How many conformers do you need?

James Holton

Department of Biochemistry and Biophysics, University of California San Francisco

Division of Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley National Lab

Structural Molecular Biology Group, Stanford Synchrotron Radiation Lab, SLAC

JMHolton@lbl.gov

Everyone knows that macromolecular structures exist as an ensemble of conformations, and that adding too many conformations into their model results in over-fitting. So, where is the balance? It helps to begin by cheating. Starting with a molecular dynamics (MD) simulation of a protein in a crystal lattice, 24,000 conformations were regarded as ground truth, and the question asked: what is the minimum number of conformers needed to represent 24,000 conformations to within experimental error (~3%). Target structure factors for refinement (F_{sim}, which was treated like F_{obs}) were derived from averaging calculated electron density of all the MD conformers, or by taking a subset such as the protein alone, the solvent alone, or even isolating one side chain at a time. Using the rotameric distribution in the MD as a guide, it was found that even the most disordered side chains can still be modeled and refined to R_{free} < 3% against F_{sim} using not more than 14 conformers, and most do not require more than two. Combining all side chains into a single, multi-conformer protein model explained F_{sim} from the protein alone to R_{free} = 4%, but refining this same model against F_{sim} from protein+solvent blows R_{free} back up to 16%. Clearly, a better model for the solvent is required. Using the MD as a guide again only 7000 water atoms are required to bring R_{free} down to 3%, but continuing the refinement for a dozen more cycles invariably blows R_{work} and R_{free} up to 55%. This divergence of refinement was in spite of extending F_{sim} to 1.0 A resolution and using perfect phases as restraints. The present conclusion is that even if a perfect model was built with R_{free} = 3%, it would not survive the refinement process. Stable modelling of macromolecules to within experimental error will therefore require not just improvements in model building, but in the refinement process itself.