Northeast Structural Genomics Consortium
Construct Design and Optimization

NESG construct design begins with bioinformatics-based analysis of the targeted protein family. First the NESG Reagent Genome database is searched for homologous protein sequences using BLAST (http://blast.ncbi.nlm.nih.gov). Then ClustalX (http://www.clustal.org/) is employed to create a multiple sequence alignment (MSA) from high scoring hits. The MSA aids the identification of targeted regions from each hit. Selected protein sequences are further analyzed with an assortment of bioinformatic tools for the purpose of identifying structural features that are known to complicate NMR and X-ray crystal structure efforts. Properties that require special consideration for structure determination efforts include disordered regions, signal peptides, metal binding sites, transmembrane segments, and low complexity regions (commonly the sequence signature of coiled coils).

Rather than sequentially submit each target to each of the individual analysis programs, the NESG has developed a centralized web-based tool called the DisMeta server (http://www-nmr.cabm.rutgers.edu/bioinformatics/disorder). The server has been designed to run standalone or interfaced directly with the target database for batch prediction and parsing of all NESG targets. Dismeta currently employs 14 different disorder predictor algorithms and 6 sequence-based structure prediction tools. We have observed that a disorder prediction based on the un-weighted consensus of several disorder predictors is more robust than any single predictor.

A typical result of our construct optimization approach is illustrated in Fig. 1., the construct optimization of SSP0609 from *Staphylococcus saprophyticus*. The consensus of the 14 disorder prediction algorithms highlighted the N-terminal 55-residues as likely to be disordered. Indeed when 49 N-terminal residues were removed from the construct, the quality of the spectra improved dramatically and the NESG was able to determine its structure by NMR methods (pdb id: 2K3A). Removal of disordered regions also generally improves protein sample behavior and crystallizablity.

The NESG has developed Construct Design scripts that incorporate DisMeta generated information on predicted regions of secondary structure, signal peptides characteristic of secreted proteins, trans-membrane segments, and disordered regions. The NESG Construct Design scripts generate multiple alternative constructs for each 'interest region' (i.e., at least 2 constructs per interest region). N- and C- termini of designed constructs will never lay in a predicted helices or strands. Predicted signal peptides, inter-membrane segments and large disordered regions are excluded from the designed construct and multiple constructs with alternative N- and C- termini are generated per interest region.

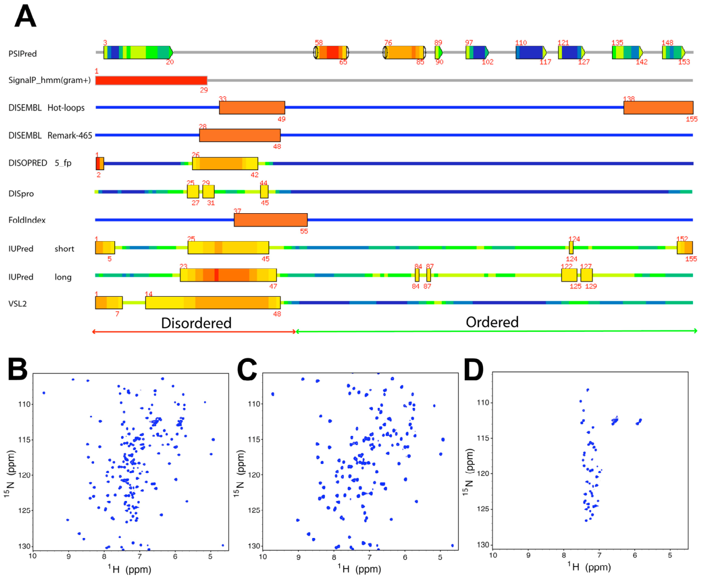

Fig. 1. Construct optimization of *Staphylococcus saprophyticus* SSP0609 protein (NESG target SyR11) and identification of a large disordered segment in the N-terminal region of the protein. (A) DisMeta report showing disorder in the N-terminal 55 residues of the sequence, (B) ^{1}H-^{15}N HSQC recorded at 30 ºC of full length SSP0609 (res. 1-155), (C) ^{1}H-^{15}N HSQC of the best truncated SSP0609 construct (res. 50-155), (D) difference spectrum shows unfolded amino terminal signal peptide. NMR structure was solved (PDB ID, 2K3A)