EXPERIMENTAL VERIFICATION OF THE TANGO PREDICTIONS

What could be the reason behind the failure to predict the aggregation behaviour of for some sequences? One possibility is that we are comparing different sequences analyzed under very different conditions. Also inaccuracies of the algorithm and force field used could explain the failures. However, it is quite possible that in some cases experimental errors could explain some of the wrong predictions. To find out which of these explanations is correct we investigated some predictions where TANGO predicts aggregation, while the published data reports no aggregation tendency at all (Supplementary Table 1). We selected the alpha4 and alpha7 peptide of glutathione S transferase P domain II (45), the ara2 peptide of Ara protein. [1], the cro5 peptide of cro repressor [2] and the Helix E of sperm whale myoglobin [3]. In Supplementary Figure 1 we show the CD spectra of these peptides measured at two concentrations (50 and 500μM) which include the concentration range previously published. Quite remarkably the behavior for 4 of the 5 peptides was the one predicted by TANGO and not the one published. The CD spectra of the cro5, ara2 and Helix E peptides show clear concentration dependence (Supplementary Figure 1) while the solubility of the alpha7 peptide was too low to permit CD spectra to be recorded. On the other hand the alpha4 peptide showed no concentration dependence although TANGO predicts this peptide to be aggregating. These blind controls thus increase the performance of TANGO on our dataset: discarding the 0.2-5% region we obtain a success rate of 95% (146 out of 154 peptides predicted correctly) yielding a correlation of 0.90 between predictions and experimental data.

Since discrepancies were found in 4 of the 5 false positives of the above set of 179 peptides, and in order to totally exclude uncertainty on the performance of TANGO due to additional errors in the literature dataset, we cross-checked the performance of TANGO by measuring an independent set of 71 peptides derived from three disease-related human proteins (prion protein, lysozyme and β2-microglobulin). All peptides were measured by CD at two concentrations, 50 μM and 1 mM in 50 mM phosphate buffer pH 7.2 at 25 ˚C (Supplementary Table 2). The performance of TANGO on this set is approximately the same as observed on the data extracted from the literature. TANGO

correctly predicts 65 of the 71 peptides of our dataset (91%) giving a correlation of 0.70 in comparison with 0.74 for the literature set.

References

1. Morozova-Roche, L.A. et al. Amyloid fibril formation and seeding by wild-type human lysozyme and its disease-related mutational variants. *J Struct Biol* **130**, 339-351 (2000).
2. Sippl, M.J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* **213**, 859-883. (1990).
3. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. Selection of Representative Protein Data Sets. *Protein Science* **1**, 409-417 (1992).