Supplementary Materials

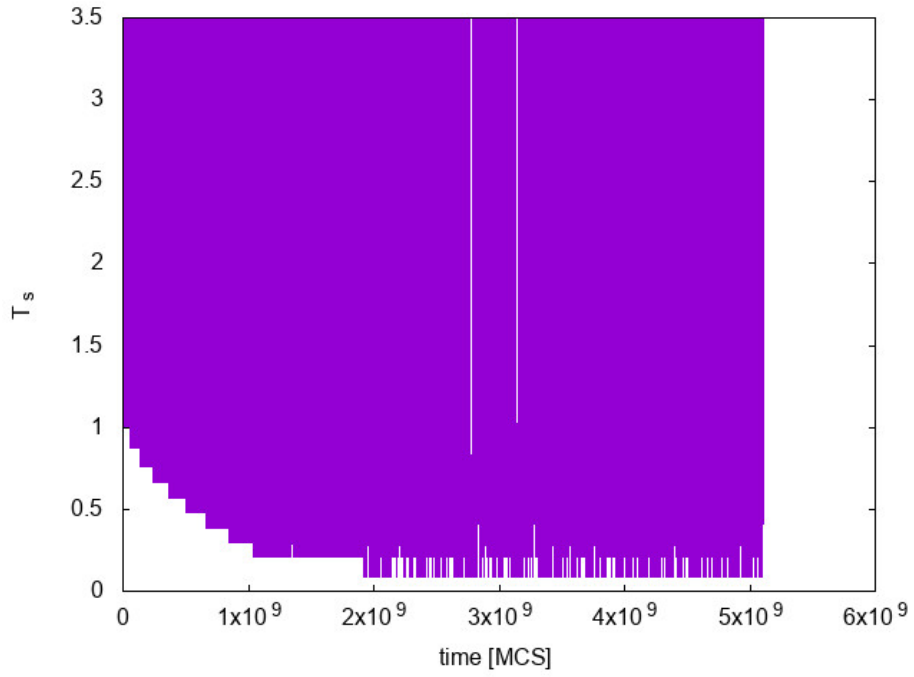


Figure S1: the fluctuations in the temperature T_s during a sampling of sequence space with the adaptive simulated tempering of 1BPI.

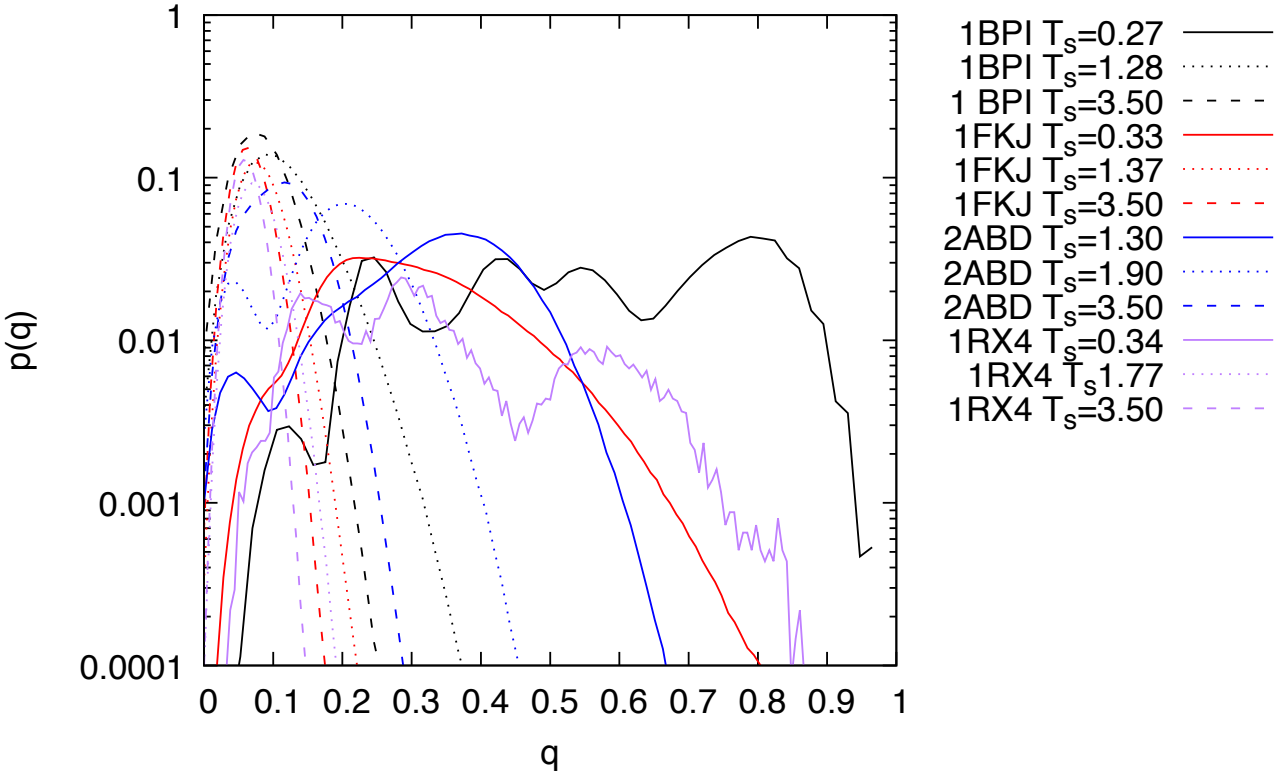


Figure S2: The distribution of the Hamming distance q between pairs of sequences in the alignments generated at different temperatures T_s for the various proteins. In the high-temperature and low-temperature phases there is a single peak at low values of q. In the frozen phase, q displays a complex distribution which suggests a hierarchical clustering in sequence space, similarly to the behaviour of spin glasses at low temperature.

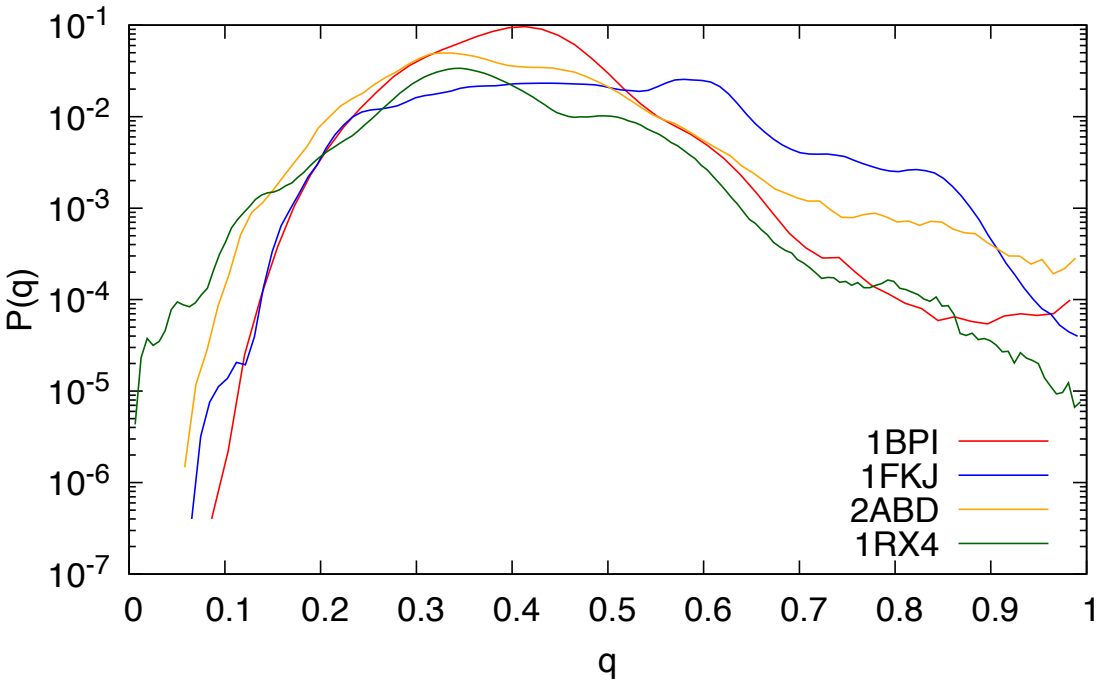
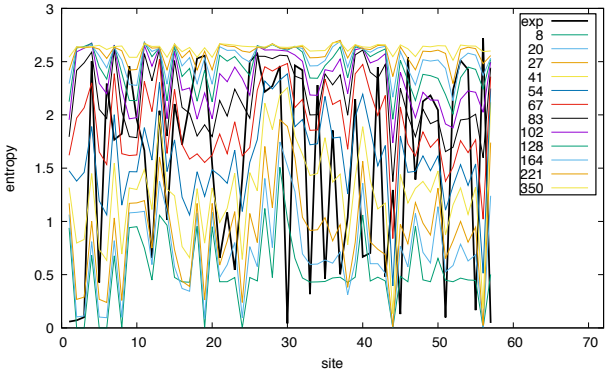
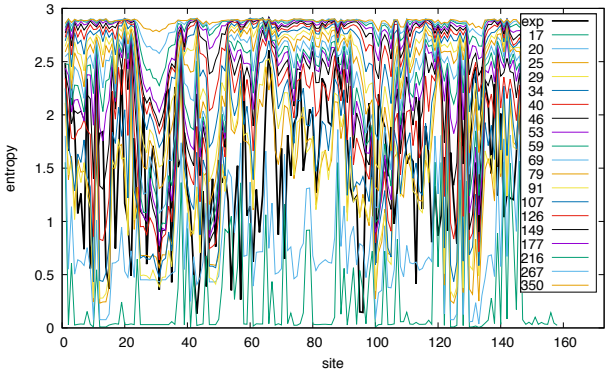


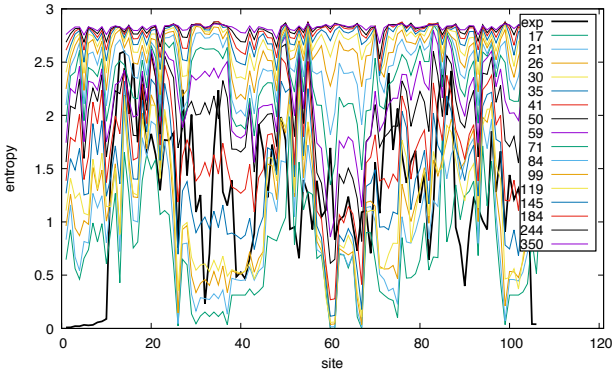
Figure S3: The distribution of Hamming distances for the actual alignments of the four proteins.

1BPI

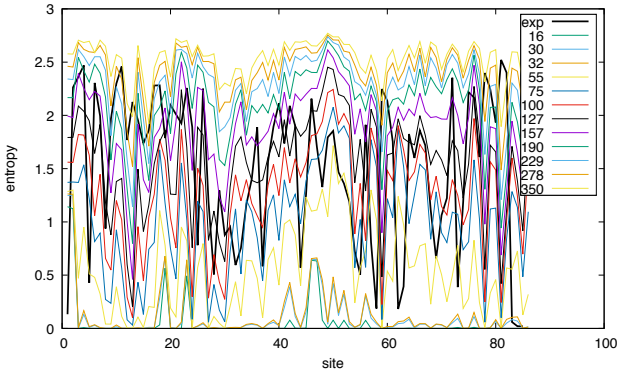


1FKJ





1RX4



2ABD

Figure S4: The entropy per site S_i and its distribution for the synthetic alignments generated by the Monte Carlo calculations at different temperatures and for the natural alignments of proteins 1BPI, 1FKJ, 1RX4 and 2ABD.

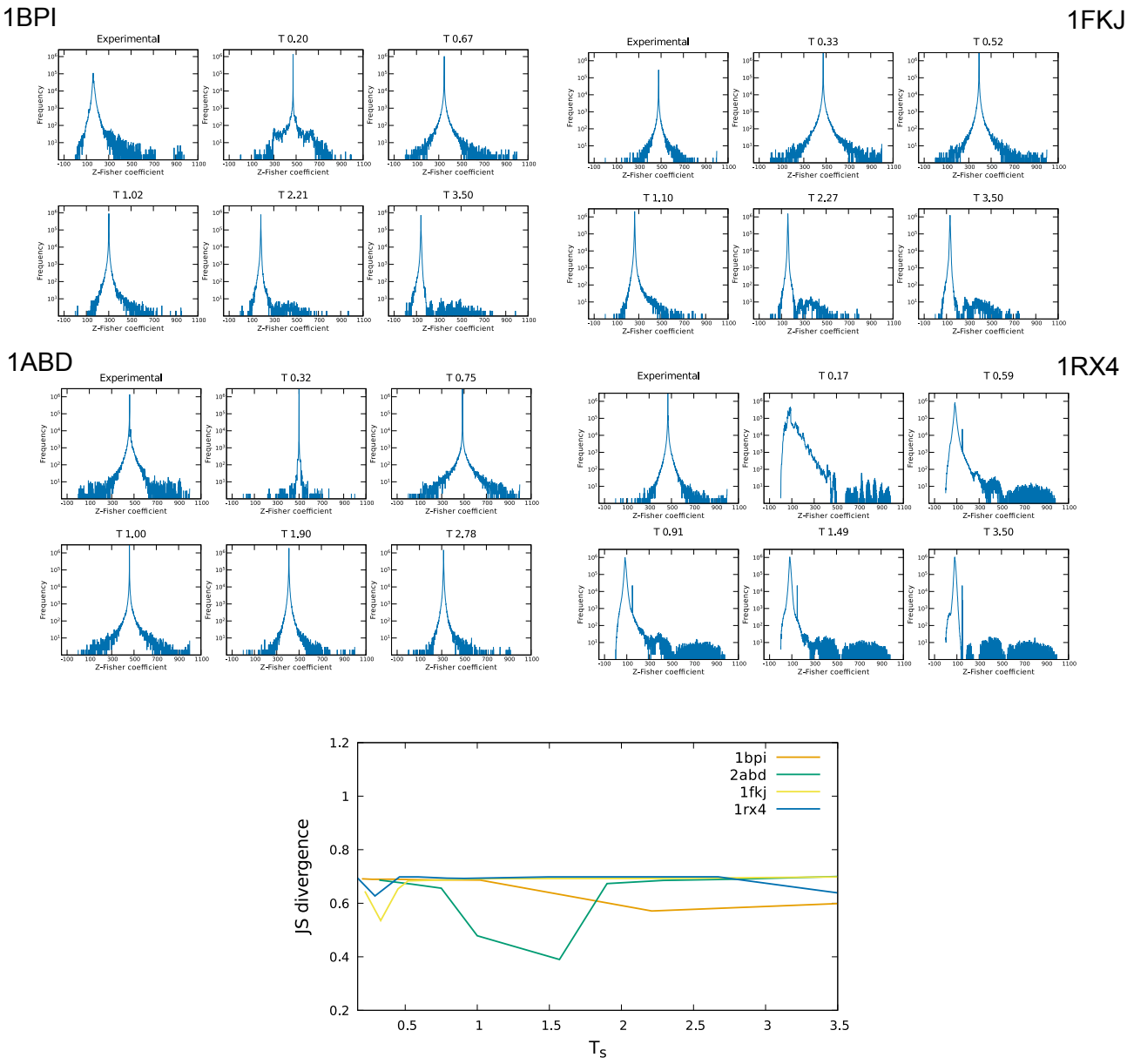


Figure S5: The distribution of Fischer-transformed two-point correlation functions $C_{ij}(\sigma_i, \sigma_j)$ calculated for the experimental alignments and for the sequences generated by the model at different temperatures T_s for the four proteins under study. The Fischer transformation is useful for the visualisation because in case of uncorrelated data would give a Gaussian distribution. The bottom plot displays the Jensen-Shannon divergence between the experimental distribution and those obtained at various temperatures.

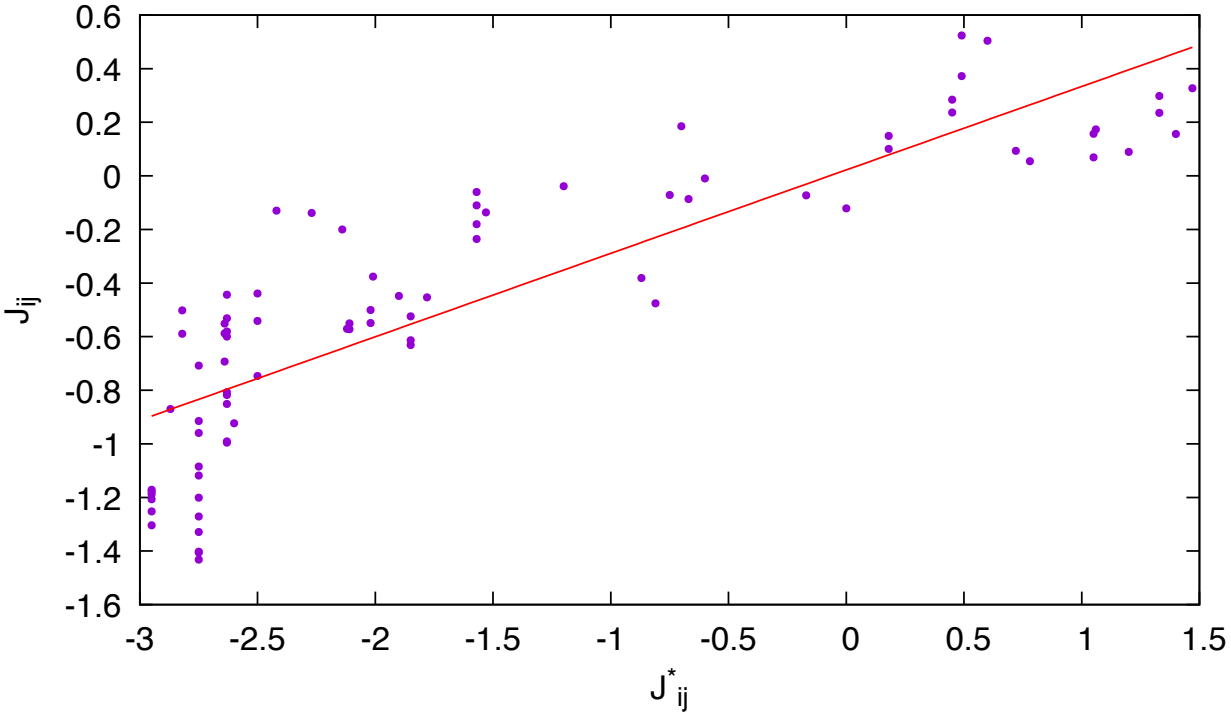


Figure S6: An example of scatter plot between model and back-calculated interaction energies, from which one calculates the Pearson correlation coefficient to assess the validity of the inversion algorithm.

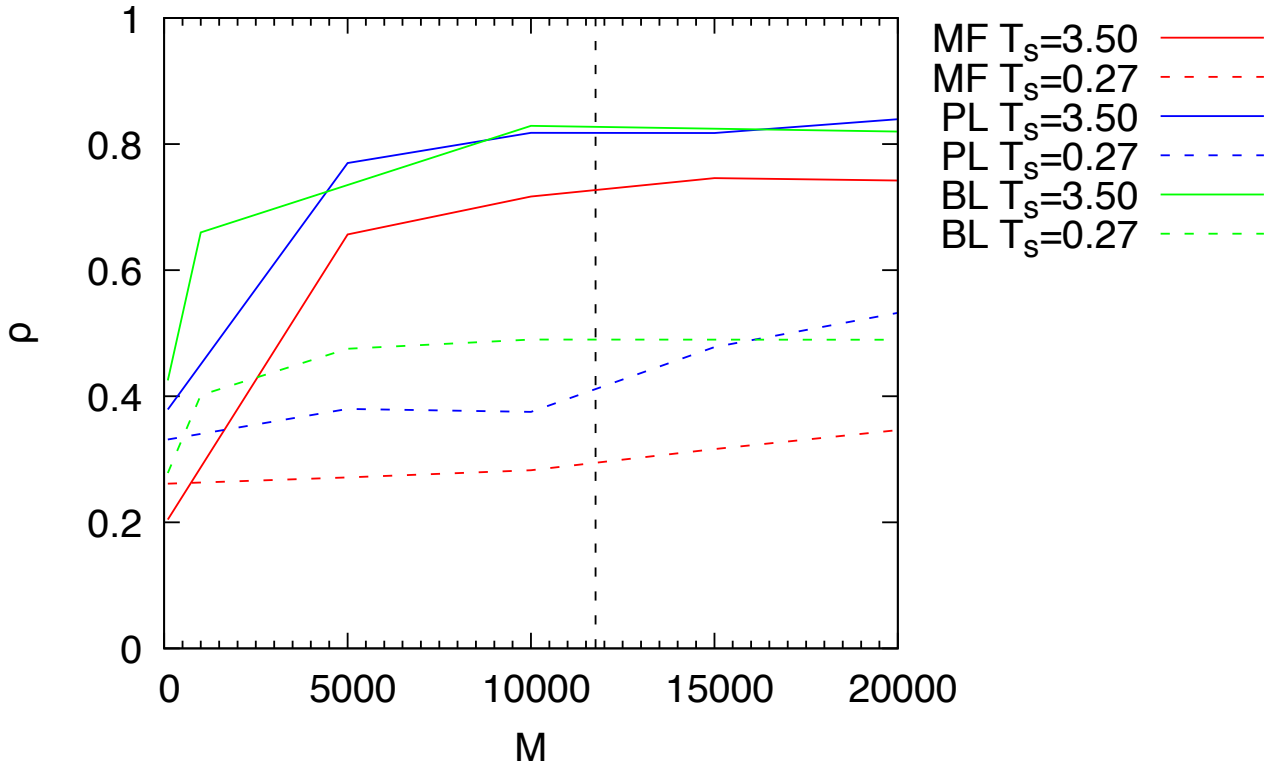


Figure S7: The correlation ρ as a function of the number of sequences M for BPI, calculated with the three algorithms at selected temperatures.

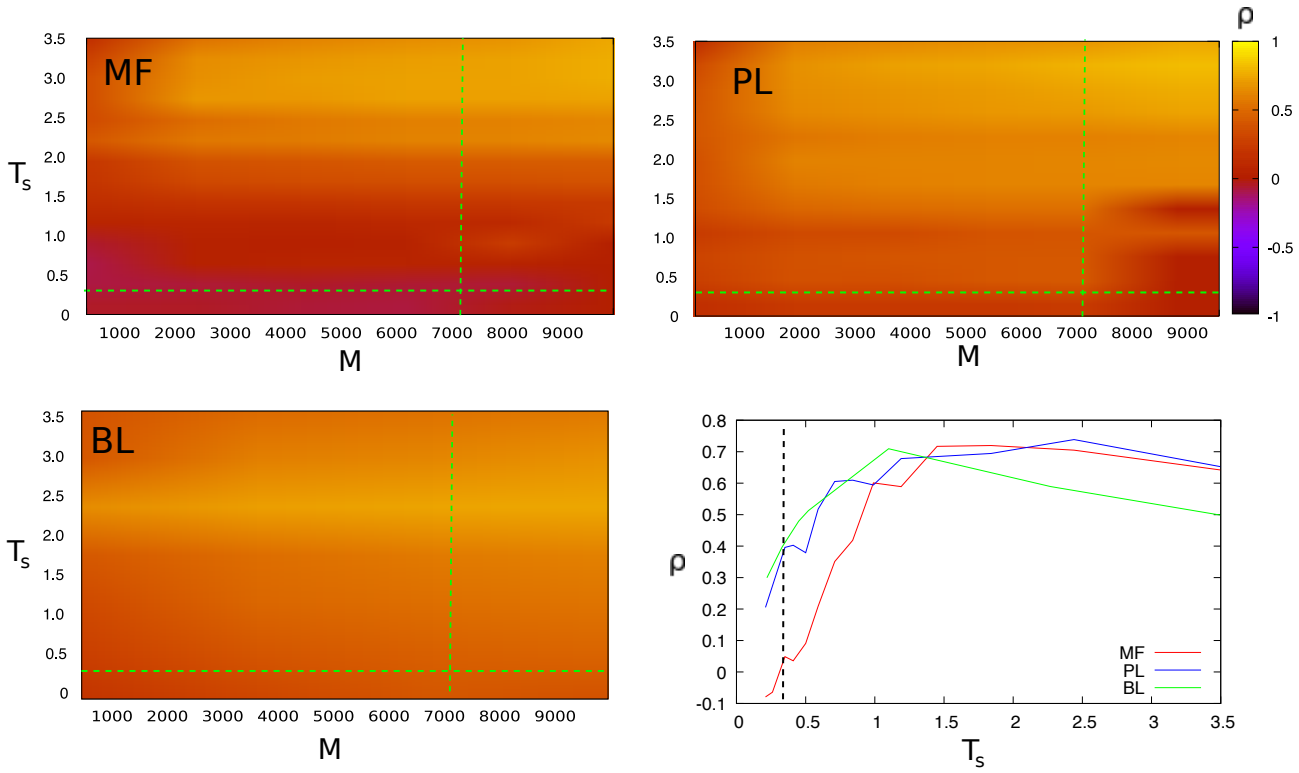


Figure S8: Analogously to Fig. 3 of the main text, the Pearson correlation coefficient between the interaction energies used to generate the synthetic sequences and those back-calculated with the mean-field (MF), the pseudolikelyhood (PL) method and the Boltzmann learning method (BL) for protein 1FKJ. Dashed lines indicate the temperature T_s^n and the number of sequences of the true alignment.

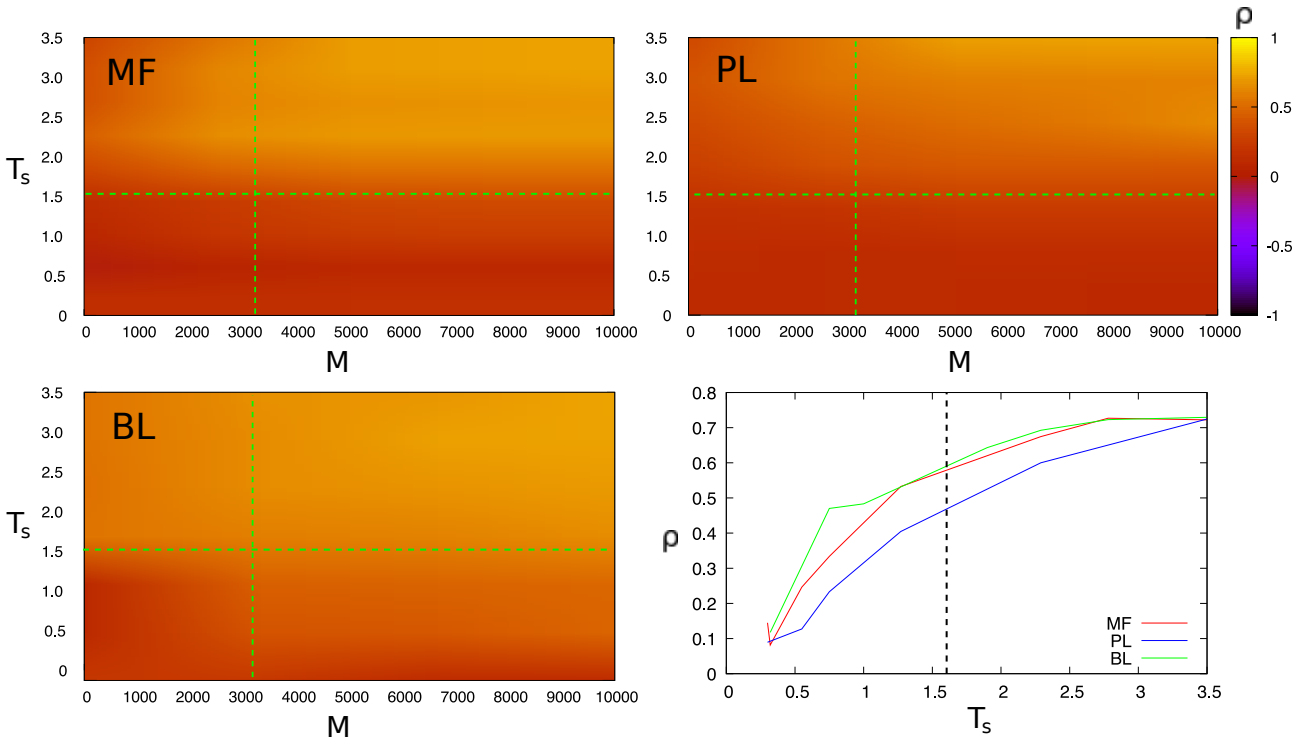


Figure S9: same as Fig. S3 for protein 2ABD.

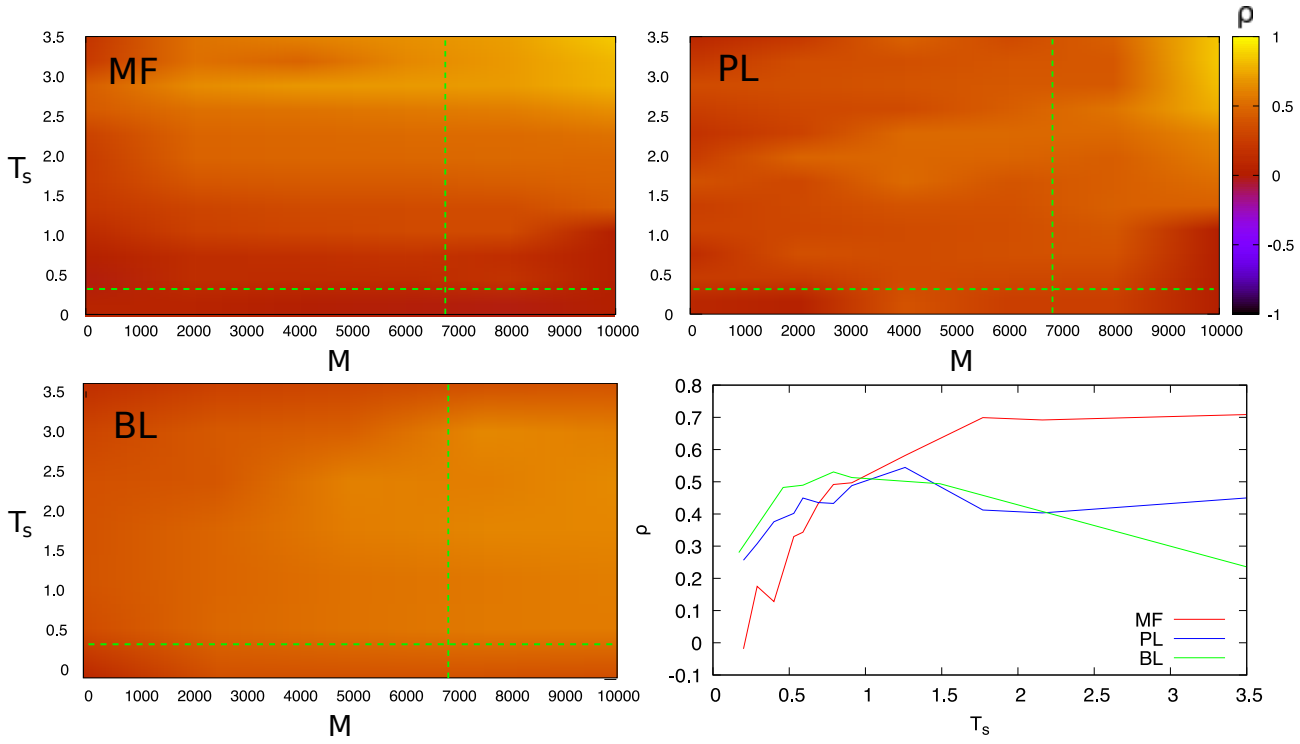


Figure S10: same as Fig. S3 for protein 1RX4.

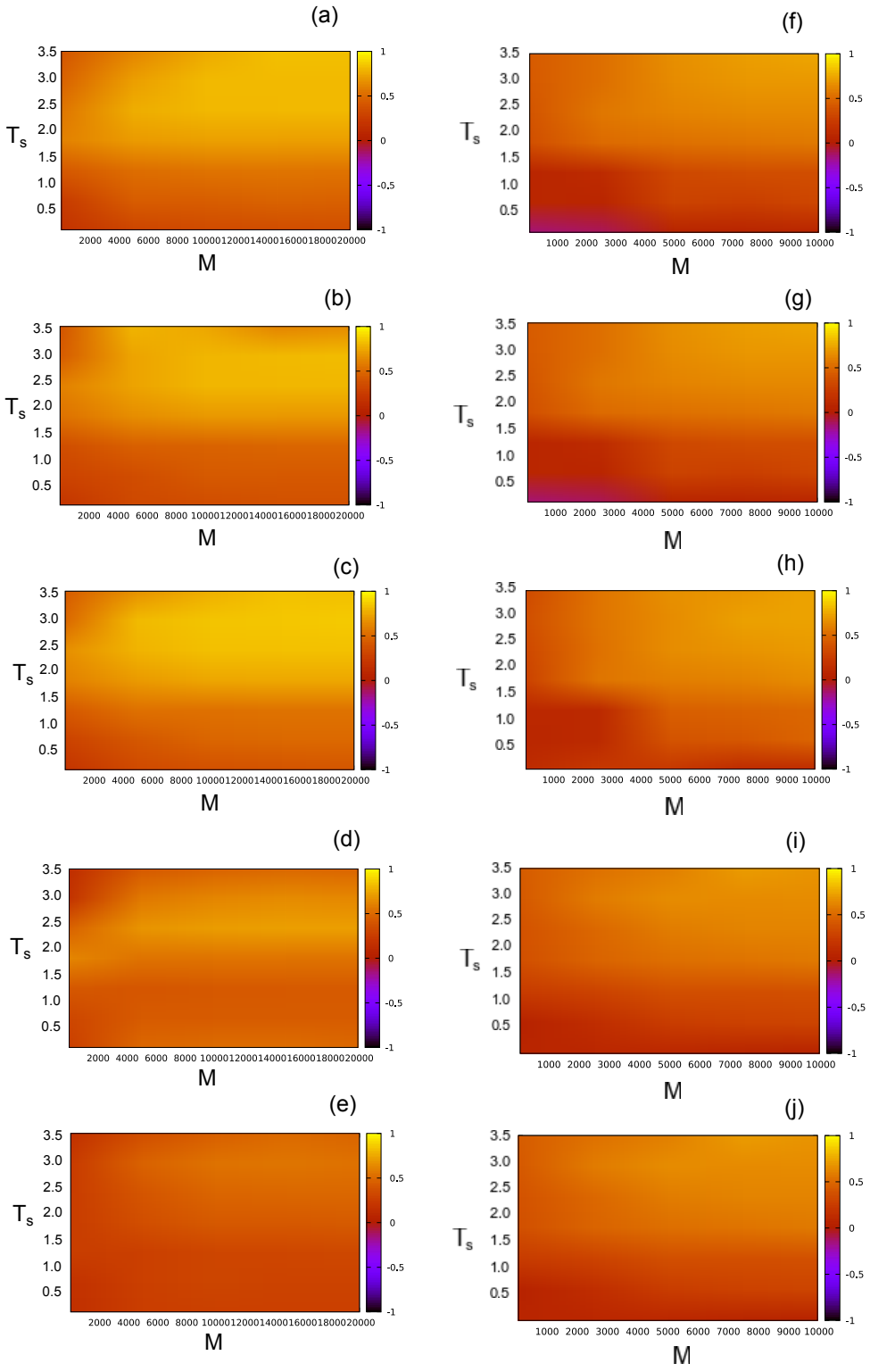


Figure S11: the correlation between original and back-calculated energies by BL for 1BPI (a-e) and 2ABD (f-j). The upper six plots (a-c, f-h) are obtained using sequences selected at random from those sampled by the algorithm, and thus can be regarded as "typical" according to the underlying potential; the bottom four plots (d, e, i, j) are obtained using sequences uncorrelated to those sampled.

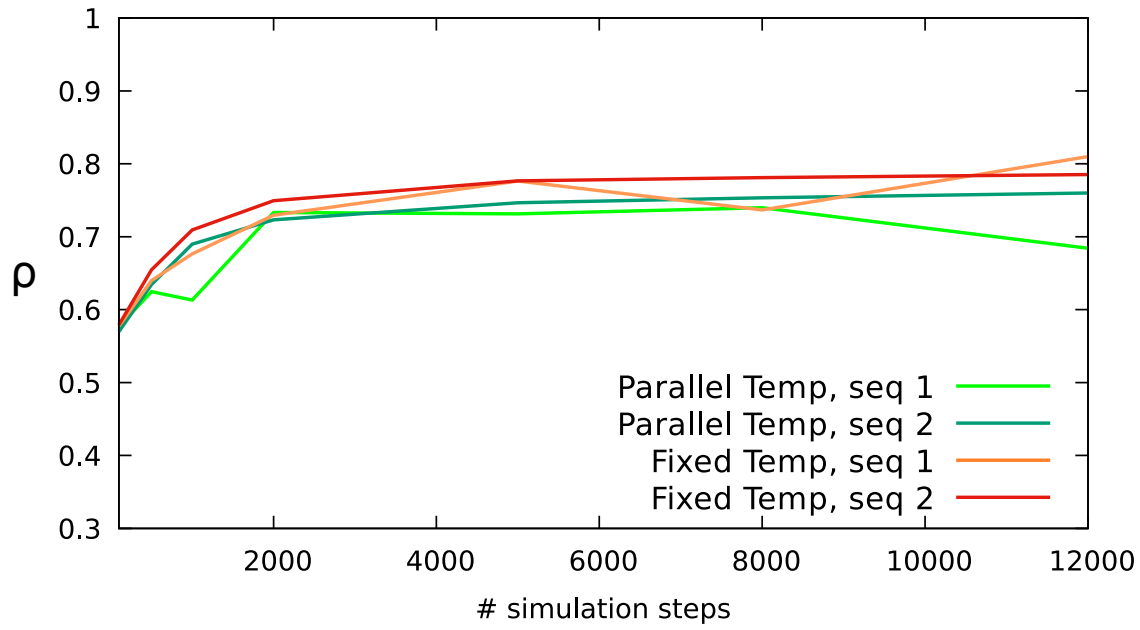


Figure S12: The correlation ρ calculated by BL for two sequences of 1BPI at $T_s = 0.5$ as a function of the number of steps of the algorithm (each step is meant as a Monte Carlo attempt at each site of the chain, followed by a step of the steepest-descent in the optimisation of the energies) with the standard implementation described in *Methods* and with a variant in which the sampling of sequence space is carried out with a parallel tempering method.

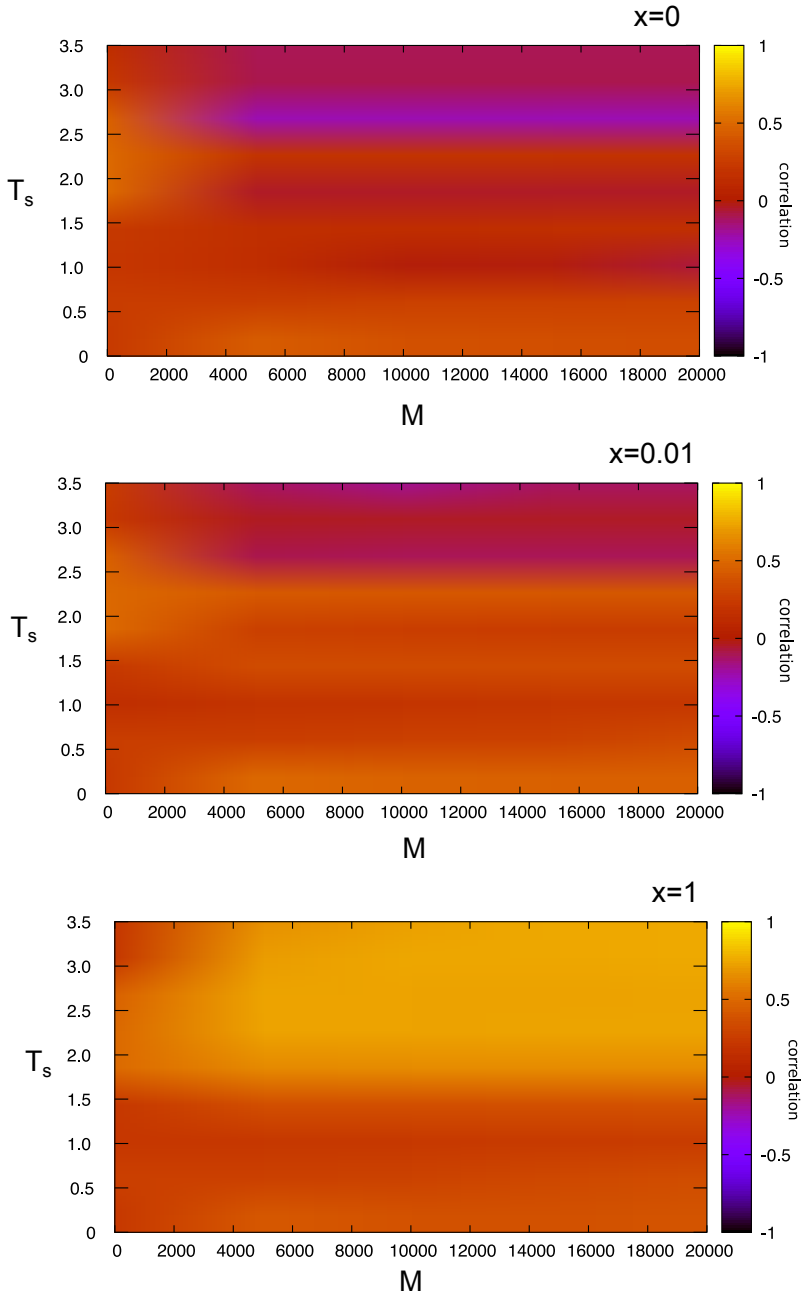


Figure S13: The correlation coefficient ρ calculate for 1BPI with the MF algorithm using three different values of the pseudocounts x.

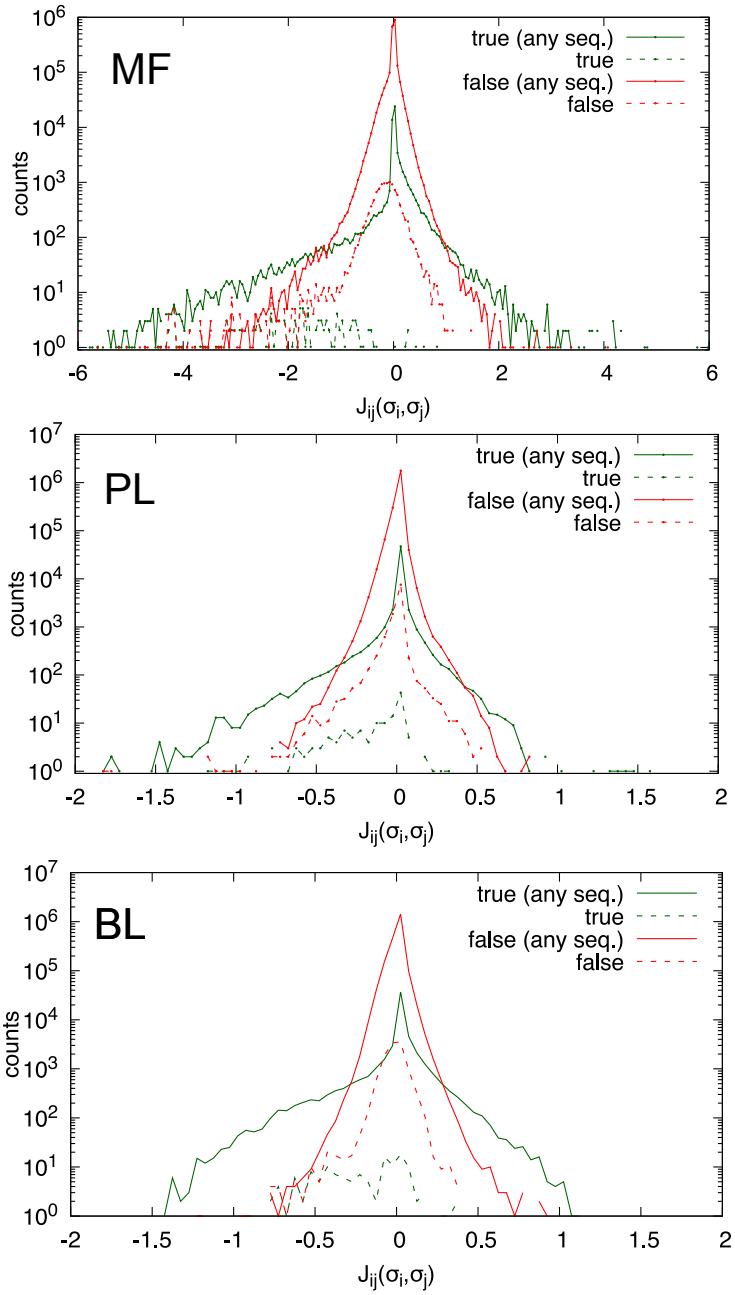


Figure S14: The distribution of interaction energies calculated by MF, PL and BL for 1FKJ at $T_s = T_s^n$ for the true native contacts (solid green curve) and for the non--native contacts (solid red curve). The dashed curves indicate the distribution of native and non--native energies associated only to a "typical" sequence (cf. Fig. S8).

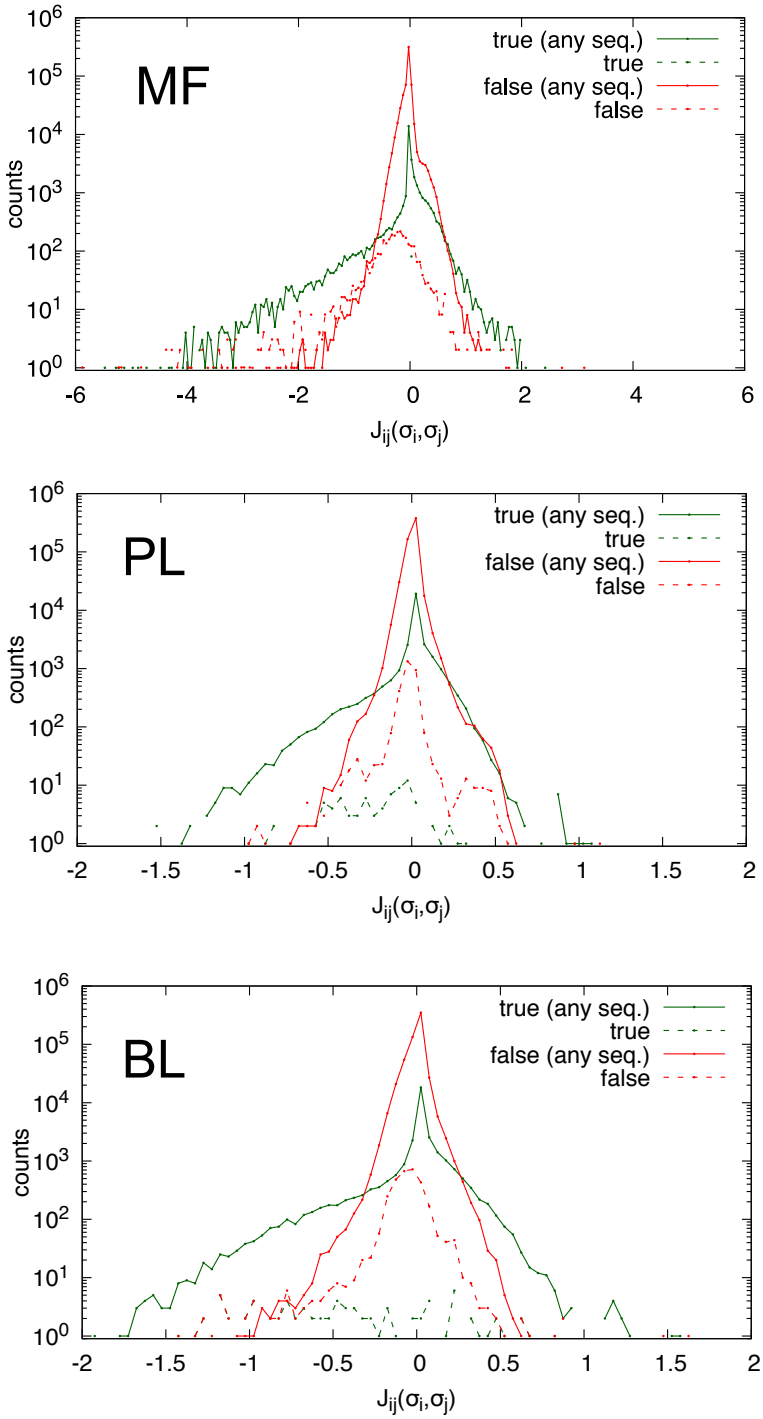


Figure S15: The distribution of interaction energies calculated by MF, PL and BL for 1BPI (analogously to Fig. 5 in the main text), calculated at $T_s = 102$.

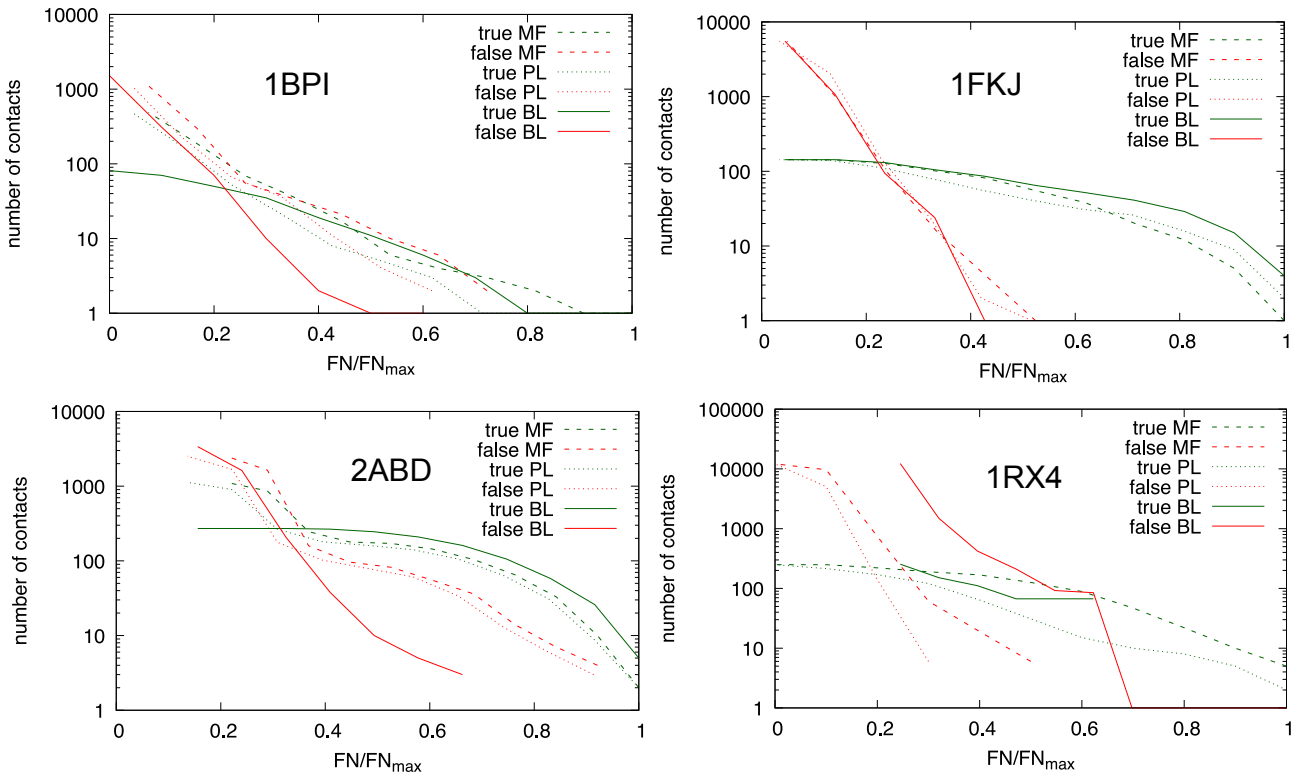


Figure S16: The prediction of the different algorithms evaluated with the Frobenius norm (FN) for the synthetic sequences.

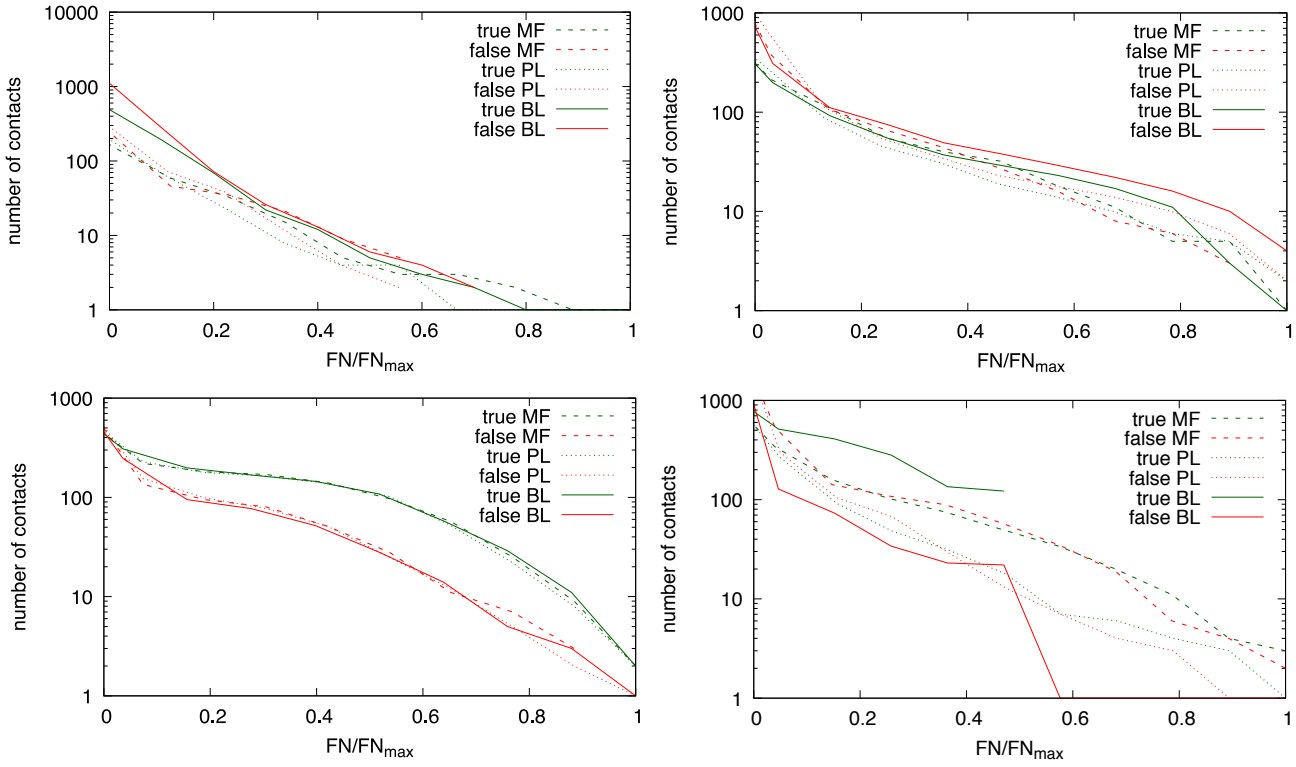


Figure S17: The prediction of the different algorithms evaluated with the Frobenius norm (FN) for the synthetic sequences including the APC correction.

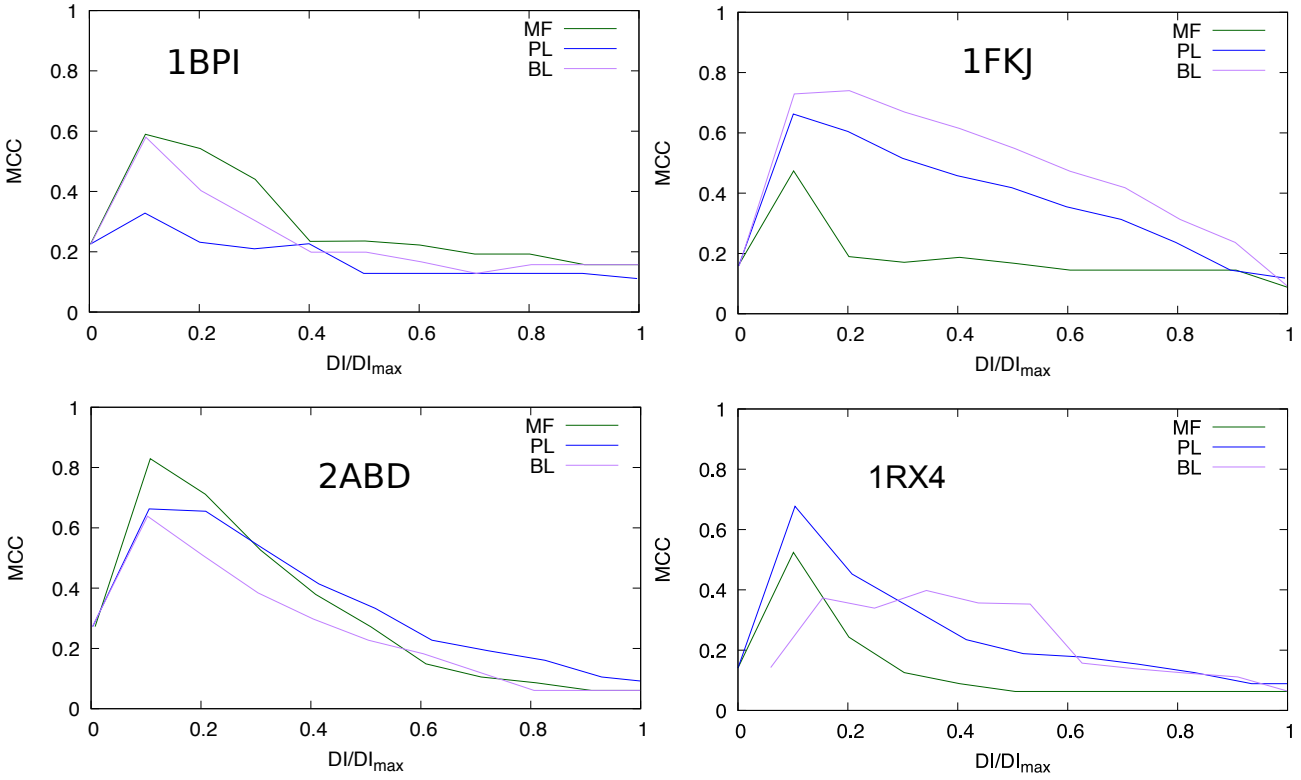


Figure S18: The MCC associated with sequences generated at T_s^n with the three algorithms, as a function of the threshold on the relative direct information.

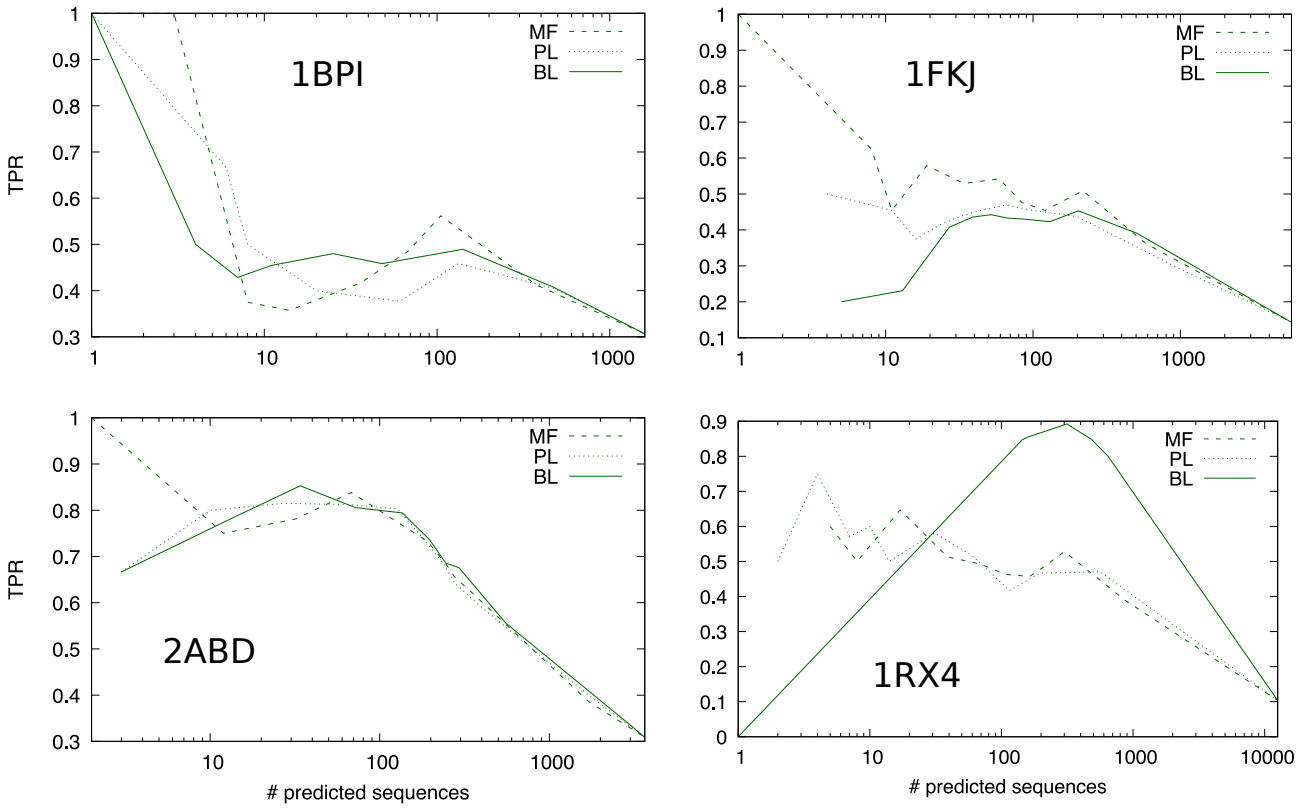


Figure S19: The prediction of the different algorithms evaluated with the TPR corrected by APC.

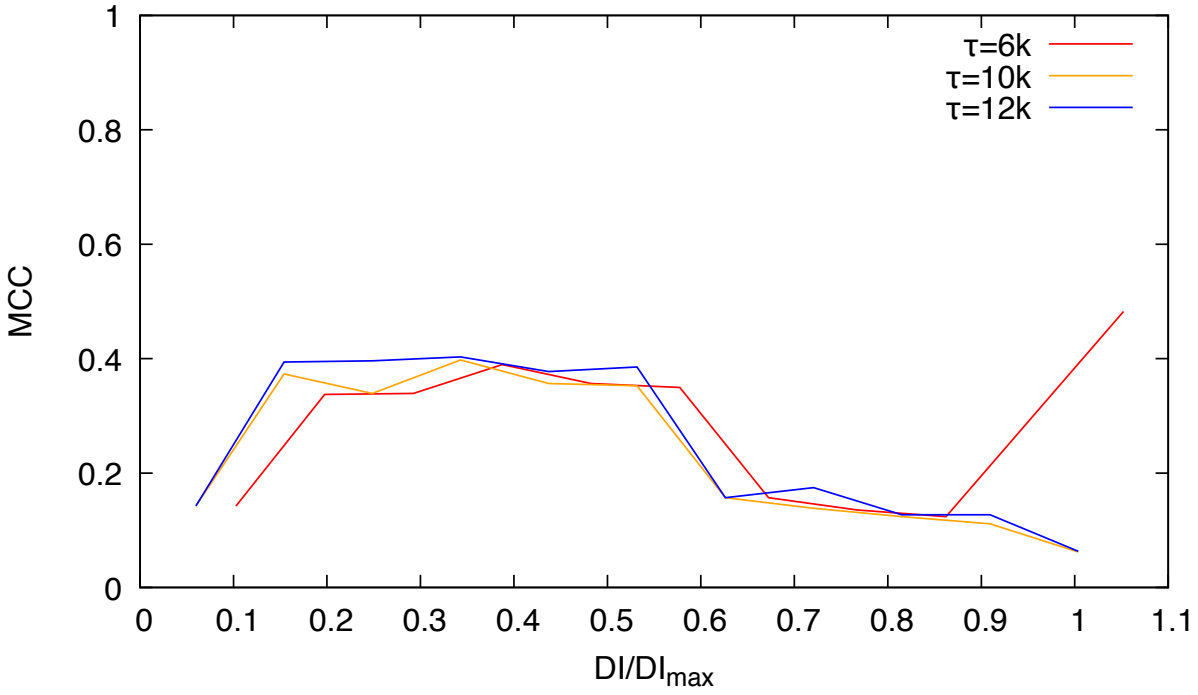


Figure S20: The MCC of 1RX4 at $T_s = T_s^n$ calculated with BL using different values of τ, and consequently different length of the optimisation of the parameters.

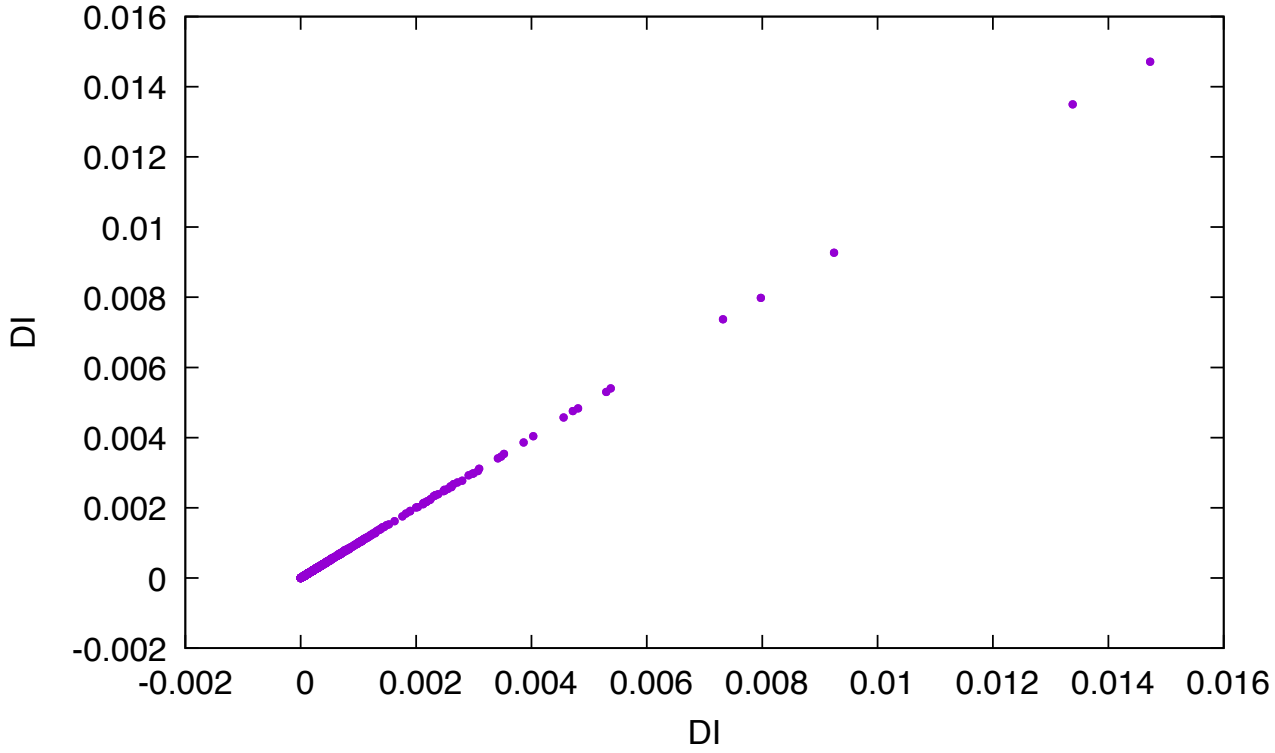


Figure S21: Direct information (DI) predicted for 1BPI at $T_s = T_s^n$ in two independent optimisation runs of BL.

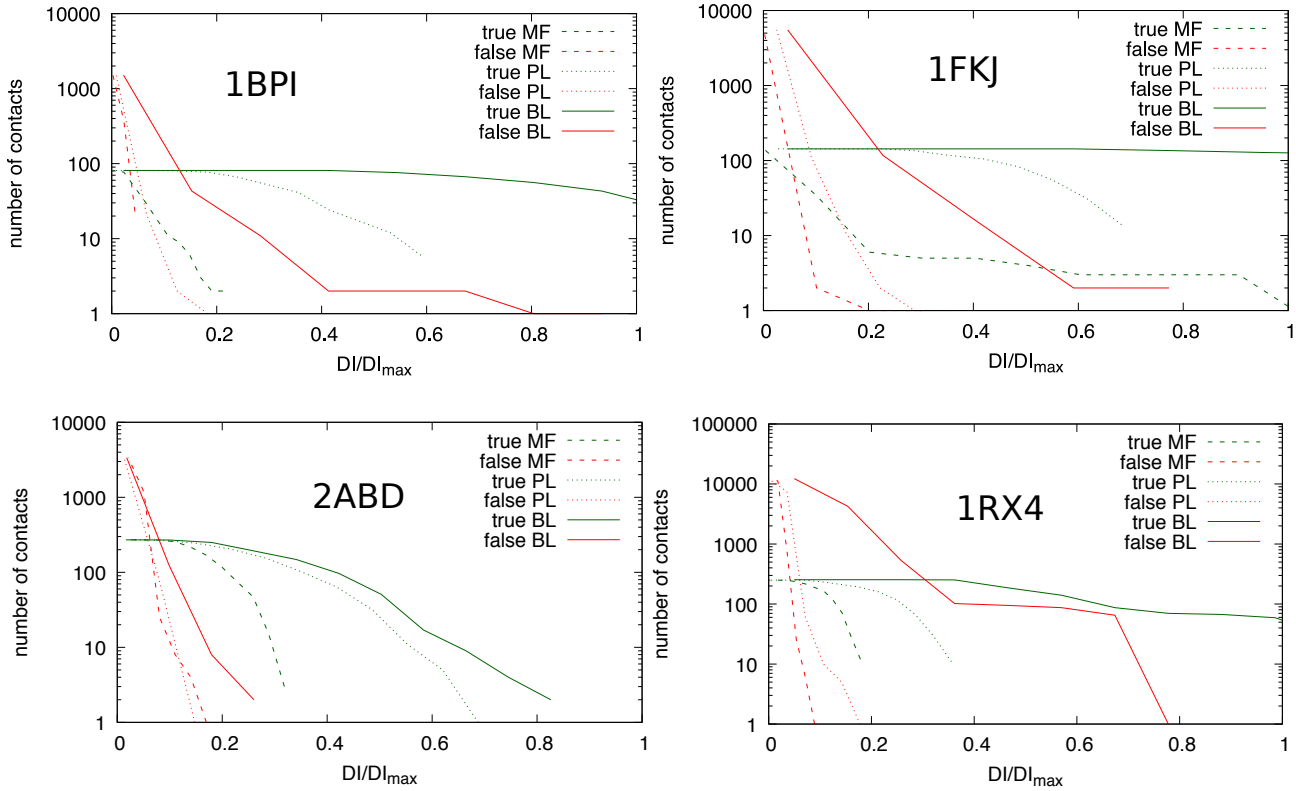


Figure S22: The number of native and non-native contacts obtained from direct information, as a function of the threshold at $T_s = 1$ for 1BPI, 1FKJ and 1RX4 and $T_s = 2$ for 2ABD, that is in the low-temperature, non-glassy phase.

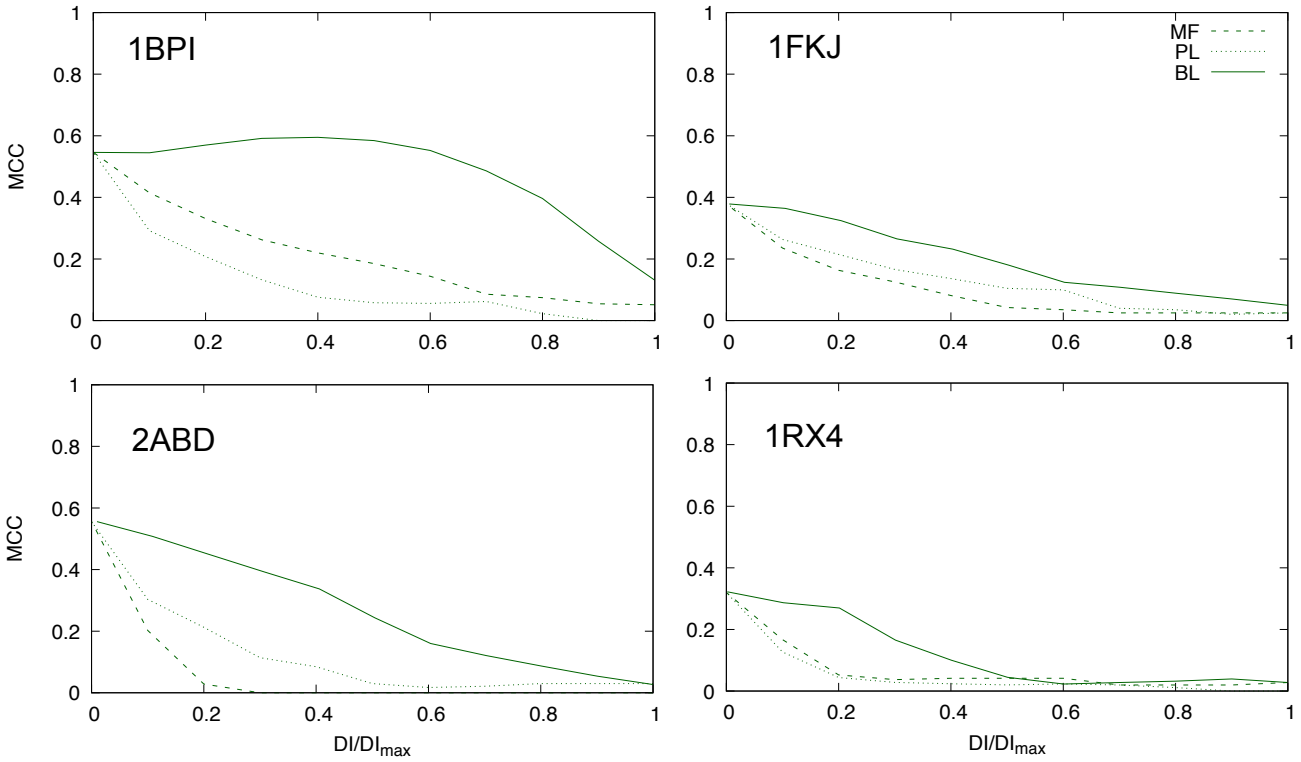


Figure S23: The MCC for the real alignment of the four proteins.

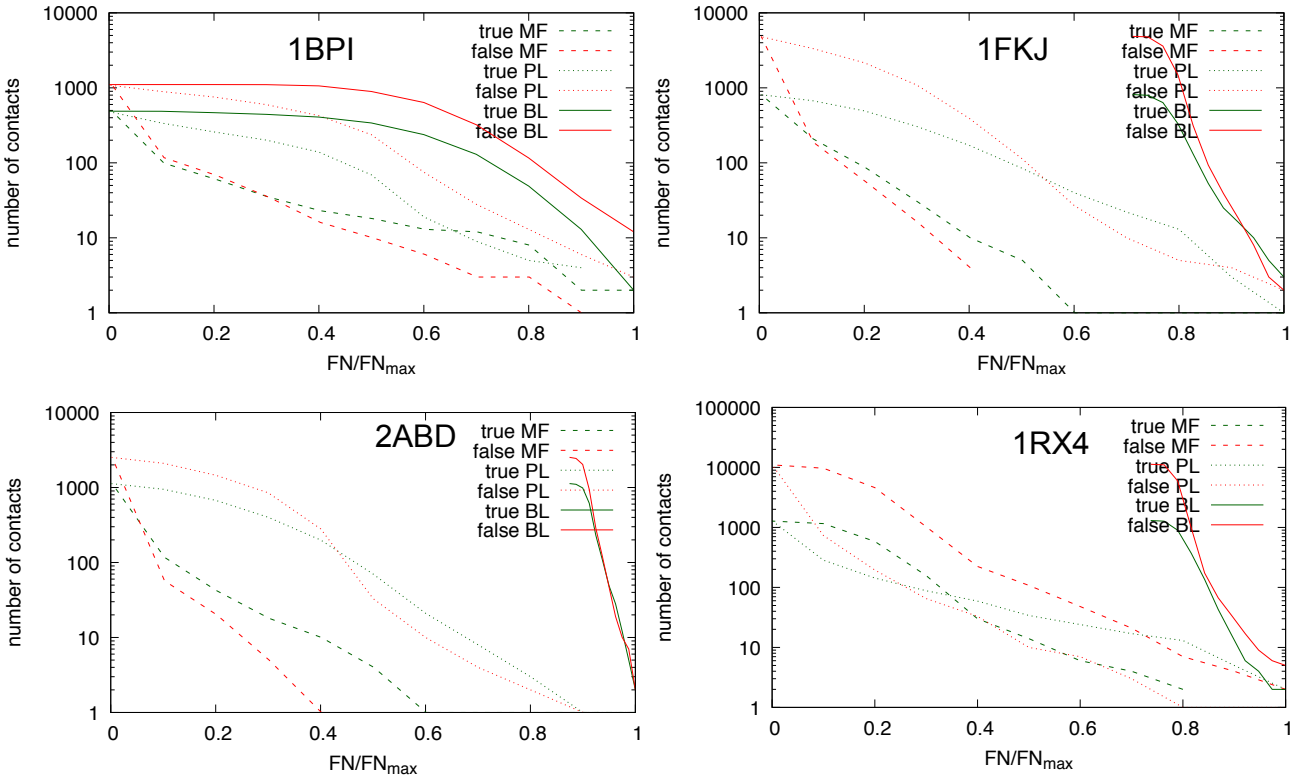


Figure S24: The performance of the three algorithms in detecting native contacts making use of the Frobenius norm (FN).

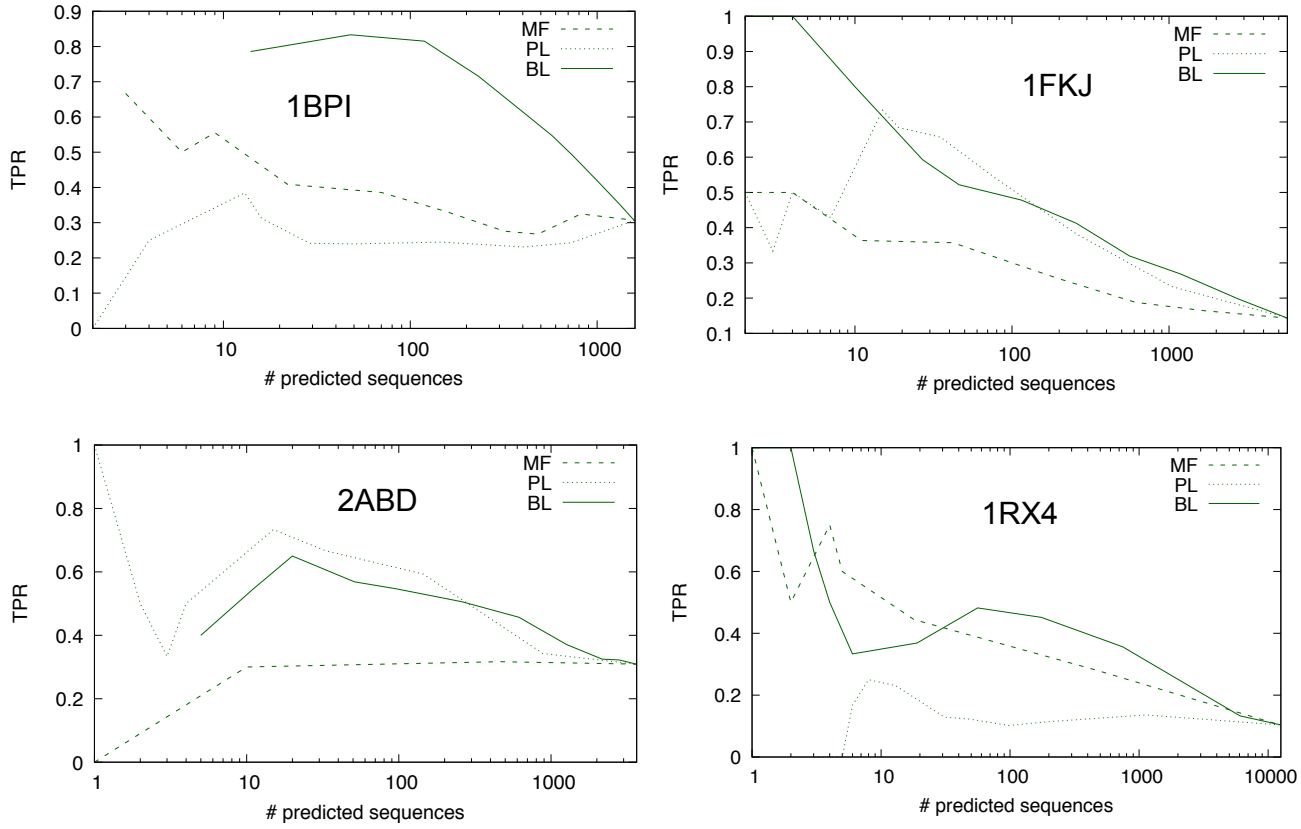


Figure S25: Estimation of the performance of the three algorithms by calculation of the TPR in case of real sequences without APC correction.

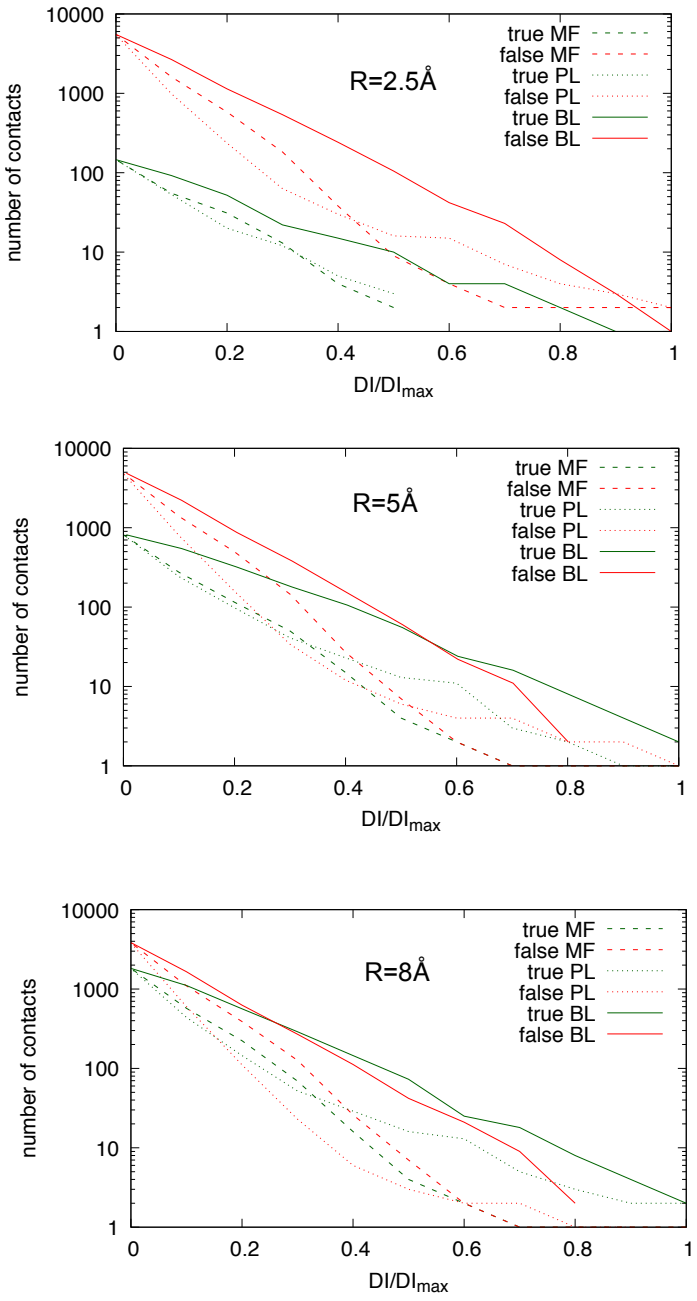


Figure S26: The dependence of the results on 1FKJ on the threshold length R used to define a contact in the crystallographic conformation.

	TPR α	TPR β	TPR tertiary
1BPI MF	0,23	0	0,03
1BPI PL	0	0,06	0,01
1BPI BL	0,45	0,15	0,15
1FKJ MF	0,14	0,13	0,04
1FKJ PL	0	0,10	0,04
1FKJ BL	0,18	0,01	0,04
2ABD MF	0,18	-	0,07
2ABD PL	0,22	-	0,17
2ABD BL	0,27	-	0,29
1RX4 MF	0,22	0,09	0,04
1RX4 PL	0,15	0,05	0,02
1RX4 BL	0,29	0,07	0,06
average	0,19	0,07	0,08

Table S1: The true positive rate (TPR) calculated for the real alignments with the three methods, selecting the top-scoring 10% of the contact, divided into contacts within α-helices, β-structures (hairpins, sheets and turns) and tertiary structure.