A typical output of the analysis of NMR restraints (e.g., run of Cyana program) is a set of few dozens candidate 3D structures of the analyzed molecule (protein). However, the analysis does not give information which of those candidates are more favored, and in what ratio in the solution. If the SAXS scattering curve is measured on the same sample, this information can be used for the purpose.

Previous implementations of "ensemble fit" (search for a mix of molecular conformations which matches the SAXS curve) were designed to choose from a huge ensemble generated by molecular dynamics. Therefore the methods must trade off accuracy for manageable speed. Typically, a genetic algorithm is used to generate a candidate subset of conformations. For such a selection a combined SAXS curve is computed, using precomputed model curves for the ensemble members, and it is fitted to the experimental data. The principal drawback of this approach is mutually independent computation of the model curves. The computation involves optimization on a few parameters (unknown constants in the model describing protein-solvent interaction typically), which may end up with quite different values. If those do not match, combining them into a single curve is a questionable approximation.

On the contrary, with a relatively small input set of candidate NMR structures we take a more accurate approach. Both the model parameters, considered globally now, and weights of individual candidate structures (reflecting their presence in the solution) become independent variables of a multidimensional global optimization problem; the optimized value is the accuracy of the fit to the experimental data. The optimization must escape from traps of many local minima therefore we use Monte Carlo search several stochastic enhancements like tunnelling or parallel tempering. We discuss the results on samples where both NMR and SAXS data are available simultaneously.

The method offers several opportunities for parallelization. The SAXS curves are represented by approx. 2,000 points, which offers fine-grain multicore
parallelism with the option to leverage also GPU. Similarly, the computation of the curve, regardless of specific method, follows the Debye formula, ie. sweeping over all pairs of atoms. Finally, the stochastic global optimization can be sped up by several parallel searches, making the algorithm suitable for cluster computation. We discuss benefits of several such options.

The final issue is user friendliness of the entire workflow, which is quite complex, involving several programs to be run, handling different file formats, and setting multiple parameters, ending up with visualization of results. We discuss possible extensions of existing web portal solutions, considering also integration to grid portals like WeNMR.

Primary authors: Dr. KRENEK, Aleš (Masaryk University) ; Dr. KUBÍN, Karel (Masaryk University) ; Dr. ŠTEFL, Richard (Masaryk University) ; Dr. FILIPOVŠIC, Jiří (Masaryk University)

Co-authors:

Presenter: Dr. KRENEK, Aleš (Masaryk University)

Session classification: Biomedicine & Life Science

Track classification: Biomedicine & Life Sciences Applications

Type: Oral