Journal Name

RSCPublishing

ARTICLE

# SUPPLEMENTARY INFORMATION: Application of the Maximum Entropy Principle to determine ensembles of Intrinsically Disordred Proteins from RDC data

Cite this: DOI: 10.1039/x0xx00000x

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

M. Sanchez-Martinez,[a] and R. Crehuet[a]

## Technical details

The Pales software was called with the default options for steric alignment with explicit hydrogens (`-H`) as our structures contain hydrogens and we checked that the fit was better using them. The steric alignment was used because the experimental measured were taken in poly(ethylene glycol), where steric interactions dominate. The concentration of the liquid crystalline medium was kept as default, as it only scales linearly the magnitude of the alignment, which was taken into account by adjusting the scaling factor of the calculated RDCs versus the measured RDCs.

## Supplementary Figures

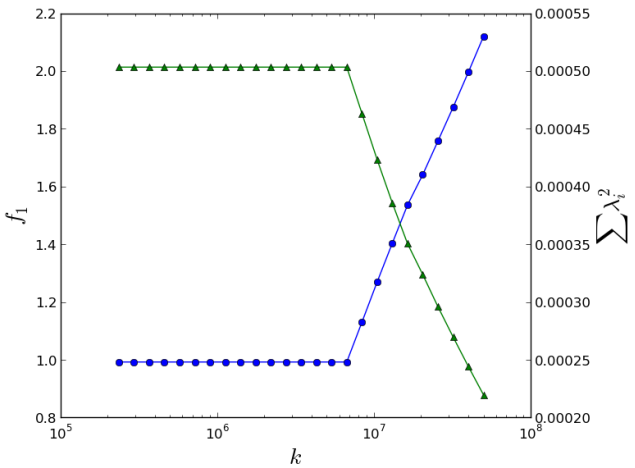


Fig. S1 Optimized $\boldsymbol{\lambda}$ (green) and $f_1$ values (blue) for different values of $k$. The optimization is run from high $k$ to low $k$, ensuring that the lambdas fall to the minimum closest to zero

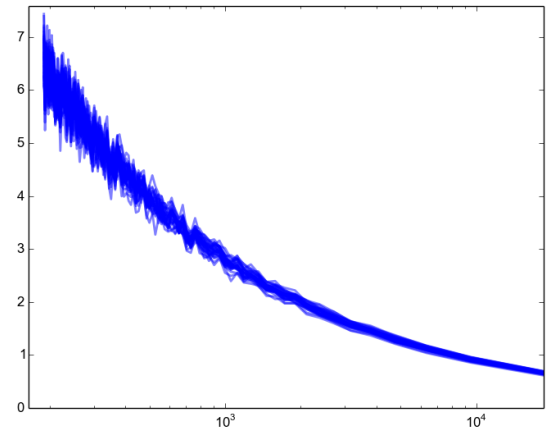


Fig. S2 Standard Error of the mean of all the N–H RDCs of Sendai virus at T01 with Profasi. The RDCs have been scaled to fit the experimental RDCs so that the error can be compared with the measured RDC. A constant scaling factor, from the fit with 20000 structures was used.

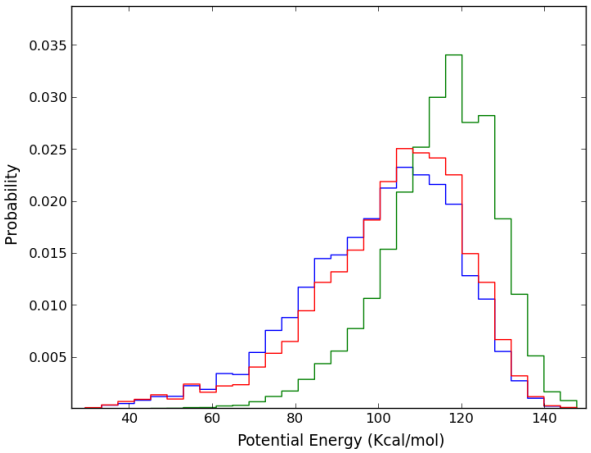


Fig. S4 Energy distribution histogram at the T0 (blue), T1 (green) and T0-reweighted (red) ensembles. The T0-reweighted ensemble is only slightly shifted towards the T1 ensemble, as would be expected from the small reweighting shown in fig S3.

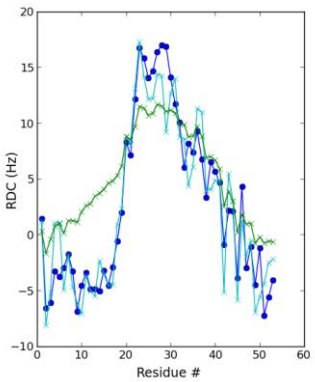


Fig. S3 N–H RDCs for the synthetic data. Blue circles are the objective RDCs, at temperature $T_1$.Green crosses are the RDCs of the ensemble to be fitted, at temperature $T_0$. Cyan crosses are the RDC of the $T_0$-ensemble re-fitted with the exact Boltzmann weight to fit the objective $T_1$-ensemble. This is the best possible fit that the $T_0$-ensemble can give. It shows that this ensemble can be reweighted to reproduce pretty accurately the RDCs at $T_1$. The RDCs also suggest that at $T_0$ there are too many long helices in the central region, compared to $T_1$, something to be expected also from the fact that $T_0$ is lower than $T_1$.

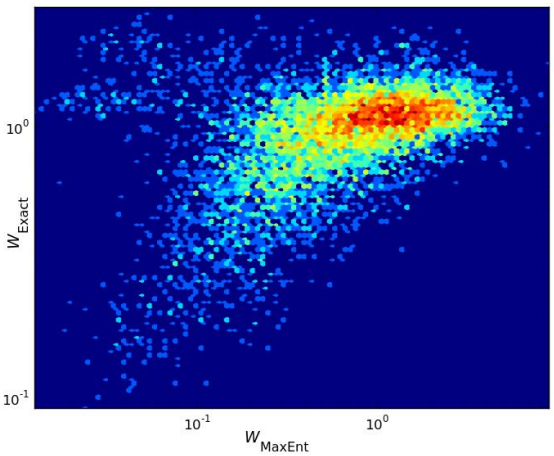


Fig. S5 Density plot (blue is low, red is high) between the exact Boltzmann weight of the from $T_0$ to $T_1$ (x-axis) and the weights from the ME fitting of the N–H RDCs (y axis). Although there is a certain correlation, most of the structures are not significantly reweighted by the N–H RDCs and have weights equal to 1.

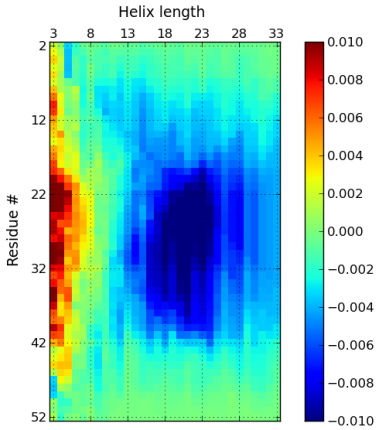


Fig. S6 SS-map difference between the initial ensemble at $T_0$ and the MaxEnt reweighted $T_0$-ensemble (to fit the $T_1$- N–H RDCs). As could be expected from the RDCs, the reweighted ensemble is depleted of long helices for residues 20 to 40, approx. Remark the different scale from Fig. 2.

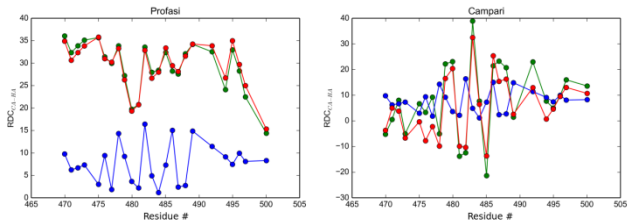


Fig. S7 Cross-validation of Cα-Hα RDCs predicted from the original (green) and N–H RDCs reweighted ensembles (red). Left: Profasi, right: Campari. The experimental Cα-Hα are plotted in blue. The same scaling factor than for the N–H was used. The root-mean-square error for the Campari reweight is slightly reduced (from 15.2 to 13.3) and for the Profasi ensemble, it remains essentially the same (22.5 and 22.5).

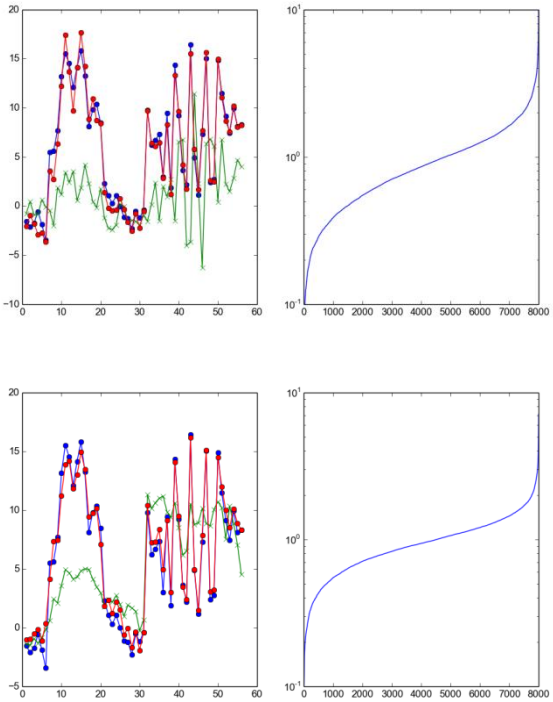

Fig. S8 MaxEnt fit of the N–H (from 1 to 31) Cα-Hα (from 21 to 56) RDCs (left) and corresponding distribution of weights(right). Campari ensemble (top) and Profasi ensemble (bottom). Because MaxEnt is fitting a larger set of data, the reweighting is higher. Compare with Figure 3 and 4.

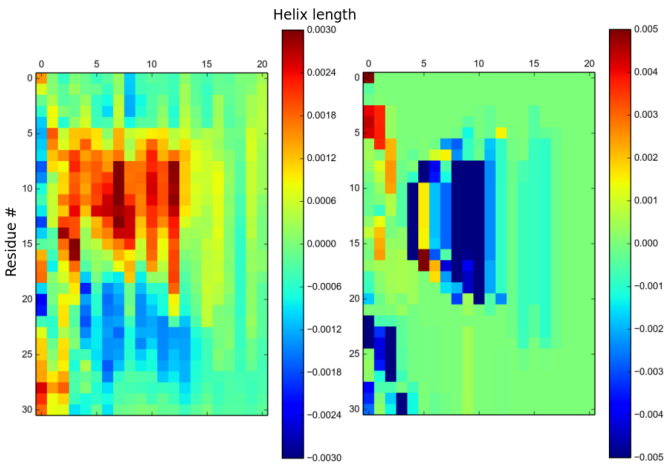


Fig. S9 SS-map difference between the initial ensemble and the MaxEnt reweighted ensemble (to fit the N–H and Ca-Ha experimental RDCs) for Profasi (left) and Campari (right) ensembles. As could be expected from the RDCs, the reweighted ensemble is depleted of long helices for residues 20 to 40, approx. Remark the different scale from Fig. 5

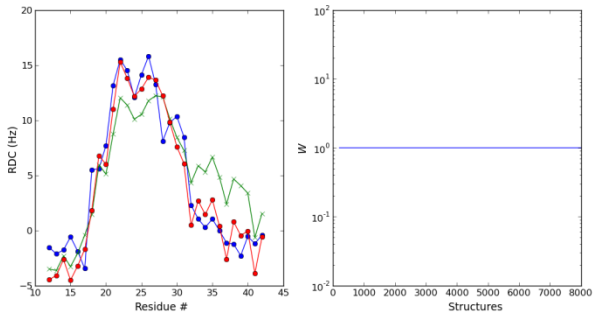


Fig. S10 Left: Fit of the Profasi ensemble to the experimental RDCs (blue) by using all the structures except the 208 structures that the MaxEnt optimization assigns weights lower than 0.75. Right: distribution of the weights (values for $w$=0 are not shown due to the logarithmic scale). ). The weights are kept unchanged, and so equal to 1, for 7792 structures. The remaining 208 structures have zero weight and so are effectively removed.

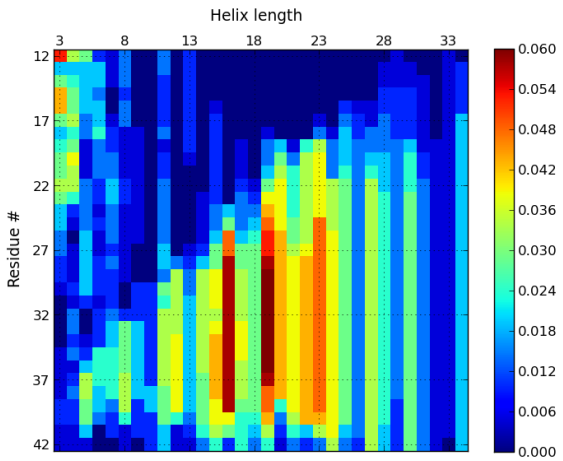


Fig. S11 SS-map of the 208 structures that have $w$<0.75 for the Profasi ensemble fitted to the experimental N–H RDC data.

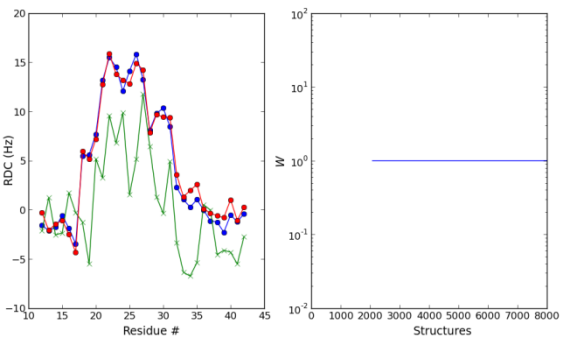


Fig. S12. Left: Fit of the Campari ensemble to the experimental N–H RDCs (blue) by using all the structures except the 2074 structures that the MaxEnt optimization assigns weights lower than 0.75. Right: distribution of the weights (values for $w$=0 are not shown due to the logarithmic scale). The weights are kept unchanged, and so equal to 1, for 5926 structures. The remaining 2074 structures have zero weight and so are effectively removed.

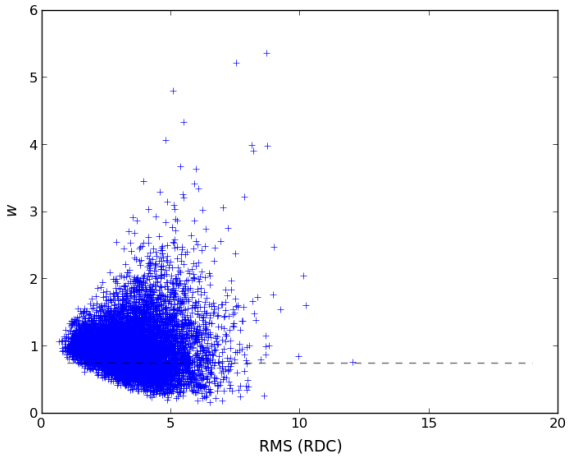


Fig. S13 Optimized weights for the Campari ensemble to fit the experimental N–H RDCs. The x-axis represents the root-mean-square of the RDCs for each of the 8000 structures, showing that the structures that get significantly reweighted are the ones that have large RDCs. The dotted lines is at $w$=0.75, and defines a small fraction of structures that, if removed, improve significantly the fit. See the text for more details.