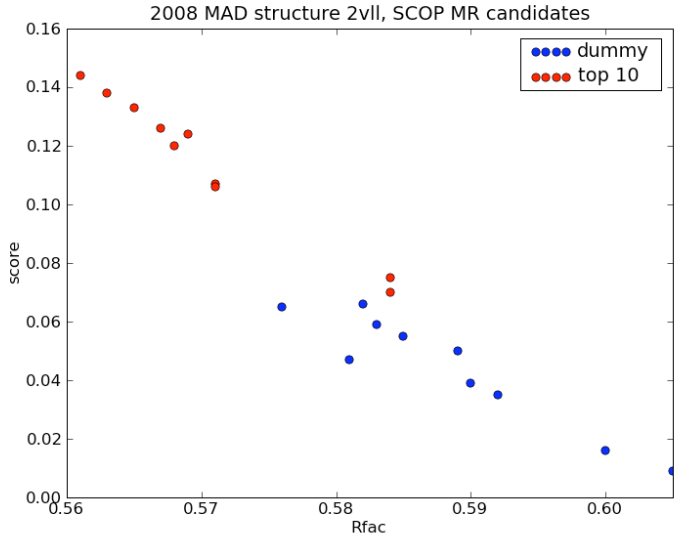
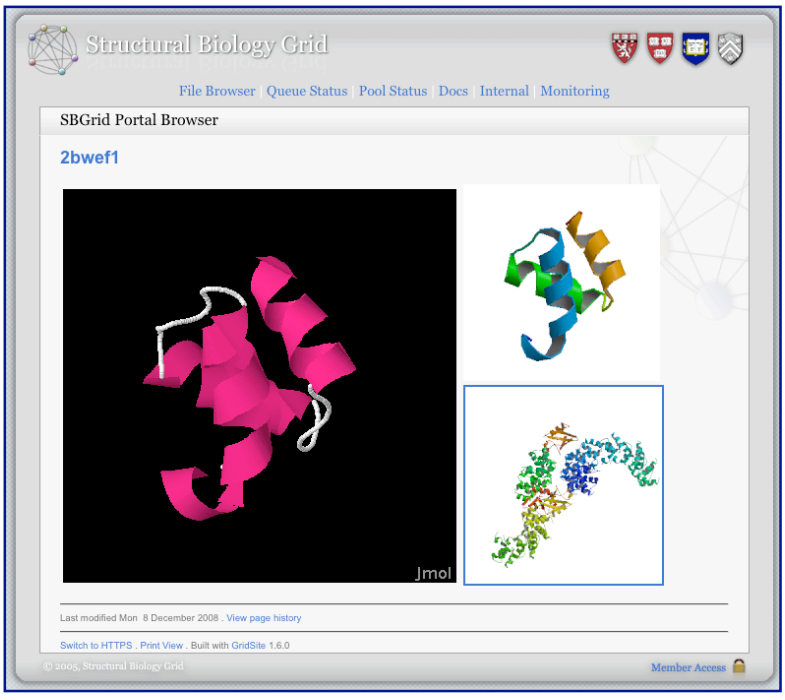
Macromolecular X-ray crystallography is a powerful method that is used to determine macromolecular structures at high resolution. The technique is employed in over 3000 structural biology laboratories, and in just 2008 over 2000 structures have been deposited to the structural biology-clearing house –the RCSB Protein Data Bank. Approximately 65% of all new structures were solved using partial information from models that were previously deposited in the protein data bank, but the remaining 35% of the structures were solved *de novo*, often using a more time consuming and laborious technique called experimental phasing – EP. We have identified some EP models that contain previously known structural elements, but were not utilized as molecular replacement models, often due to minimal or no sequence identity. We hypothesize that the Structural Classification of Proteins (SCOP) database, with over 100,000 macromolecular domains, could be used to determine structures of a subset of EP macromolecules. This would allow automation and acceleration of the structure determination process.



In order to analyze the general utility of our approach we have recently analyzed all 500 EP structures from 2008 and found that many of them include fragments that were previously available in the RCSB. Subsequently we have performed molecular replacement computations that included 10 models with high similarity to each EP structure along with 10 domains randomly selected from the Structural Classification of Proteins (SCOP) database. In approximately 30% of cases the SCOP fragment produced a molecular replacement hit clearly separated from the clustered decoy models. Figure one illustrates this for a newly discovered protein structure. We are now extending our analysis. We have selected 3000 unique and representative domains form the SCOP database and use all of them as a decoy set against the best models. This large-scale computation is currently underway and will be soon completed.

To date we have run this workflow over 600 times and utilized in excess of 20,000 CPU hours at four different Open Science Grid sites (Harvard, University of California at San Diego, University of Wisconsin Madison, and FermiLab). Some sample results and reports can be found at these URLs:

http://abitibi.sbgrid.org/~ijstokes/mr-jobs/2ets/2ets-summary.html
http://abitibi.sbgrid.org/~ijstokes/tmscan-jobs/mad/2-molrep-fast-10/plots.html



We have worked with a team from Open Science Grid to improve efficiency and reliability both of our underlying grid infrastructure and also this specific workflow. This has enabled us to consider more ambitious computational problems, and we are developing web-based interfaces to allow scientists associated with SBGrid to submit various computational jobs to Open Science Grid (see Fig 2 - sample page from our current molecular replacement portal).