Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

NMR Probing and Visualization of Correlated Structural Fluctuations in Intrinsically Disordered Proteins

Dennis Kurzbach, [*,a, b] Andreas Beier, [c] Agathe Vanas,[c] Andrea G. Flamm, [c] Gerald Platzer, [c] Thomas C. Schwarz[c] and Robert Konrat[*,c]

Supporting Information

Interpretation of Correlation and Covariance matrices of PRE and PRI data.

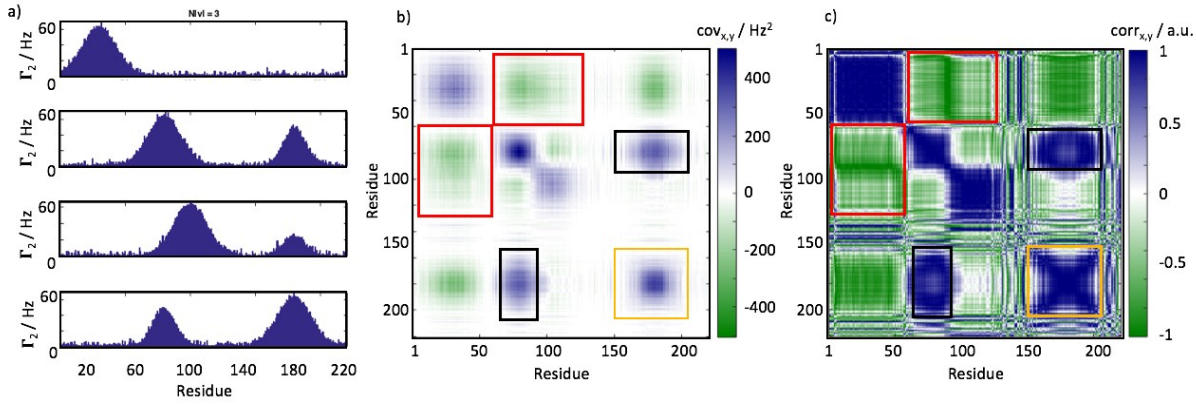

Figure S1. a) Simulated, hypothetic PRE profiles for 4 labelling sites: aa 30, 80, 100 and 180. Simulated noise was added to the profiles. b) Covariance matrix obtained from the data in a). c) Correlation coefficients obtained from the data in a).

Fig. S1a shows simulated PRE (or PRI) residue plots for a hypothetic, 220 amino acids (aa) long IDP with labelling sites at aa 30, 80, 100 and 180. Around each labelling site an area of increased PRE (or PRI) rates is observable (here referred to as short-range PRE). For labelling site 180 we introduced a long-range PRE remote to the labelling site around aa 80 (and vice versa). Fig. S1b shows the covariance matrix obtained from the four data sets in Fig. S1a according to eq. 2a in the main text. Fig S1c shows the corresponding correlation coefficients obtained according to eq. 2b of the main text, which corresponds to the normalized covariance between residues. In Fig. S2b and S2c the yellow square marks the region of the matrices that features positive covariance / correlation due to short-range PREs around the labelling site 180. We can expect such positive matrix entries around each labelling site as the residues close (in the primary sequence) to the spin label are intrinsically correlated via their spatial proximity and will therefore compulsorily experience an effect from the unpaired electron. The black squares indicate patches of positive covariance / correlation between labelling site 80 and 180. These positive correlations can be traced back to the long-range PRE observed around aa 80 for labelling site 180. Via these long-range PREs the label at aa 180 is correlated with the label at aa 80 (and vice versa). As indicated by the red squares for labelling sites 30, 80 and 100, one only observes negative covariance / anti-correlation between different labelling sites, if there are no long-range PREs connecting two labelling sites.
Generally, short-range PREs do not appear correlated in the covariance / correlation matrices also if the two regions spanned by short-range PREs overlap as in the case of the labels attached to aa 80 and 100.
Note that the correlation coefficients shown in Fig. S1c tend to values close to-1 or 1 (while zero values become depleted) due to the normalization to the standard deviations. Such, the correlation coefficients show clear positive or negative correlation also for weak PREs, while covariance matrix elements depend stronger on the intensity of the PRE.
The correlation analysis, however, is also more prone to correlation of noise in the input data,[1] especially for small standard deviations, a fact that must be carefully considered when performing such an analysis. Yet, correlation coefficients have the important advantage that a quantitative comparison with other data (e. g., between PRE and PRI) is simplified and that the comparison can be performed also in cases of varying number of input data sets (denoted as N in eq. 2 in the main text).

The exactly same reasoning presented here for PRE data is also valid for PRI data, since the underlying residue plots do not differ in shape, but only in units.

Influence of Noise on the Correlation Analysis

To account for the influence of noise on the correlation coefficient we performed a correlation analysis in dependence of the noise levels of the input data. In Fig. S2 we show the simulated PRE data for the four labelling sites at aa 30, 80, 100 and 180 with varying signal-to-noise ratios (SNR) between 6 and 2. (The SNR in Fig S1a is 10.) Fig. S3a-S3c shows the resulting correlation coefficients. Although the matrices naturally become noisier, the overall distribution of positive and negative correlations remains clearly discernible. The average signal-to-noise ratio (SNR) of a correlation matrix is higher than that of the input data sets as the four data sets are combined into one representation. Such that $\langle n \rangle \propto N^{1/2}$, where N is the number of input data sets (i.e., independent residue plots). This dependence is graphically depicted in Fig. S3d. Note that the average noise level, $\langle n \rangle$, of a correlation matrix, **Corr**, with elements $corr_{x,y}$ is defined as the variance of all matrix elements, i.e., $\langle n \rangle = Var(\mathbf{Corr})$. A detailed treatment of the noise in covariance analyses of NMR data can be found in the Supporting Material of reference 2. The same principles apply here.

Note that due to the normalization by the standard deviation the average noise level of a correlation matrix is independent of the size of the matrix and the absolute rates of the underlying PRE or PRI.

For the analyses presented in this work we find for all cases $N > 3$, such that $\langle n \rangle < 0.5$ holds. This dependence is highlighted by the dashed line in Fig. S3d. We therefore cut the matrices in the main text and set all values for which $-0.5 < corr_{x,y} < 0.5$ to zero in order to avoid confusion through correlated noise.

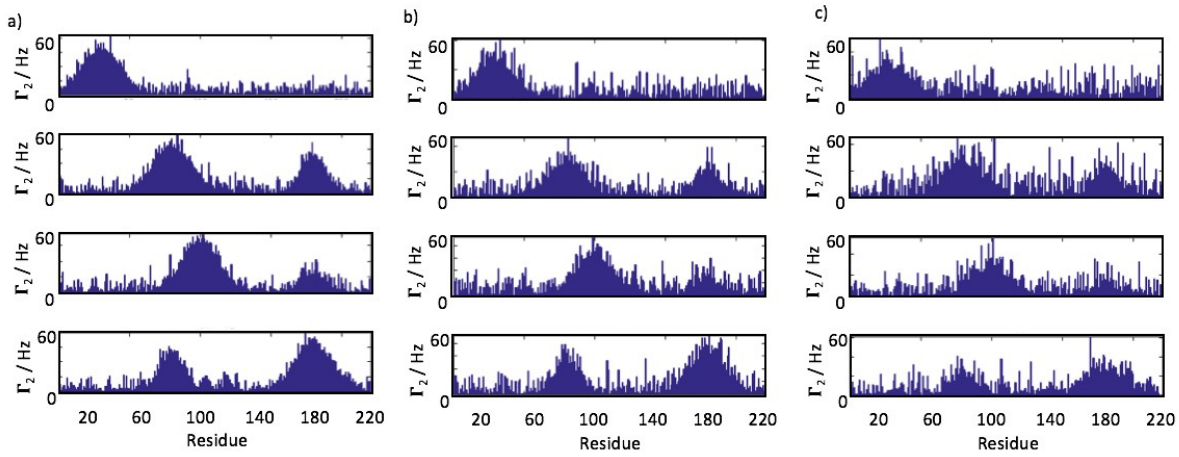

Figure S2. a) Simulated PRE data for labelling sites at aa 30, 80, 100 and 180 with an average SNR of 6. b) As in a), but with an average SNR of 3. c) As in a), but with an average SNR of 2.

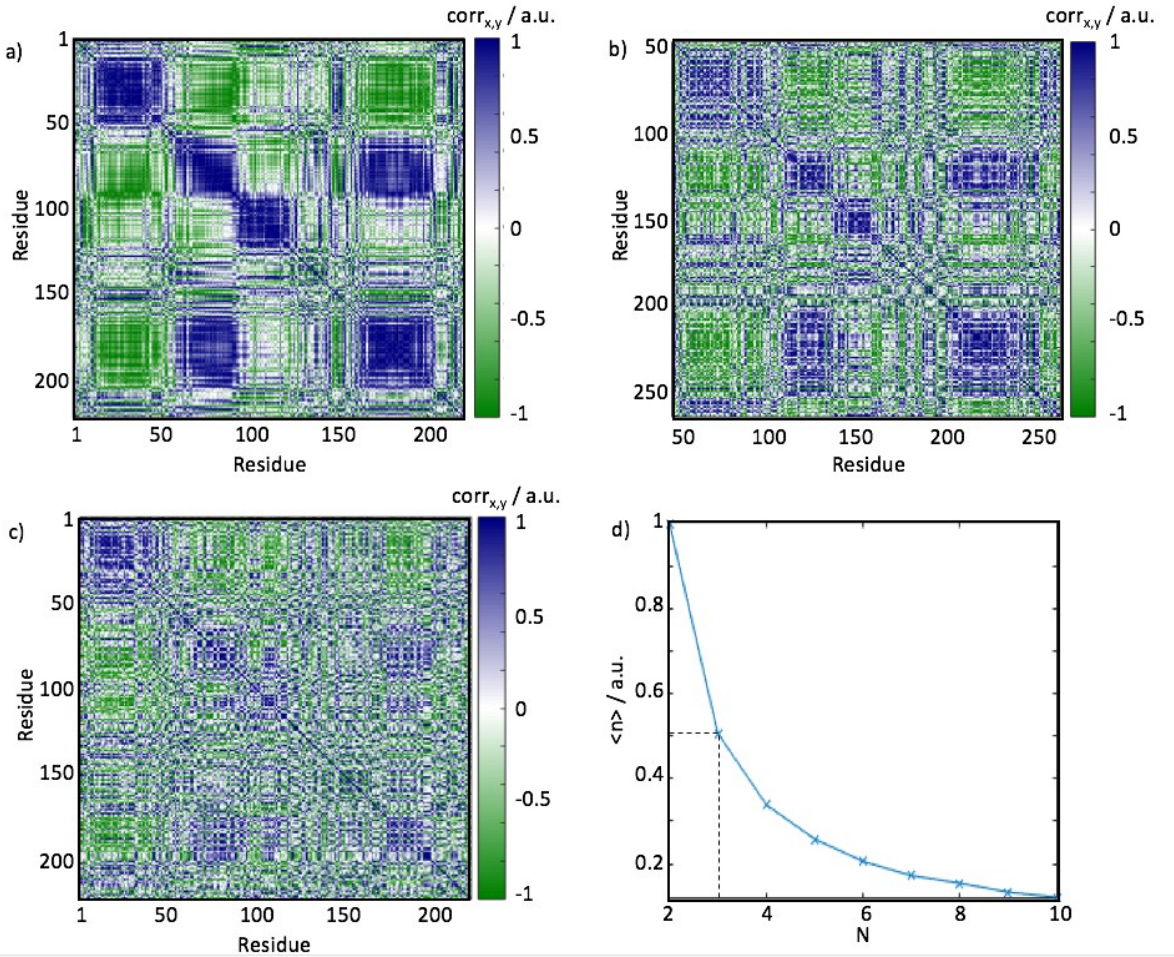

Figure S3. a) Correlation coefficients obtained from the simulated data shown in Figure S2a. b) Correlation coefficients obtained from the simulated data shown in Figure S2b. c) Correlation coefficients obtained from the simulated data shown in Figure S2c. d) The average noise level, <n>, of a correlation matrix vs. the number of input data sets. The dashed line indicates the average noise level for N =3.

Robustness of the Correlation Analysis

Fig. S4 shows PRI rates for OPN and BASP1[3] that underlie the here presented correlation analysis. In Fig S5a-S5c it is shown how the resulting correlation matrices appear if not all data sets are used for the calculation of the correlation coefficients. Fig S5a the correlation coefficients are shown when all data are used as in the main text. However, here we do not digitize the matrix as in the main text, but retain all entries. For BASP1 we observed non-zero correlation in the C-terminal region because of correlated noise; one can see from Fig. S4b that there is no PRI detectible between aa 100 and 200, yet we observe large correlation coefficients as the corresponding standard deviations used for data normalization become small (cf. eq. 2 in the main text). This clearly indicates that the digitization (as in the main text) of the correlation coefficients according to the average noise level of the matrices is necessary to avoid false correlations.

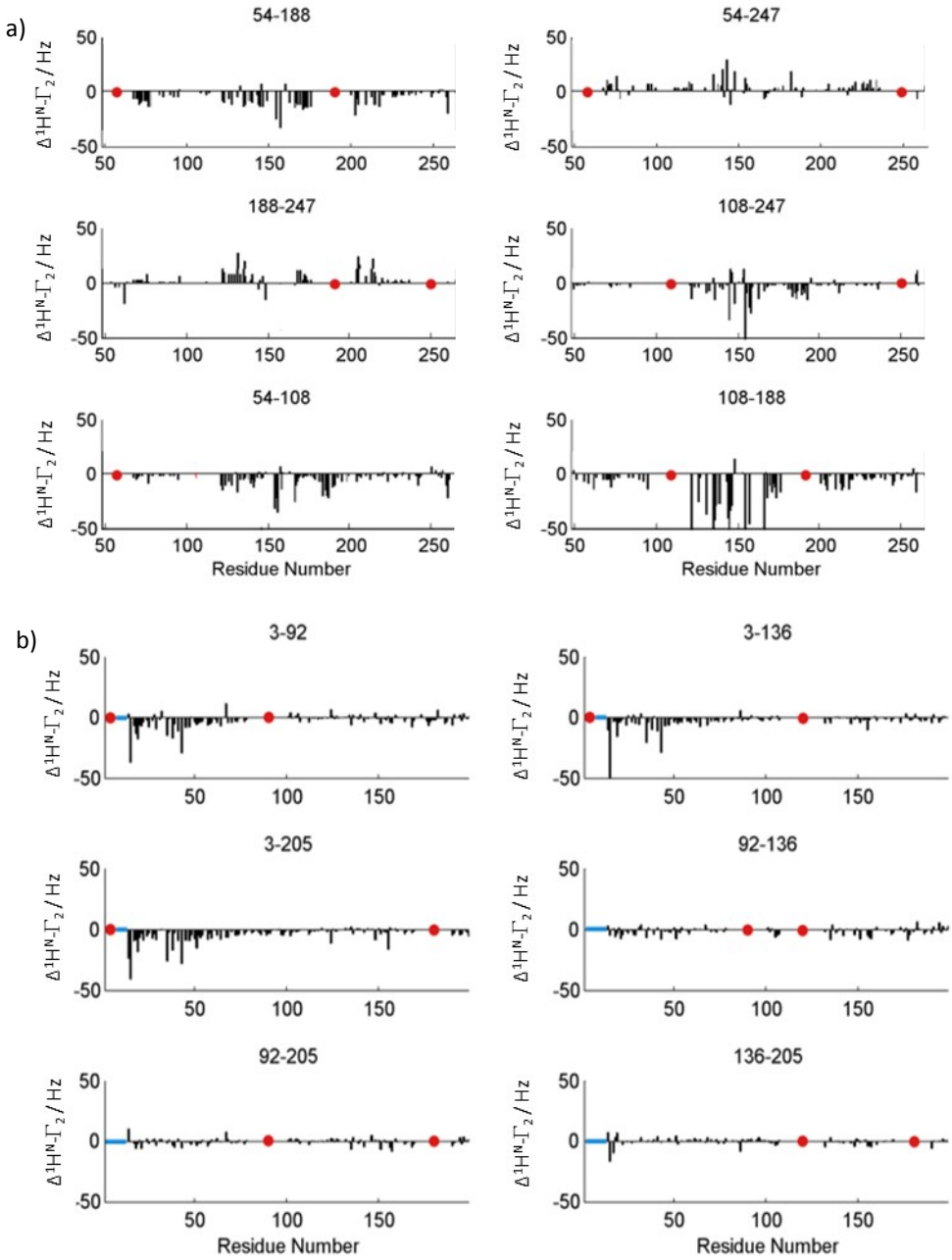

Figure S4. a) PRI rates for six pairs of labelling sites in OPN (see reference 2). b) a) PRI rates for six pairs of labelling sites in BASP1 (see reference 2). The labelling sites are indicated as red dots. Blue marked residues were excluded from the analysis, due to possible intermolecular biases.

In Fig. S5b and S5c the PRI-derived correlation coefficients for N = 5 and N = 4 are shown for OPN and BASP1. One can clearly observe that the overall shape of the matrices is retained even if a significant share of the underlying data is not used for their calculation. This indicates the robustness of the correlation analysis for PRI data since the PRI effect is based on a cooperative / collective phenomenon of concerted fluctuation along the protein chain, such that the information contained in the PRI values is present in each residue plot shown in Fig. S4.

This is, yet, not the case for PRE data, for which every set residue-resolved data obtained from a different spin label unrelatedly reflects compaction of the sub domains of a protein. These individual phenomena are, yet, removed from PRI rates as the PRI is defined as the deviation from independent effects of individual PREs, i.e., $\Delta\Gamma_2 = {}^1H^N\text{-}\Gamma_2(D) - [{}^1H^N\text{-}\Gamma_2(S_1) + {}^1H^N\text{-}\Gamma_2(S_2)]$, as explained in the main text.

In Fig. S6 the PRE data for the four labelling sites of OPN and BASP1 are shown and in Fig. S7 the resulting correlation matrices for N = 4 and N = 3. It is clearly observable that the shape of the matrices changes when one data set is left out for their calculation. This exemplifies the dependence of the PRE correlation coefficients on the input data; each data set contains individual information that is missing in the final matrix, if left out in the calculation of the resulting correlation coefficients, while PRI data is based on cooperative phenomena that relate all detected PRI rates.

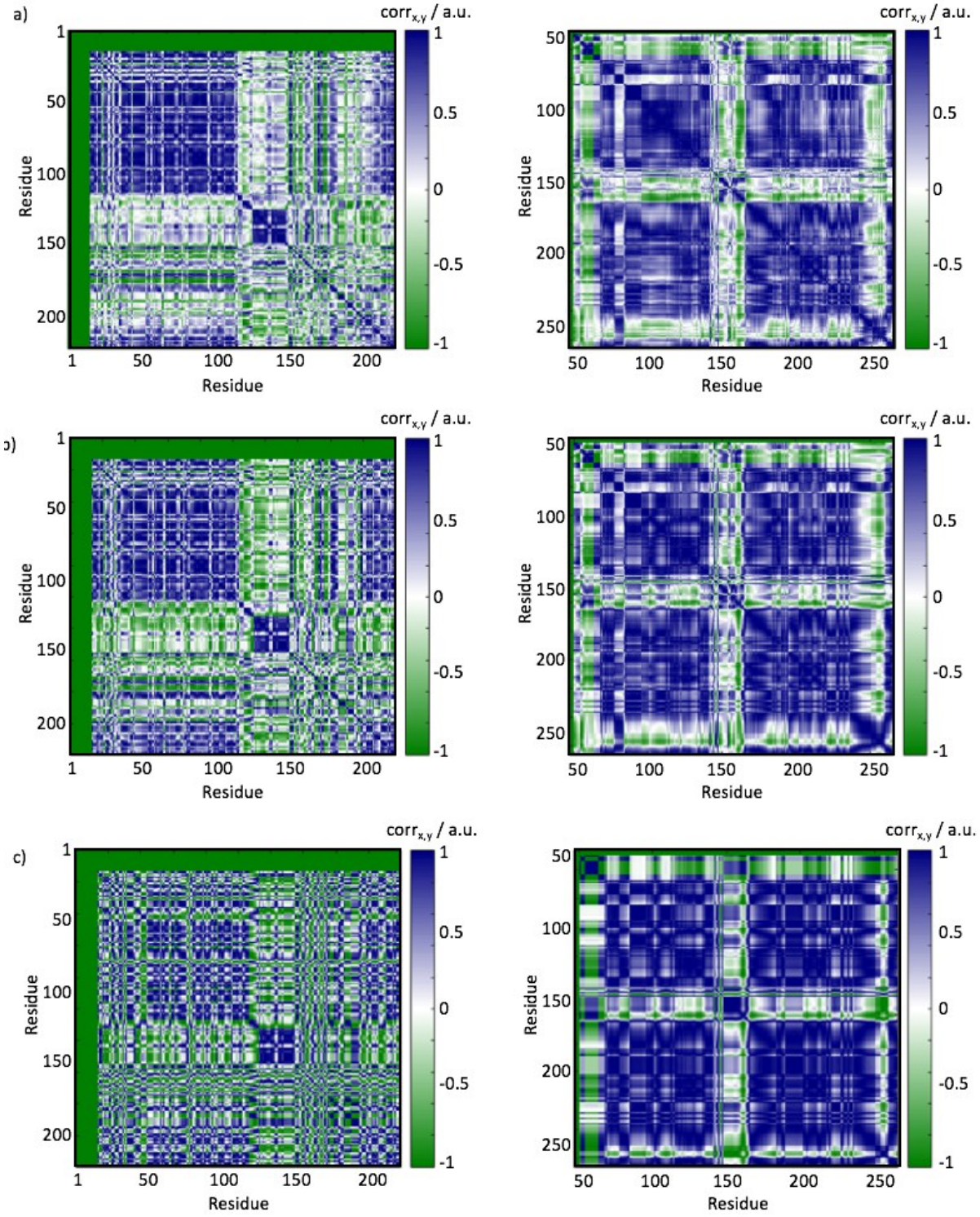

Figure S5. a) Correlation coefficients obtained from the data in Figure S4 for BASP1 (left) and OPN (right) utilizing all six (N = 6) available data sets (i.e., pairs of labelling sites). b) The same as in a), but utilizing only 5 data sets (N = 5). c) b) The same as in a), but utilizing only 4 data sets (N = 4).

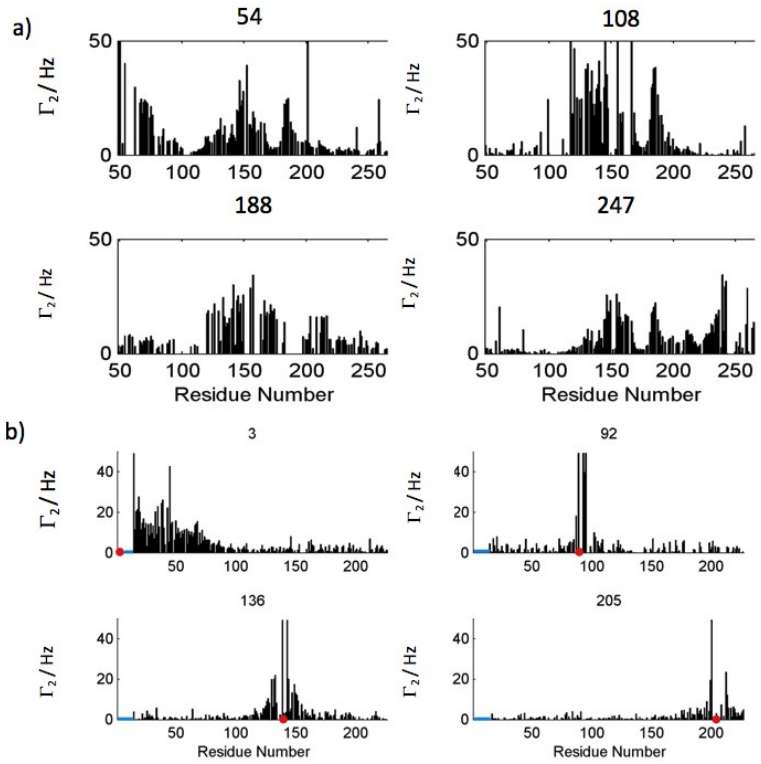

Figure S6. a) PRE residue plots for the four OPN single mutants as indicated on top of each panel reproduced from reference 1.b) PRE residue plots for the four BASP single mutants as indicated on top of each panel reproduced from reference 1.

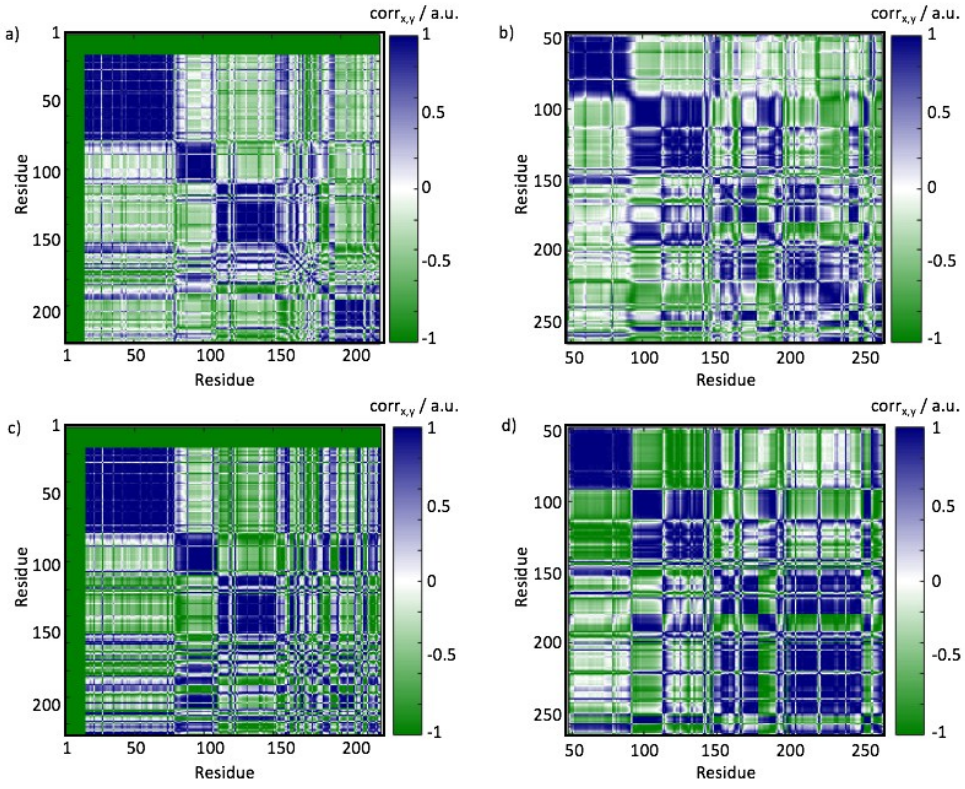

Figure S7. a) Correlation coefficients for BASP1 and b) OPN obtained from the PRE data shown in Figure S6 utilizing all four data sets (N = 4) in both cases. c) Correlation coefficients for BASP1 and d) OPN obtained from the PRE data shown in Figure S6 utilizing three data sets (N = 3) in both cases.

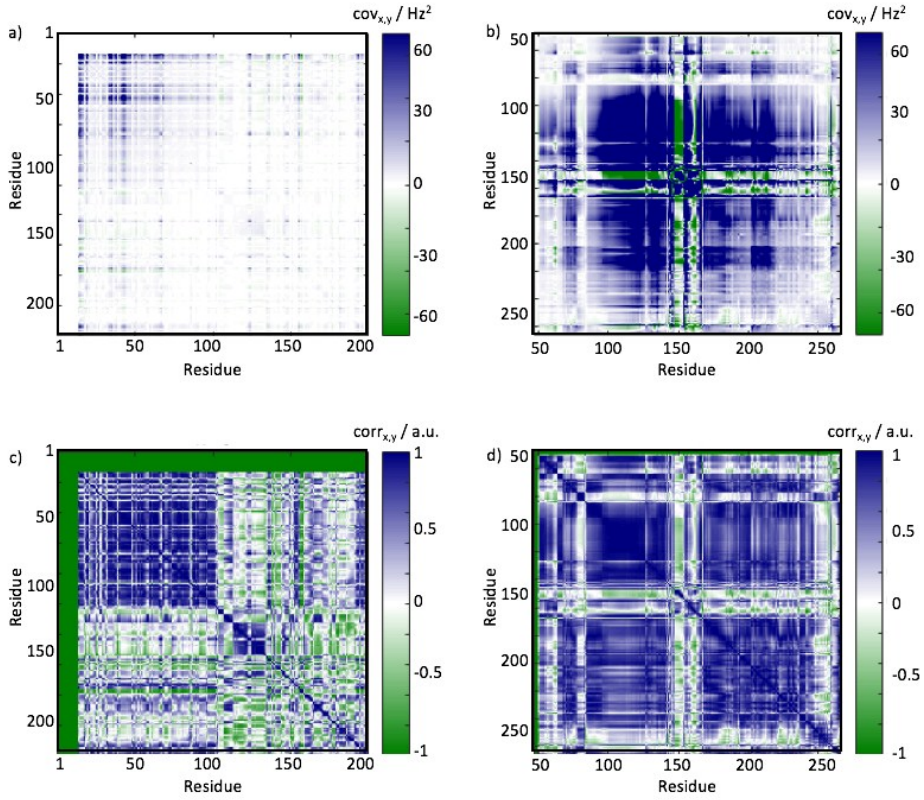

Figure S8. a) Covariance matrices for BASP1 and b) OPN obtained from the PRI data shown in Figure S4 utilizing all six data sets (N = 6) in both cases. c) Correlation coefficients for BASP1 and f) OPN obtained from the PRI data shown in Figure S4 utilizing all six data sets (N = 6) in both cases.

Comparison between Covariance and Correlation Matrices of Experimental Data

In Fig. S8a and S8c covariance matrices of the PRI data shown in Fig S4 are shown. In Fig. S8b and S8d the corresponding correlation coefficients are shown for BASP1 and OPN. Covariance and correlation matrices have a similar form even though correlation coefficients are prone to correlate noise as explained above. The information content prevalent in the covariance and correlation matrices is, thus, comparable, if one carefully considers the influence of data normalization on the average matrix noise level.

Details on Cluster Analyses

Cluster Analysis for Fig. 2a and 3a were performed as proposed in references [2] and [4]. A Ward-type inner square distance hierarchical clustering algorithm was used employing a Euclidean distance norm. Other clustering algorithms led to comparable results. The cluster analyses were performed on the non-truncated data set.

References

1. Everitt, B. S.; Landau, S.; Leese, M.; Stahl, D., *Cluster Analysis*. Wiley: 2011; Vol. 5.
2. Kurzbach, D., *Prot. Sci.* **2016,** *25*, 1617–1627.
3. Kurzbach, D.; Vanas, A.; Flamm, A. G.; Tarnoczi, N.; Kontaxis, G.; Maltar-Strmecki, N.; Widder, K.; Hinderberger, D.; Konrat, R., *Phys Chem Chem Phys* **2016,** *18* (8), 5753-8.
4. Selvaratnam, R.; Chowdhury, S.; VanSchouwen, B.; Melacini, G., *Proc Natl Acad Sci* **2011,** *108* (15), 6133-8.