by the $^3J_{\mathrm{H}\alpha\mathrm{N}}$ couplings not present in the previous study, which increased the content of extended conformations. However, even after reweighting by COPER with these additional data, the helical cluster 1 remains the most highly populated cluster, albeit closely followed by clusters 3 and 4 (Figure 5A).

The relative changes in the cluster populations due to the COPER reweighting are shown in Figure 5B as $\ln(p_{\mathrm{COPER}}/p_{\mathrm{equal}})$, where p_{equal} and p_{COPER} represent the cluster populations before and after reweighting, respectively. Values for $\ln(p_{\mathrm{COPER}}/p_{\mathrm{equal}})$ range between about -0.9 to 2.4 corresponding to errors on the order of less than 3 kT in the free energy of the individual clusters.

Results for the EGAAIAASS nonapeptide

As indicated, in contrast to the kinked form of EGAAXAASS peptides with aromatic amino acids X in their center, peptides with other amino acids besides proline and glycine showed extended conformations from the sequence profile of their NMR parameters.[23] We further tested the reweighting of a 10000-conformation trajectory of the prototypical extended EGAAIAASS peptide, for which the published $^1D_{\mathrm{NH}}$, $^1D_{\mathrm{C}\alpha\mathrm{H}\alpha}$, $^3J_{\mathrm{HNH}\alpha}$ values were used as input for the COPER ME method. As for the EGAAWAASS the conformations were clustered into 20 clusters based on the ϕ and ψ torsion angles of the residues A3-A7. Figure 6A shows the ϕ/ψ distributions of the four most highly populated clusters before reweighting. In this case, the most highly populated (18 %) cluster 1 has almost completely extended conformations with only a slight admixture of helical conformations for residue A4. Clusters 2-4 have about 10 % populations and are mostly extended (cluster 2), mixed extended/helical (cluster 3), and mostly helical (cluster 4). COPER reweighting reduced the total χ^2 value from 8.8 to 1.0, but changed the individual cluster populations by less than 2 % (Figure 6B). Accordingly, the reweighting factors $\ln(p_{\mathrm{COPER}}/p_{\mathrm{equal}})$ only ranged from about -0.2 to 0.1 (Figure 6C), showing that the free energy adjustment is less than 0.2 kT. The total entropy loss due to the reweighting was only 0.11. Apparently the AMBER03 force field in conjunction with the TIP4P water model

reproduced the extended conformations of the EGAAIAASS peptide almost quantitatively, whereas it significantly exaggerated the more helical conformations of the EGAAWAASS peptide.

Comparison of COPER with other ensemble reweighting algorithms

The COPER approach may be compared to the previously proposed Maximum Entropy[11,21] and Bayesian[22] ensemble reweighting algorithms. These previous methods all contained non-deterministic random sampling algorithms and due to computational efficiency had to be restricted to smaller subsets (at most several thousand structures) from computed ensembles of tens of thousands of structures. In contrast, due to the efficiency of the inner point convex optimization method and the use of only gradients of the objective and constraining functions,[36] COPER can calculate globally optimized weights in a very efficient, numerically stable, and deterministic manner for very large ensembles of so far up to 70'000 structures. We note that this limit is rather dictated by numerical precision and not by computational speed.

Besides this advantage in efficiency and the well-definedness of the solution, the underlying mathematical target of COPER also differs from the previous approaches. The described Maximum Entropy approaches[11,21] minimize a free energy, in which an entropy term was subtracted from χ^2. Thus Hummer and colleagues[11] use the free energy function $G = \chi^2 - \Theta S$ where Θ is a tunable temperature parameter that balances the agreement between experimental and back-calculated data with conformational diversity. Θ is then varied until the corresponding free energy change matches an expected error in the force field. In COPER, we rather define the error of the parameters, i.e. χ^2, and obtain as a result the entropy change ΔS and the concomitant free energy change ΔG. Thus as shown in Supporting Information Figure S2, the relation between ΔS and χ^2 can easily be established. This relation can also be used to achieve a certain ΔG ($=kT\,\Delta S$), which matches expected force field errors. Using our error estimates, the free energy changes of less than 3-kT were in the range of the expected force field errors.

In contrast to the Maximum Entropy approaches, the Bayesian[22] ensemble reweighting algorithm determines the population weights from assumed prior distributions of the weights and likelihood functions of the parameters based on experimental, theoretical or assumed errors. This approach also provides estimates of the uncertainties in the weights, which are not easily obtained by other methods. However, the computational cost is rather high and so far it has only been applied to small ensembles of hundreds of conformations.

Extension of the inner point convex optimization to Maximum Occurrence

Bertini and colleagues previously have introduced the method of maximum occurrence (MO)[10] for the analysis of ensembles of flexible macromolecules. The method tries to determine the maximum time or occurrence that a molecule can spend in a given conformation k such that the weighted average over all conformations of a theoretical ensemble is still compatible with the experimental average data. The problem can thus be formulated as

$$\text{Maximize} \quad p_k \tag{11a}$$

$$\text{Subject to} \quad \chi^2(\boldsymbol{p}) \le 1 \tag{11b}$$

$$0 \le p_i \le 1, \quad i = 1,...,N \tag{11c}$$

$$\sum_{i=1}^{N} p_i = 1 \tag{11d}$$

where the populations p_i and the constraining function χ^2 are defined as in Eq. 4. Previously, this problem could only be solved by using a non-deterministic, simulated annealing procedure on smaller subsets (480 families of 50 members) of a large ensemble (56000 structures).[10] However, since the target function p_k (Eq. 11a) and the constraints (Eq. 11b-d) are convex or affine, the entire problem is a convex optimization problem that can be solved efficiently by the described inner point method.

While it is beyond the scope of the present work to perform a detailed analysis of the EGAAWAASS peptide conformations by the MO method, we have tested the efficiency of this inner point solution to the MO problem on ensembles of random conformations generated for this peptide by the program Flexible-Meccano.[8] The ensembles ranged in size from 10000 to 70000 members and were subjected to the MO optimization using the experimental RDC and *J*-coupling backbone constraints described in Figure 2. The CPU time necessary to calculate one MO population increased approximately linearly with the ensemble size and amounted to 850 seconds on a single core of a 2.6-GHz Intel Xeon CPU for the 70000-member ensemble. This compares very favorable with the 6 hours reported previously for subsets of a 56000-member ensemble.[10]

Conclusions

We have presented the ME method COPER using inner point convex optimization to reweight large simulated conformational data sets by average experimental data. Compared to

previous methods, COPER can analyze full, very large ensembles of 10^4 to 10^5 conformers and not only smaller subsets thereof in a deterministic, fast, and robust manner. The convex optimization guarantees a global unique optimal solution, and hence a reliable determination of the final population weights for the full ensembles. Hence, such reweighted ensembles constitute the best representation of the information contained in both the simulated ensemble and the experimental data. Since the final entropy is determined reliably, its loss relative to the unconstrained ensemble can be used as a quantitative measure of the information content of experimental data relative to the theoretical ensemble. A large reduction in entropy will indicate that the theoretical ensemble is not a good representative of the real-world situation and hence the simulation needs to be improved. However, the measure can also be used to judge the information content of individual data types, e.g. a comparison of the entropy reduction induced by the different NMR data types clearly revealed the much higher information content of RDCs relative to three-bond *J*-couplings. Furthermore, the reproducibility of the entropy reduction on different subsets of the large ensembles provides an estimate for its density of sampling of the conformational space. Thus if the reproducibility becomes low, a larger number of structures needs to be generated in the initial ensemble to cover the space adequately.

The application to the reweighting of the MD trajectories of small peptides by NMR data showed that the AMBER03 force field overestimated the helical content for the turn-forming EGAAWAASS peptide, but not for the extended EGAAIAASS peptide. The reduction in entropy was in all cases smaller than 3, indicating that adjustments of the force field of less than 3 kT-units would be needed to bring the MD trajectory into agreement with the experimental data. An ensemble created by the Flexible-Meccano statistical coil model of the EGAAWAASS peptide needed stronger reweighting than the MD-derived ensemble to fit the experimental data, consistent with the cruder nature of this model. Eventually, such COPER-reweighted populations may be used via projection onto some essential coordinates to improve existing MD force fields by free energy perturbation methods.[37] Compared to pure χ^2 minimization for force field optimization,[38,39] this may have the advantage to reduce the risk of overfitting,[40] since the entropy is maximized.

While the application of COPER was shown here for average NMR data, in fact it is applicable to any experimental average data that can be predicted from a set of molecular conformations, such as small-angle X-ray scattering[41] or Förster resonance energy transfer[42] data. Furthermore, convex optimization can provide similar advantages of well-defined, robust

solutions and large-scale efficiency for other ensemble analysis methods with convex target functions and constraints such as constrained χ^2 minimization and MO.[10]

Materials and Methods

Sample preparation

Uniformly $^{15}N/^{13}C$-labeled peptide EGAAWAASS was prepared by expression in *E. coli* as a C-terminal fusion with the immunoglobulin-binding domain of streptococcal protein G as described previously.[43] The peptide was cleaved bluntly from the fusion by factor Xa. NMR samples were prepared as 1 mM (0.25 mM) peptide, 25 mM acetate, pH 4.5 in 5/95 % D_2O/H_2O for measurement under isotropic (anisotropic) conditions. Residual alignment of peptides was achieved by introducing the peptide solutions into 10 % (w/v) polyacrylamide gels and horizontal compression.[44,45]

NMR experiments

All NMR experiments were carried out at 298 K on a Bruker Advance III 600 MHz spectrometer equipped with a TXI probe. Spectra were processed using NMRPipe.[46] $^3J_{NH\alpha}$ couplings were obtained from a quantitative-J version of the $^3J_{NH\beta}$-HNHB experiment using a 27 ms ^{15}N-$^1H^\alpha$ dephasing delay.[34,47] The resonance line shapes were fitted with the NLINLS program contained in NMRPipe and $^3J_{NH\alpha}$ coupling constants were determined from the ratios of cross and reference peak heights as described.[34] The $^3J_{HNH\alpha}$ values were taken from the work by Dames et al.[23] $^3J_{NC\gamma}$ and $^3J_{C'C\gamma}$ scalar couplings of the central W5 residue were determined by quantitative-J 2D constant-time ^{15}N-{$^{13}C^\gamma$} and $^{13}C'$-{$^{13}C^\gamma$} spin-echo difference experiments.[33] Error estimates for the quantitative-J measurements were obtained from the noise of the spectra.

$^1D_{C\alpha C'}$ RDCs were calculated as the difference in $^{13}C'$-$^{13}C^\alpha$ doublet splittings observed under anisotropic and isotropic conditions, which had been measured with a modified version of HNCO experiment, in which the 180-degree C^α decoupling pulse in the C' evolution was removed. Similarly, ^1H-^{15}N RDCs were obtained from ^1H-^{15}N HSQCs without ^1H decoupling during the ^{15}N evolution. A modified version of the HN(CO)CA experiment without ^1H decoupling in the $^{13}C^\alpha$ evolution period was used to detect $^1H^\alpha$-$^{13}C^\alpha$ RDCs. Each RDC experiment was carried out twice, and the reported values and the error estimates refer to mean and standard deviations derived from such repeated experiments.

MD simulations

MD simulations were carried out with the GROMACS simulation package[48] using the AMBER03 force field.[49] Extended input starting structures of the peptides EGAAXAASS were generated using MOLMOL[50] and solvated in a dodecahedron box containing about 8700 TIP4P water molecules, three sodium ions and two chloride ions. The energy of the system was first minimized by the steepest descent method, followed by a 500-ps simulation for equilibration of solvent molecules with the position of the peptide kept fixed. Electrostatic interactions were implemented by particle-mesh Ewald (PME) summation with a grid spacing of 0.12 nm,[51] while the Lennard-Jones interactions had a cut-off at 1.4 nm. The integration time step was 2 fs. Production runs for 100 ns were carried out at a constant temperature of 300 K and pressure of 1 bar. 35000 (X=W) or 10000 (X=I) conformations were obtained as 20-ps frames sampled uniformly from 7 (X=W) and 2 (X=I) 100-ns trajectories started with different random seeds.

Back-calculation of NMR parameters

For every snapshot of the MD trajectory, theoretical RDCs were predicted based on a steric alignment model using an efficient algorithm described previously.[6] The RDC values of each conformation were scaled by a constant determined by a least square fit between the average RDCs of all conformations and the experimental RDC values of the peptide. Theoretical ^3J values (in Hz) were calculated using the following Karplus relations: $^3J_{\mathrm{HNH}\alpha}$ = 8.40 $\cos^2(\phi - 60°)$ -1.36 $\cos(\phi - 60°)$ + 0.33,[31] $^3J_{\mathrm{H}\alpha\mathrm{N}}$ = -1.00 $\cos^2(\psi - 120°)$ +0.65 $\cos(\psi - 120°)$ - 0.15,[52] $^3J_{\mathrm{NC}\gamma}(\mathrm{W5})$ = 1.29 $\cos^2(\chi_1)$ -0.49 $\cos(\chi_1)$ + 0.34 and $^3J_{\mathrm{C'C}\gamma}(\mathrm{W5})$ = 2.31 $\cos^2(\chi_1 - 120°)$ - 0.87 $\cos(\chi_1 - 120°)$ + 0.49.[53]

Clustering of MD conformations

To obtain structural insights, the ensemble of MD conformations for the EGAAXAASS peptides were divided into 20 clusters using the hierarchical clustering function of MATLAB (MathWorks, Inc) and a ϕ/ϕ angle distance metric $d(i,j)$ between individual conformations i and j

$$d(i,j) = \sqrt{\sum_{res} d_{ang}^2(\phi_{res}(i),\phi_{res}(j)) + d_{ang}^2(\psi_{res}(i),\psi_{res}(j))}$$

where the summation runs over the central residues A3 to A7 of the peptide to emphasize their conformation and the periodic angular distance metric d_{ang} is defined as

$$d_{ang}(\alpha,\beta) = \min\left(|\alpha-\beta|, 360° - |\alpha-\beta|\right).$$

The distance between two clusters was defined as the average of all the individual distances of their members.

Implementation of COPER

COPER was implemented using the IPOPT[36] open source software package written in C++ for large-scale nonlinear optimization. The IPOPT algorithm utilizes primal-dual interior-point methods[26] to find local solutions of optimization problems. COPER objective functions, constraints and their derivatives, as well as data input and output were coded in C and linked to IPOPT. To speed up the search for the maximum entropy solution, the optimization was implemented as a minimization of the convex function e^{-S} rather than as a maximization of the entropy S. COPER source code and compiled executables for several platforms are available from the authors upon request.

Default tolerances and the maximum numbers of iterations for the chi square minimization (entropy maximization) were set to 1×10^{-3} and 20000 (1×10^{-5} and 80000), respectively. Using these parameters, the total reweighting of the 35000-member EGAAWAASS peptide ensemble with 35 constraints took 560 seconds on a single core of a 2.6-GHz Intel Xeon CPU.

Acknowledgments

We gratefully acknowledge Prof. Olaf Schenk for initially pointing out the IPOPT algorithm to us and for very helpful discussions. This work was supported by a stipend from the Croucher Foundation (H.T.A.L) and Swiss National Science Foundation grant 31-149927 (S.G.). Computational resources were provided by the Basel Computational Biology Center (http://www.bc2.ch/).

Supporting Information Available

Table of experimental NMR constraints used for the conformational analysis of the EGAAWAASS peptide. This material is available free of charge via the Internet at http://pubs.acs.org.

Figure legends

Figure 1: Solution of constrained convex optimization problem by interior point method. A) Example of a convex and a not convex set. B) Illustration of interior point method. The intersection of all convex constraints (constraint 1 (green), constraint 2 (magenta), …) defines the convex set of feasible points (yellow). The value of the convex objective function is shown by the blue, dashed contour lines. The search starts from an interior point (red circle) within the set of feasible points and follows the gradient of the objective function until the boundary of the set of feasible points is reached. The search then continues along the boundary following the gradient of the objective function until the optimal solution (red circle) is attained (see text).

Figure 2: Comparison of experimental RDCs and backbone scalar couplings obtained on the nonapeptide EGAAWAASS to values back-calculated from its 35000-frame MD trajectory. Experimental data are shown as blue circles and the unbiased (equal population) average of the predicted observables from the trajectory as green lines. The reweighted averages after χ^2 minimization and the COPER entropy maximization are indicated as dashed magenta lines and red lines respectively.

Figure 3: Cross validation of the COPER-reweighted populations by χ_1 rotamer populations determined independently from ${}^3J_{NC\gamma}$ and ${}^3J_{C'C\gamma}$ scalar couplings for the side chain of W5 in the EGAAWAASS peptide. Experimentally determined populations are shown in blue, unbiased populations from the 35000-frame MD trajectory in green, COPER-reweighted populations according to backbone RDCs and *J*-couplings (${}^1D_{NH}$, ${}^1D_{C\alpha H\alpha}$, ${}^1D_{C\alpha C'}$, ${}^3J_{HNH\alpha}$, ${}^3J_{H\alpha N}$) in red, and COPER-reweighted populations according to backbone and side chain RDCs and *J*-couplings (${}^1D_{NH}$, ${}^1D_{C\alpha H\alpha}$, ${}^1D_{C\alpha C'}$, ${}^3J_{HNH\alpha}$, ${}^3J_{H\alpha N}$, ${}^3J_{NC\gamma}$, ${}^3J_{C'C\gamma}$) in magenta.

Figure 4: Clustering of the 35000 conformations from the MD trajectory of the EGAAWAASS peptide according to the ϕ/ψ angles of its central five residues (A3-A7). A) Ramachandran population plots of the four most highly populated clusters are shown with contour levels spaced by a factor of 2.5. The most highly populated cluster 1 has α-helical conformations for residues A4 to A6 and partially α-helical conformations for residues A3 and A7. B) Overlay of 8 representative conformations from cluster 1 where residues A3 to A7 form a helical turn. Backbone hydrogen contacts in this turn are shielded from external water by the side chain of W5.[24]

Figure 5: Reweighting of populations for the 20 clusters from the 35000 MD conformations of the EGAAWAASS peptide. A) Populations of the clusters before reweighting are shown in red and after COPER-reweighting in blue. For testing the robustness the 35000 conformations were split into two 17500-conformation sets. Averages and standard deviations of the cluster populations of these two subsets after COPER reweighting are shown in green. B) Reweighting factors for the cluster populations indicated as $\ln(p_{COPER}/p_{equal})$, where p_{equal} and p_{COPER} represent the populations before and after COPER reweighting, respectively. Data for the COPER analysis of the 35000 conformations and of the two 17500-conformation subsets are shown in blue and green, respectively.

Figure 6: Analysis of the 10000 conformations from the MD trajectory of the EGAAIAASS peptide. The conformations were clustered into 20 subsets according to the ϕ/ψ angles of its central five residues (A3-A7). A) Ramachandran population plots of the four most highly populated clusters are shown with contour levels spaced by a factor of 2.5. The most highly populated cluster 1 has extended conformations for residues A3 and I5 to A7 and partially α-helical conformations for residue A4. B) Populations of the 20 clusters before reweighting are shown in red and after COPER-reweighting in blue. C) Reweighting factors for the cluster

populations indicated as $\ln(p_{COPER}/p_{equal})$, where p_{equal} and p_{COPER} represent the populations before and after COPER reweighting, respectively.

Tables

Table 1: χ^2 and entropy values[a] before reweighting, after χ^2 minimization and after entropy maximization of the frames of the EGAAWAASS nonapeptide MD trajectory or its Flexible-Meccano data set using different NMR observables as constraints

		before reweighting[b]	after χ^2 minimization		after entropy maximization			
constraints	N[c]	$\sum_\alpha \chi_\alpha^2$ [d]	$\sum_\alpha \chi_\alpha^2$	ΔS[e]	$\sum_\alpha \chi_\alpha^2$	ΔS[e]	ΔS average[f]	ΔS S.D.[f]
$^3J_{HNH\alpha}$	7	3.52	0.00	-0.201	1.00	-0.045	-0.044	0.000
$^3J_{H\alpha N}$	5	4.41	0.00	-0.878	1.00	-0.215	-0.199	0.006
$^3J_{NC\gamma}$	1	4.42	0.00	-0.043	0.00	-0.043	-0.008	0.001
$^3J_{C'C\gamma}$	1	11.77	0.00	-0.277	0.98	-0.195	-0.210	0.070
$^1D_{NH}$	8	498.97	0.00	-1.415	1.00	-0.668	-0.652	0.008
$^1D_{C\alpha H\alpha}$	7	383.92	0.00	-0.905	1.00	-0.540	-0.540	0.013
$^1D_{C\alpha C'}$	8	35.56	0.01	-1.387	1.00	-0.459	-0.459	0.000
all backbone ^3J[g]	12	4.41	0.00	-1.037	1.00	-0.206	-0.206	0.006
all ^1D[h]	23	959.94	0.31	-4.194	3.01	-2.297	-2.311	0.068
all ^1D + all backbone ^3J[i]	35	967.87	0.37	-4.870	4.94	-2.512	-2.581	0.158
all ^1D + all backbone ^3J (FM ensemble)[j]	34	1158.08	0.98	-5.738	4.94	-3.611	-3.685	0.035
all ^1D + all ^3J[k]	37	1024.19	0.20	-5.203	6.82	-2.633	-2.656	0.057

[a] Unless noted otherwise, all values correspond to a calculation of the 35000-frame MD data set.

[b] The entropy value of the equal population distribution for the 35000 conformations is $S(\boldsymbol{p}^{\text{equal}}) = \ln(35000) = 10.463$.

[c] Number of experimental constraints.

[d] The values correspond to the sum of individual χ^2 values for the different data types.

[e] The entropy difference is calculated as the deviation from $S(\boldsymbol{p}^{\text{equal}})$.

[f] The 35000 conformation data set was randomly divided into two mutually exclusive data sets, each containing 17500 conformations. The calculation was repeated on both data sets and entropy differences were calculated as the deviation from $S(\boldsymbol{p}^{\text{equal}}) = \ln(17500) = 9.770$. ΔS average (S.D.) corresponds to the average (standard deviation) of both entropy differences.

[g] The constraints consist of the backbone scalar couplings $^3J_{HNH\alpha}$ and $^3J_{H\alpha N}$.

[h] The constraints consist of the RDCs $^1D_{NH}$, $^1D_{C\alpha H\alpha}$, and $^1D_{C\alpha C'}$.

[i] The constraints consist of $^3J_{HNH\alpha}$, $^3J_{H\alpha N}$, $^1D_{NH}$, $^1D_{C\alpha H\alpha}$, and $^1D_{C\alpha C'}$.

[j] The calculation was carried out on a 35000-conformation ensemble generated by Flexible-Meccano using $^3J_{HNH\alpha}$, $^3J_{H\alpha N}$, $^1D_{NH}$, $^1D_{C\alpha H\alpha}$, and $^1D_{C\alpha C'}$ constraints. Due to the nature of the

Flexible-Meccano simulation the ϕ angle of the last residue is fixed and its $^{3}J_{HNH\alpha}$ constraint is not meaningful.

[k] The constraints consist of $^{3}J_{HNH\alpha}$, $^{3}J_{H\alpha N}$, $^{3}J_{NC\gamma}$, $^{3}J_{C'C\gamma}$, $^{1}D_{NH}$, $^{1}D_{C\alpha H\alpha}$, and $^{1}D_{C\alpha C'}$.

References

(1) Zwanzig, R.; Szabo, A.; Bagchi, B. Levinthal's Paradox. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 20–2.

(2) Tompa, P. Unstructural Biology Coming of Age. *Curr. Opin. Struct. Biol.* **2011**, *21*, 419–425.

(3) Frauenfelder, H.; Sligar, S.; Wolynes, P. The Energy Landscapes and Motions of Proteins. *Science* **1991**, *254*, 1598–1603.

(4) Boehr, D. D.; McElheny, D.; Dyson, H. J.; Wright, P. E. The Dynamic Energy Landscape of Dihydrofolate Reductase Catalysis. *Science* **2006**, *313*, 1638–1642.

(5) Manglik, A.; Kobilka, B. The Role of Protein Dynamics in GPCR Function: Insights From the β2AR and Rhodopsin. *Curr. Opin. Cell Biol.* **2014**, *27*, 136–143.

(6) Huang, J.-R.; Grzesiek, S. Ensemble Calculations of Unstructured Proteins Constrained by RDC and PRE Data: a Case Study of Urea-Denatured Ubiquitin. *J. Am. Chem. Soc.* **2010**, *132*, 694–705.

(7) Feldman, H. J.; Hogue, C. W. A Fast Method to Sample Real Protein Conformational Space. *Proteins* **2000**, *39*, 112–131.

(8) Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R.; Blackledge, M. A Structural Model for Unfolded Proteins From Residual Dipolar Couplings and Small-Angle X-Ray Scattering. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 17002–17007.

(9) Jha, A.; Colubri, A.; Freed, K.; Sosnick, T. R. Statistical Coil Model of the Unfolded State: Resolving the Reconciliation Problem. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13099–13104.

(10) Bertini, I.; Giachetti, A.; Luchinat, C.; Parigi, G.; Petoukhov, M. V.; Pierattelli, R.; Ravera, E.; Svergun, D. I. Conformational Space of Flexible Biological Macromolecules From Average Data. *J. Am. Chem. Soc.* **2010**, *132*, 13553–13558.

(11) Różycki, B.; Kim, Y. C.; Hummer, G. SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. *Structure* **2011**, *19*, 109–116.

(12) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(13) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12562–12566.

(14) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated Molecular Dynamics: a Promising and Efficient Simulation Method for Biomolecules. *J. Chem. Phys.* **2004**, *120*, 11919–11929.

(15) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Systematic Validation of Protein Force Fields Against Experimental Data. *PLoS ONE* **2012**, *7*, e32131.

(16) Lindorff-Larsen, K.; Trbovic, N.; Maragakis, P.; Piana, S.; Shaw, D. E. Structure and Dynamics of an Unfolded Protein Examined by Molecular Dynamics Simulation. *J. Am. Chem. Soc.* **2012**, *134*, 3787–3791.

(17) Chen, Y.; Campbell, S. L.; Dokholyan, N. V. Deciphering Protein Dynamics From NMR Data Using Explicit Structure Sampling and Selection. *Biophys. J.* **2007**, *93*, 2300–2306.

(18) Bernadó, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. Structural Characterization of Flexible Proteins Using Small-Angle X-Ray Scattering. *J. Am. Chem. Soc.* **2007**, *129*, 5656–5664.

(19) Nodet, G.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M. R.; Blackledge, M. Quantitative Description of Backbone Conformational Sampling of Unfolded Proteins at Amino Acid Resolution From NMR Residual Dipolar Couplings. *J. Am. Chem. Soc.* **2009**, *131*, 17908–17918.

(20) Berlin, K.; Castañeda, C. A.; Schneidman-Duhovny, D.; Sali, A.; Nava-Tudela, A.; Fushman, D. Recovering a Representative Conformational Ensemble From Underdetermined Macromolecular Structural Data. *J. Am. Chem. Soc.* **2013**, *135*, 16595–16609.
(21) Choy, W. Y.; Forman-Kay, J. D. Calculation of Ensembles of Structures Representing the Unfolded State of an SH3 Domain. *J. Mol. Biol.* **2001**, *308*, 1011–1032.
(22) Fisher, C. K.; Huang, A.; Stultz, C. M. Modeling Intrinsically Disordered Proteins with Bayesian Statistics. *J. Am. Chem. Soc.* **2010**, *132*, 14919–14927.
(23) Dames, S. A.; Aregger, R.; Vajpai, N.; Bernadó, P.; Blackledge, M.; Grzesiek, S. Residual Dipolar Couplings in Short Peptides Reveal Systematic Conformational Preferences of Individual Amino Acids. *J. Am. Chem. Soc.* **2006**, *128*, 13508–13514.
(24) Bignucolo, O.; Leung, H. T. A.; Grzesiek, S.; Bernèche, S. Backbone Hydration Determines the Folding Signature of Amino Acid Residues. *J. Am. Chem. Soc.* **2015**, *137*, 4300–4303.
(25) Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* **1948**, *27*, 379–423.
(26) Boyd, S.; Vandenberghe, L. Convex Optimization; Cambridge University Press, 2004.
(27) Karmarkar, N. A New Polynomial-Time Algorithm for Linear Programming. *Combinatorica* **1984**, *4*, 373–395.
(28) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics* **1951**, *22*, 79–86.
(29) Koenig, B. W.; Kontaxis, G.; Mitchell, D. C.; Louis, J. M.; Litman, B. J.; Bax, A. Structure and Orientation of a G Protein Fragment in the Receptor Bound State From Residual Dipolar Couplings. *J. Mol. Biol.* **2002**, *322*, 441–461.
(30) Meier, S.; Blackledge, M.; Grzesiek, S. Conformational Distributions of Unfolded Polypeptides From Novel NMR Techniques. *J. Chem. Phys.* **2008**, *128*, 052204.
(31) Vogeli, B.; Ying, J.; Grishaev, A.; Bax, A. Limits on Variations in Protein Backbone Dynamics From Precise Measurements of Scalar Couplings. *J. Am. Chem. Soc.* **2007**, *129*, 9377–9385.
(32) Lohr, F.; Schmidt, J. M.; Maurer, S.; Rüterjans, H. Improved Measurement of 3J(Hαi, Ni+1) Coupling Constants in H2O Dissolved Proteins. *J. Magn. Reson.* **2001**, *153*, 75–82.
(33) Hu, J.-S.; Grzesiek, S.; Bax, A. Two-Dimensional NMR Methods for Determining Ξ1 Angles of Aromatic Residues in Proteins From Three-Bond J C'Cγ and J NCγ Couplings. *J. Am. Chem. Soc.* **1997**, *119*, 1803–1804.
(34) Vajpai, N.; Gentner, M.; Huang, J.-R.; Blackledge, M.; Grzesiek, S. Side-Chain Chi(1) Conformations in Urea-Denatured Ubiquitin and Protein G From (3)J Coupling Constants and Residual Dipolar Couplings. *J. Am. Chem. Soc.* **2010**, *132*, 3196–3203.
(35) Best, R. B.; Hummer, G. Optimized Molecular Dynamics Force Fields Applied to the Helix-Coil Transition of Polypeptides. *J. Phys. Chem. B* **2009**, *113*, 9004–9015.
(36) Wächter, A.; Biegler, L. T. On the Implementation of an Interior-Point Filter Line-Search Algorithm for Large-Scale Nonlinear Programming. *Math. Programming* **2005**, *106*, 25–57.
(37) Zwanzig, R. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420–1426.
(38) Norgaard, A. B.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Experimental Parameterization of an Energy Function for the Simulation of Unfolded Proteins. *Biophys. J.* **2008**, *94*, 182–192.
(39) Li, D.-W.; Brüschweiler, R. NMR-Based Protein Potentials. *Angewandte Chemie International Edition* **2010**, *49*, 6778–6780.
(40) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D.

Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone Φ, Ψ and Side-Chain X(1) and X(2) Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257–3273.

(41) Svergun, D.; Barberato, C.; Koch, M. H. J. CRYSOL– a Program to Evaluate X-Ray Solution Scattering of Biological Macromolecules From Atomic Coordinates. *J. Appl. Crystallogr.* **1995**, *28*, 768–773.

(42) Kalinin, S.; Peulen, T.; Sindbert, S.; Rothwell, P. J.; Berger, S.; Restle, T.; Goody, R. S.; Gohlke, H.; Seidel, C. A. M. A Toolkit and Benchmark Study for FRET-Restrained High-Precision Structural Modeling. *Nat. Methods* **2012**, *9*, 1218–1225.

(43) Koenig, B. W.; Rogowski, M.; Louis, J. M. A Rapid Method to Attain Isotope Labeled Small Soluble Peptides for NMR Studies. *J. Biomol. NMR* **2003**, *26*, 193–202.

(44) Sass, H.; Musco, G.; Stahl, S.; Wingfield, P.; Grzesiek, S. Solution NMR of Proteins Within Polyacrylamide Gels: Diffusional Properties and Residual Alignment by Mechanical Stress or Embedding of Oriented Purple Membranes. *J. Biomol. NMR* **2000**, *18*, 303–309.

(45) Chou, J.; Gaemers, S.; Howder, B.; Louis, J.; Bax, A. A Simple Apparatus for Generating Stretched Polyacrylamide Gels, Yielding Uniform Alignment of Proteins and Detergent Micelles*. *J. Biomol. NMR* **2001**, *21*, 377–382.

(46) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. NMRPipe: a Multidimensional Spectral Processing System Based on UNIX Pipes. *J. Biomol. NMR* **1995**, *6*, 277–293.

(47) Archer, S. J.; Ikura, M.; Torchia, D. A.; Bax, A. An Alternative 3D NMR Technique for Correlating Backbone 15 N with Side Chain Hβ Resonances in Larger Proteins. *J. Magn. Reson.* **1991**, *95*, 636–641.

(48) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(49) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

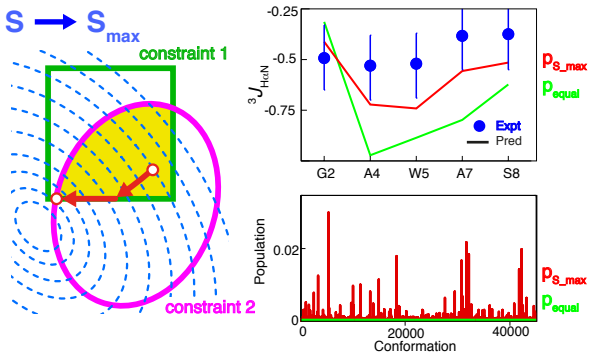
(50) Koradi, R.; Billeter, M.; Wuthrich, K. MOLMOL: a Program for Display and Analysis of Macromolecular Structures. *J. Mol. Graph.* **1996**, *14*, 51–55, 29–32.

(51) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(52) Lohr, F.; Schmidt, J. M.; Maurer, S.; Rüterjans, H. Improved Measurement of 3 J(H A I, N I+ 1) Coupling Constants in H 2 O Dissolved Proteins. *J. Magn. Reson.* **2001**, *153*, 75–82.

(53) Pérez, C.; Löhr, F.; Rüterjans, H.; Schmidt, J. M. Self-Consistent Karplus Parametrization of 3J Couplings Depending on the Polypeptide Side-Chain Torsion Chi1. *J. Am. Chem. Soc.* **2001**, *123*, 7081–7093.

Toc Figure


S → S max
constraint 1
constraint 2
-0.25
-0.5
-0.75
$^3J_{HaN}$
Expt
Pred
G2
A4
W5
A7
S8
p_{S_max}
p_{equal}
Population
0.02
0
20000
40000
Conformation

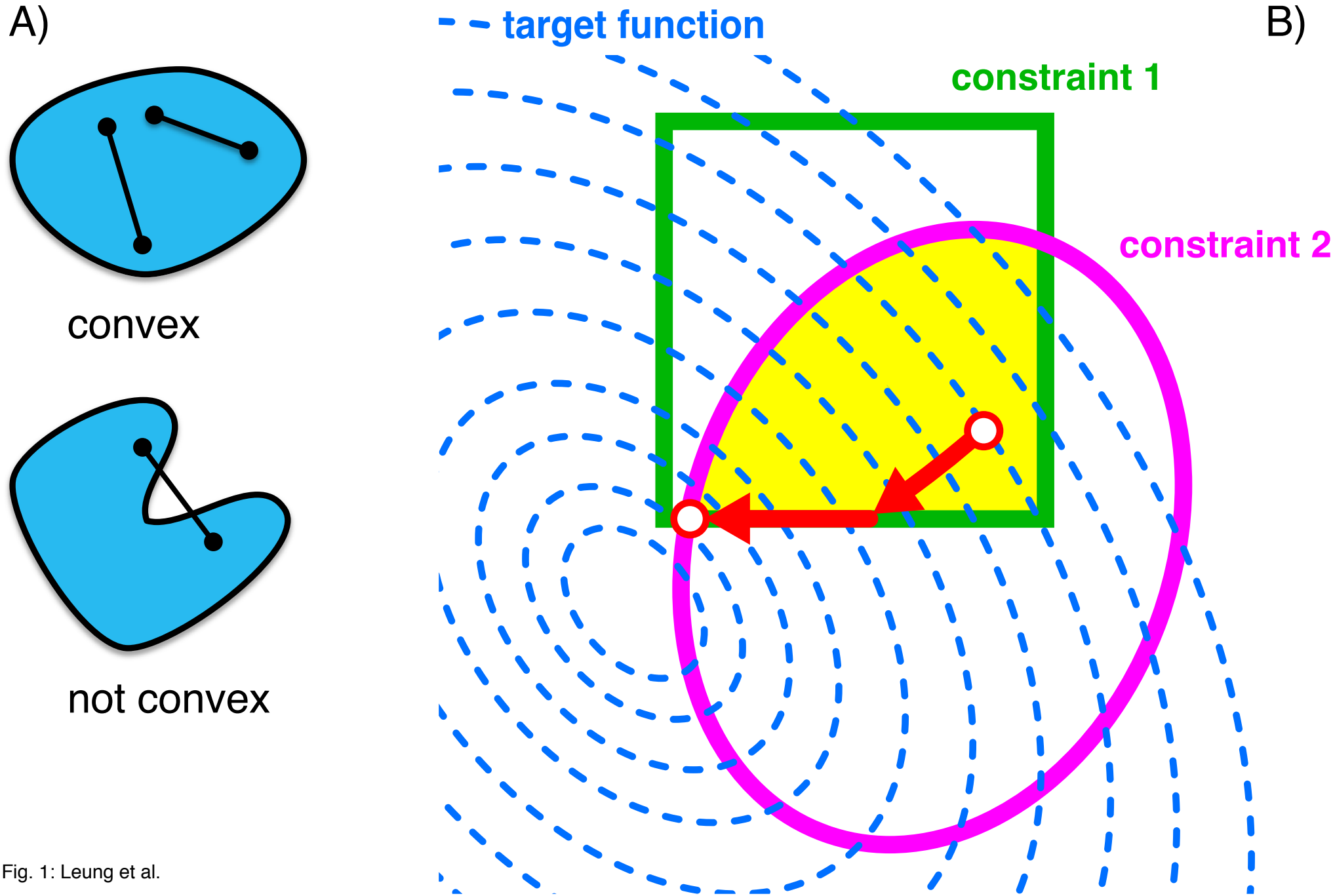


Fig. 1: Leung et al.

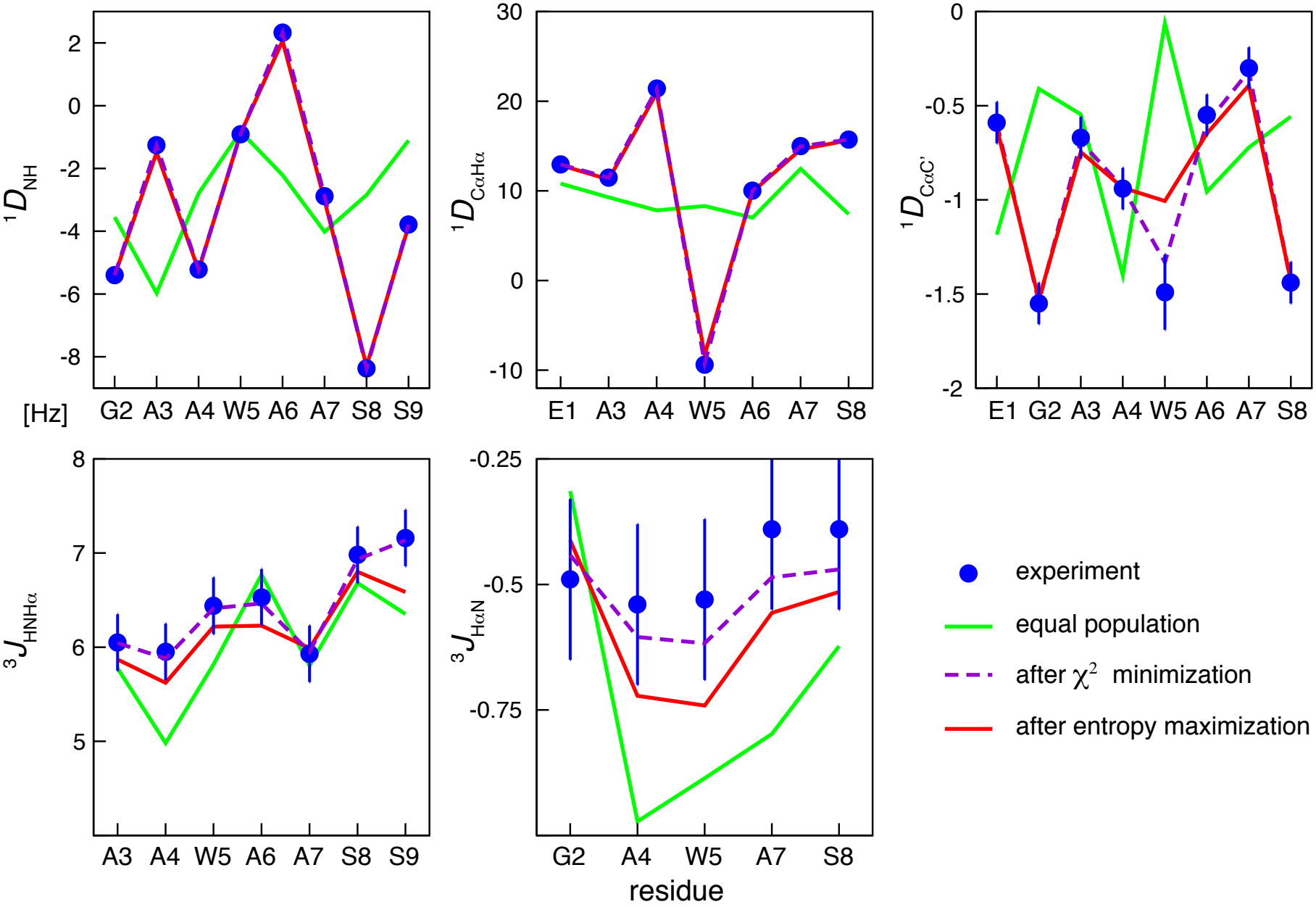


Fig. 2: Leung et al.

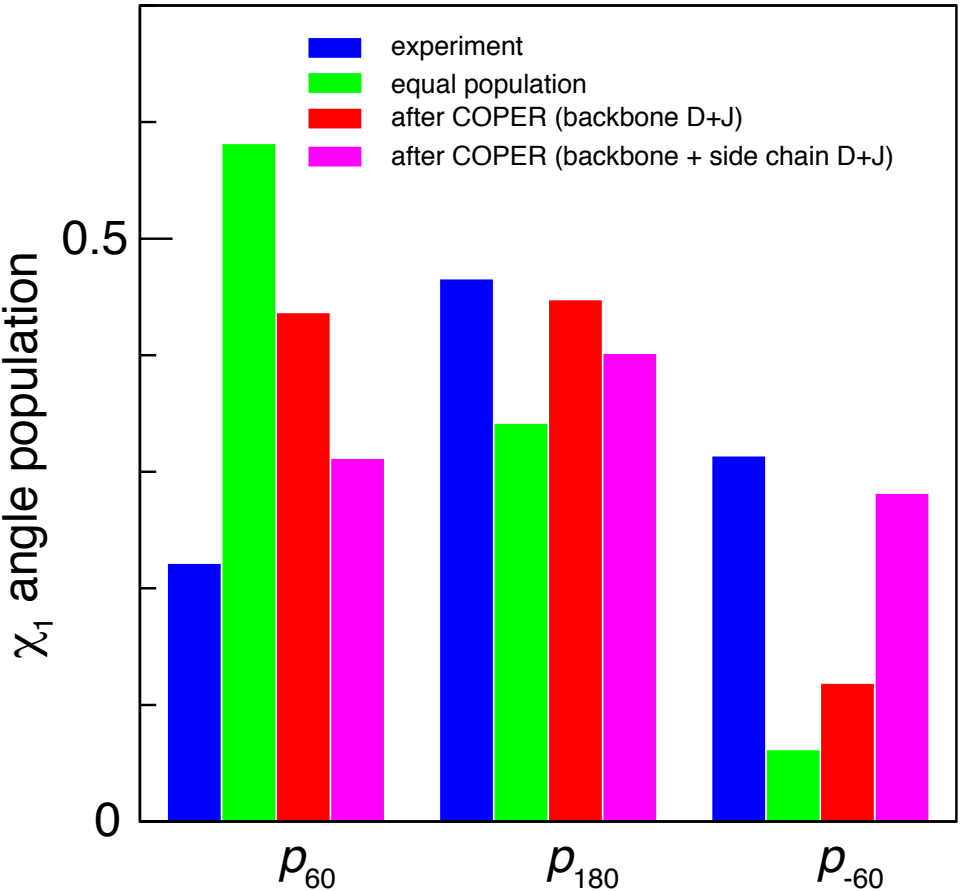


Fig. 3: Leung et al.

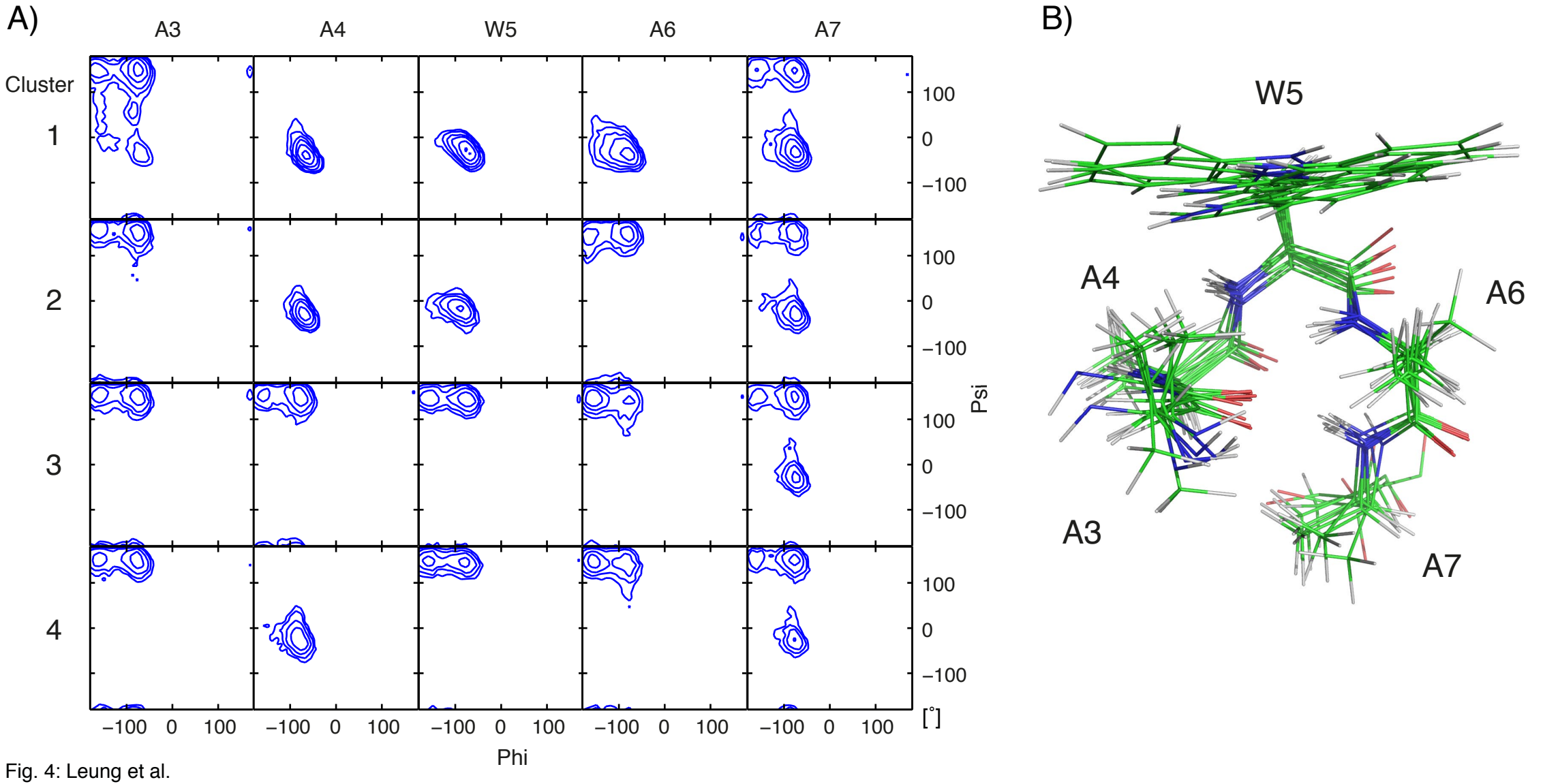


Fig. 4: Leung et al.

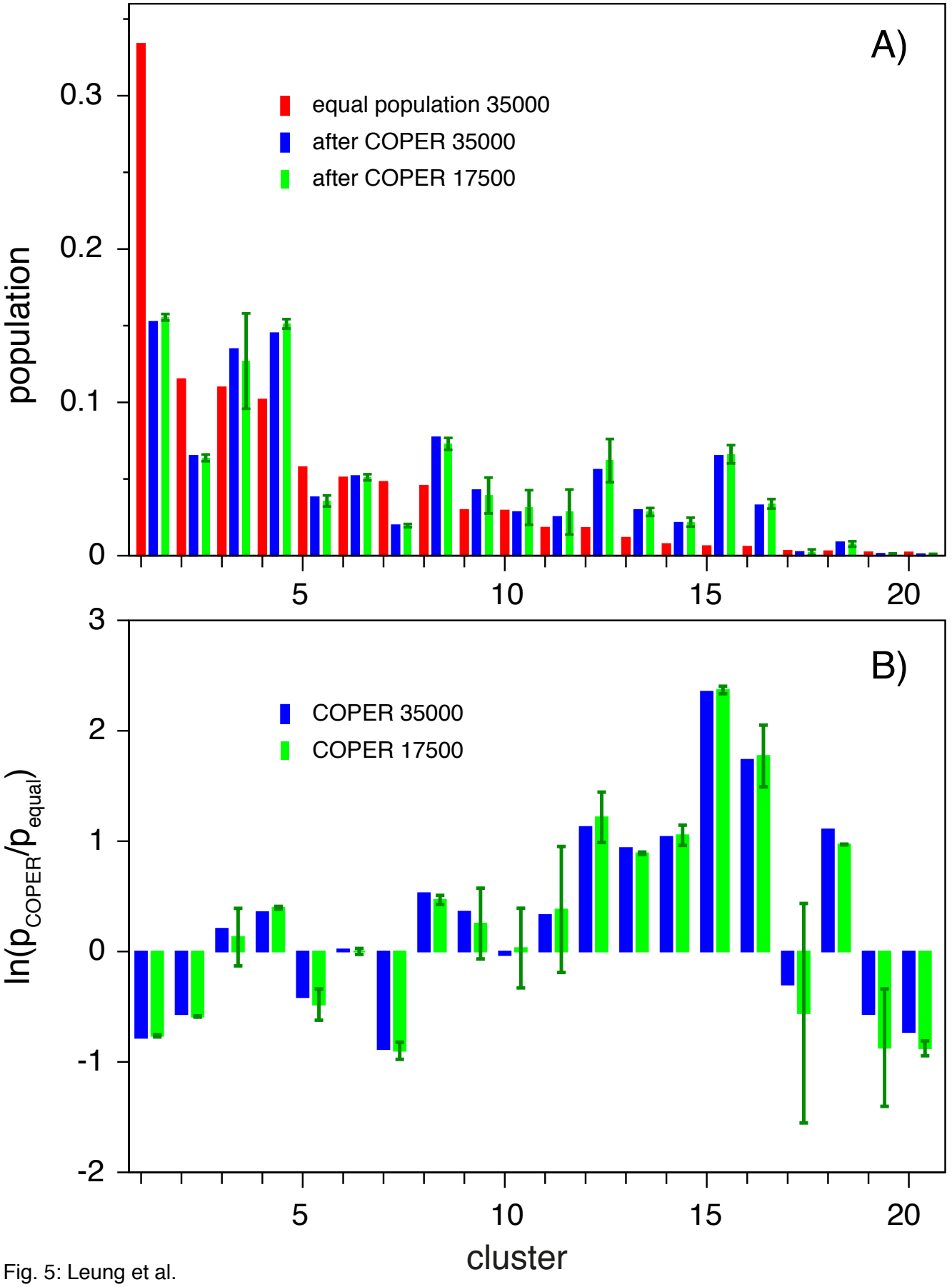


Fig. 5: Leung et al.

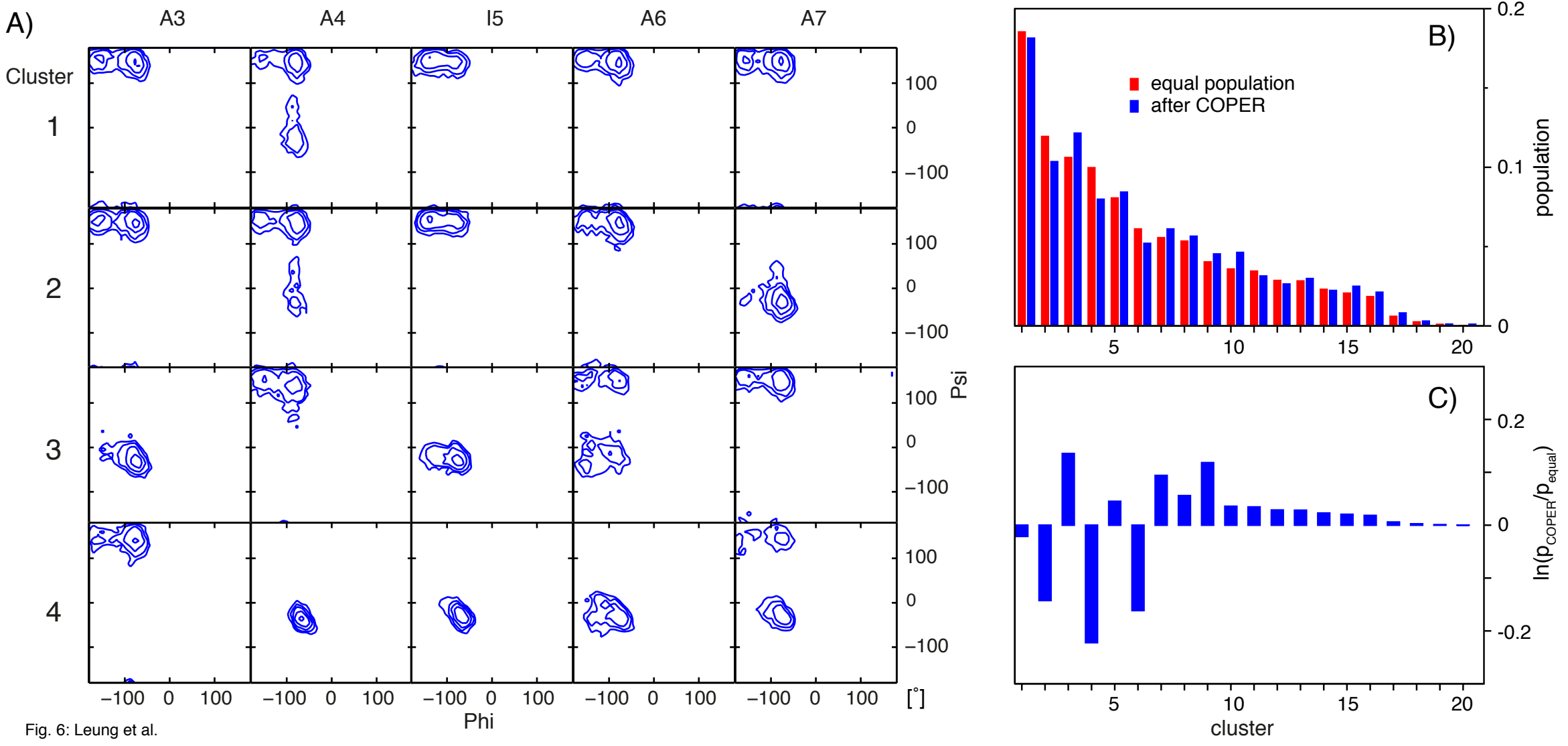


Fig. 6: Leung et al.

A rigorous and efficient method to reweight very large conformational ensembles using average experimental data and to determine their relative information content

Supporting Information

Hoi Tik Alvin Leung[1], Olivier Bignucolo[2], Regula Aregger[3], Sonja A. Dames[4], Adam Mazur[1], Simon Bernèche[2], and Stephan Grzesiek[1,]*

[1]Focal Area Structural Biology and Biophysics, Biozentrum, University of Basel, CH-4056 Basel, Switzerland
[2]SIB Swiss Institute of Bioinformatics, University of Basel, Klingelbergstrasse 50/70, CH-4056 Basel, Switzerland
[3]Institut für Biochemie, University of Leipzig, D-04103 Leipzig, Germany
[4]Department of Chemistry, Technische Universität München, 85748 Garching, Germany & Institute of Structural Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany

*Address correspondence to:
Stephan Grzesiek
Focal Area Structural Biology and Biophysics, Biozentrum
University of Basel, CH-4056 Basel, Switzerland
Phone: ++41 61 267 2100
FAX: ++41 61 267 2109
Email: Stephan.Grzesiek@unibas.ch

Table S1A: Backbone RDC ($^1D_{NH}$, $^1D_{C\alpha H\alpha}$, $^1D_{C\alpha C'}$) and *J*-coupling ($^3J_{HNH\alpha}$, $^3J_{H\alpha N}$) data (in Hz) of the isotope-labeled EGAAWAASS peptide

Residue	$^3J_{H\alpha N}$	Err.[a)]	$^3J_{H\alpha HN}$	Err.[b)]	$^1D_{NH}$	Err.	$^1D_{C\alpha H\alpha}$	Err.	$^1D_{C\alpha C'}$	Err.
E1							12.95	0.46	-0.59	0.11
G2	-0.49	0.16			-5.4	0.15			-1.55	0.11
A3			6.05	0.3	-1.26	0.15	11.5	0.46	-0.67	0.11
A4	-0.54	0.16	5.95	0.3	-5.22	0.15	21.42	0.46	-0.94	0.11
W5	-0.53	0.16	6.44	0.3	-0.91	0.15	-9.37	0.46	-1.49	0.2
A6			6.53	0.3	2.33	0.15	10.01	0.46	-0.55	0.11
A7	-0.39	0.16	5.93	0.3	-2.88	0.15	15.01	0.46	-0.3	0.11
S8	-0.39	0.16	6.98	0.3	-8.37	0.15	15.73	0.46	-1.44	0.11
S9			7.16	0.3	-3.78	0.15				

[a)]For comparison, the RMSD between experimental $^3J_{H\alpha N}$-couplings and values back-calculated from an x-ray structure was 0.13 Hz using Karplus parameters determined by Löhr et al.[1]

[b)]For comparison, the RMSD between the experimental $^3J_{HNH\alpha}$-couplings and values back-calculated from a structural ensemble is 0.36 Hz using Karplus parameters determined by Vögeli et al.[2]

Table S1B: Side chain $^3J_{NC\gamma}$ and $^3J_{C'C\gamma}$ scalar coupling constants (in Hz) of W5 in the EGAAWAASS peptide

	J	Err.
$^3J_{C'C\gamma}$	1.59	0.1
$^3J_{NC\gamma}$	1.21	0.1

Using these coupling constants and respective Karplus parameters,[3] the populations of the χ_1 +60˚, +180˚ and -60˚ rotamers are determined as 0.221, 0.464, and 0.312, respectively. Details of this calculation are given in Vajpai et al.[4]

Table S1C: Chemical shifts (in ppm) determined for the EGAAWAASS peptide[a)]

	H^N	N	H^α	C^α	C'	H^β	C^β
E1			4.103	55.83	173.15	2.152	29.99
G2	8.780	111.42	4.034[b]	45.12	173.46		
A3	8.353	124.31	4.285	52.35	177.72	1.277	19.31
A4	8.344	123.67	4.287	52.68	177.58	1.361	19.07
W5	8.008	119.98	4.612	57.37	175.80	3.308[b)]	29.50
A6	7.833	126.18	4.224	52.04	176.69	1.247	19.73
A7	8.055	123.48	4.241	52.49	177.78	1.429	19.37
S8	8.283	115.37	4.511	58.27	173.82	3.930[b]	64.13
S9	8.024	122.84		59.91	178.50		

[a)]Chemical shift assignments were derived from a set of standard HNCO, HNCA, CBCACONH, and HBHACONH experiments. ^{1}H, ^{15}N and ^{13}C chemical shifts are referenced

relative to the frequency of the ^{2}H lock resonance of water.
[b)]Methylene resonances overlap.

Figure S2

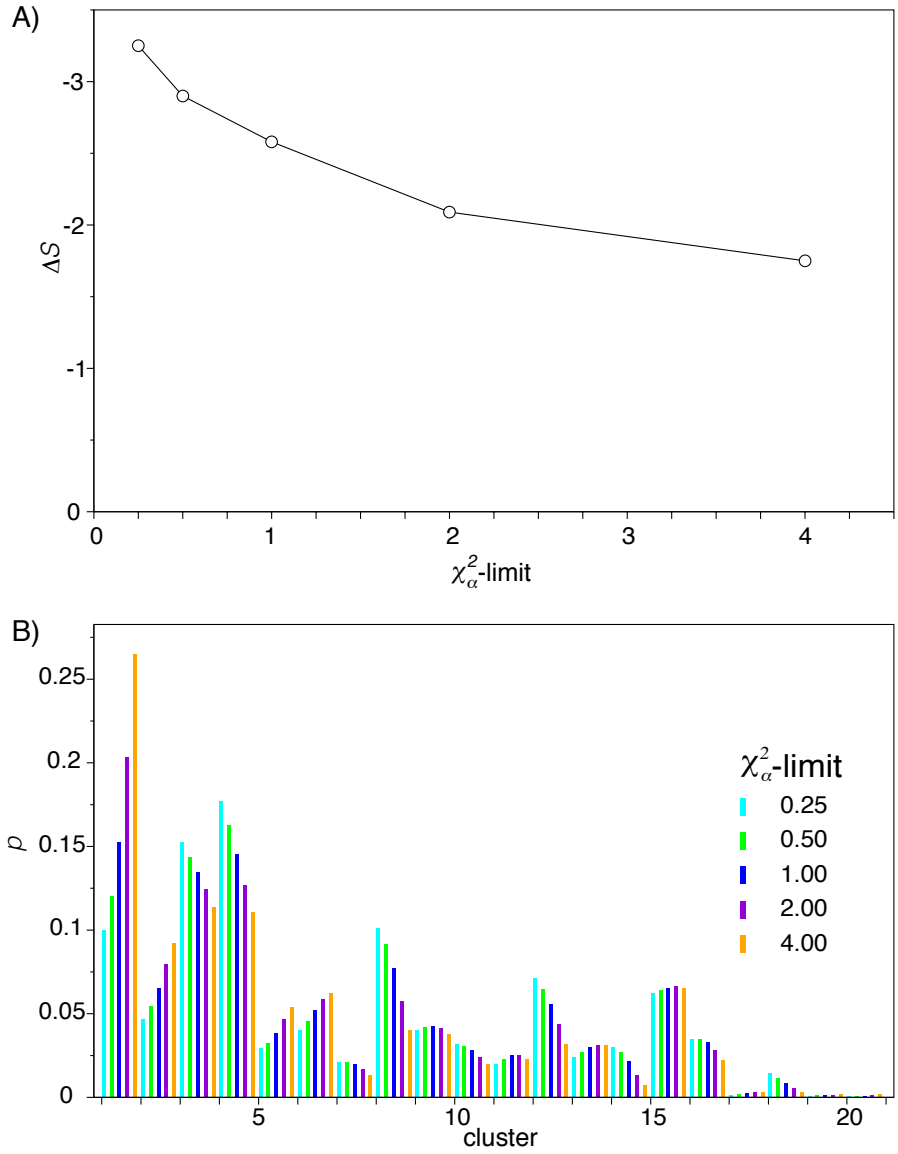


Figure S2: Analysis of the 35000 conformations from the MD conformations of the EGAAWAASS peptide with different χ^2_α-limits. A) The reduction in total entropy ΔS for χ^2_α-limits of the backbone RDC and *J*-coupling constraints ranging from 0.25 to 4.0. B) Populations of the 20 clusters calculated with χ^2_α-limits ranging from 0.25 to 4.0.

References

(1) Lohr, F.; Schmidt, J. M.; Maurer, S.; Rüterjans, H. *J. Magn. Reson.* **2001**, *153*, 75.
(2) Vogeli, B.; Ying, J.; Grishaev, A.; Bax, A. *J. Am. Chem. Soc.* **2007**, *129*, 9377.
(3) Pérez, C.; Löhr, F.; Rüterjans, H.; Schmidt, J. M. *J. Am. Chem. Soc.* **2001**, *123*, 7081.
(4) Vajpai, N.; Gentner, M.; Huang, J.-R.; Blackledge, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2010**, *132*, 3196.