Northeast Structural Genomics Consortium
Structural Proteomics In the NorthEast (SPINE)
http://spine.nesg.org

SPINE (Structural Proteomics In the NorthEast) is the NESG's inter-institutional project coordination database. SPINE is based on a three-tier software architecture. It is a web-based data management application built upon a MySQL database driven by a series of Perl scripts and Perl modules hosted by an Apache web server. Initially developed to manage the NESG protein production pipeline, SPINE has evolved to include components designed to coordinate activities and projects in the several laboratories across the consortium. The Oracle-based PLIMS (Protein Laboratory Information System) manages the majority of the NESG's protein production pipeline at Rutgers University.

The SPINE data model (Bertone et al., 2001; Goh et al., 2003) closely mirrors the current pepcDB data model, with database tables for tracking target, construct, expression, purification, biophysical characterization, X-ray and NMR data, and protein structure data. Spine generates weekly targetDB and pepcDB data deposition files which are retrieved by the PSI-KB. Spine also generates meta-data required for archiving NESG constructs at the PSI-MR, and provides key NESG protein production pipeline, sample distribution, and structure determination statistics. It is the primary tool for moving information from the NESG project to public databases.

Consortium Coordination

As a web-based application, SPINE is also the primary tool for communicating information between the laboratories of the NESG consortium. The SPINE database tracks *protein samples* created for structure screening and structure determination throughout the consortium via PST (Protein Sample Tube) ids. The entire sample creation and evolution history are available via PST ids (Figure 1). Researchers across the NESG consortium use SPINE to retrieve protein sample data, track the progress of construct and fermentation requests, and for various data mining projects.

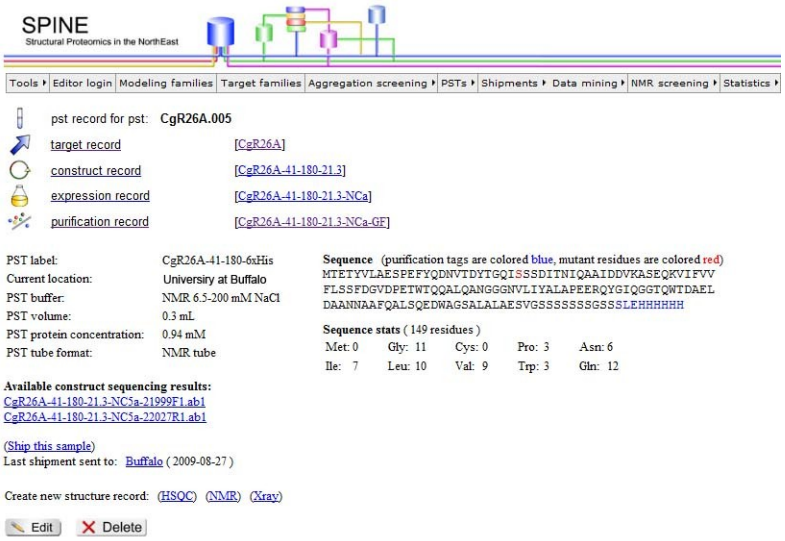

Fig. 1. SPINE's Protein Sample Tube (PST) records provide details about NESG protein samples. and link to the sample's Target, Construct, Expression, Purification, Crystallization, NMR Screening, and 3D Structure SPINE records.

SPINE organizes samples created for NMR screening into 96 sample blocks, programs NMR screening robots, and reports and archives the results of NMR screening runs. Targets that yield well dispersed ^{1}H-^{15}N HSQC spectra, which can be viewed by co-investigators across the consortium, are enriched with ^{15}N,^{13}C isotopes and distributed to the NESG NMR groups. Spine also tracks NMR sample distribution and NMR structure progress (Figure 2). Spine organizes and prepares shipments of selenomethionine labeled targets to Hauptman-Woodward Medical Research Institute (HWI) for high throughput crystallization trials. Successfully Identified crystallizations conditions are reported to both SPINE the NESG's crystallization database (PROTEUS).

Data Integration

Spine is the central archive of all reagent and structure data generated by the NESG. It is closely intergrated with ZebaView (Wunderlich et al., 2004)– the public access point to the NESG protein target list. Target, clone and expression data generated in the sample production pipeline (Acton et al., 2005) are reported to Spine from PLIMS. Predicted (e.g. Zn-binding motifs and disorder predictions) and experimental (e.g. gel filtration with static light scattering) biophysical data are also stored for each protein target. Crystallization progress is reported to Spine from the PROTEUS database, while NMR progress across the

consortium is tracked directly by SPINE. Atomic coordinates, structure validation reports, and raw structural data including FID, chemical shift, constraint, and crystallographic structure factor data, are also archived in SPINE using HarvestDB, the protein structure deposition interface to Spine. Target data and experimental data sets such as the publicly-accessible NESG RDC library and NMR FID libraries are also made freely available to the scientific community through SPINE. The resulting archive, containing data for more than 20,000 cloned constructs, is a valuable resource for datamining the structural genomics process (see for example Goh et al., 2004, Price et al., 2009)

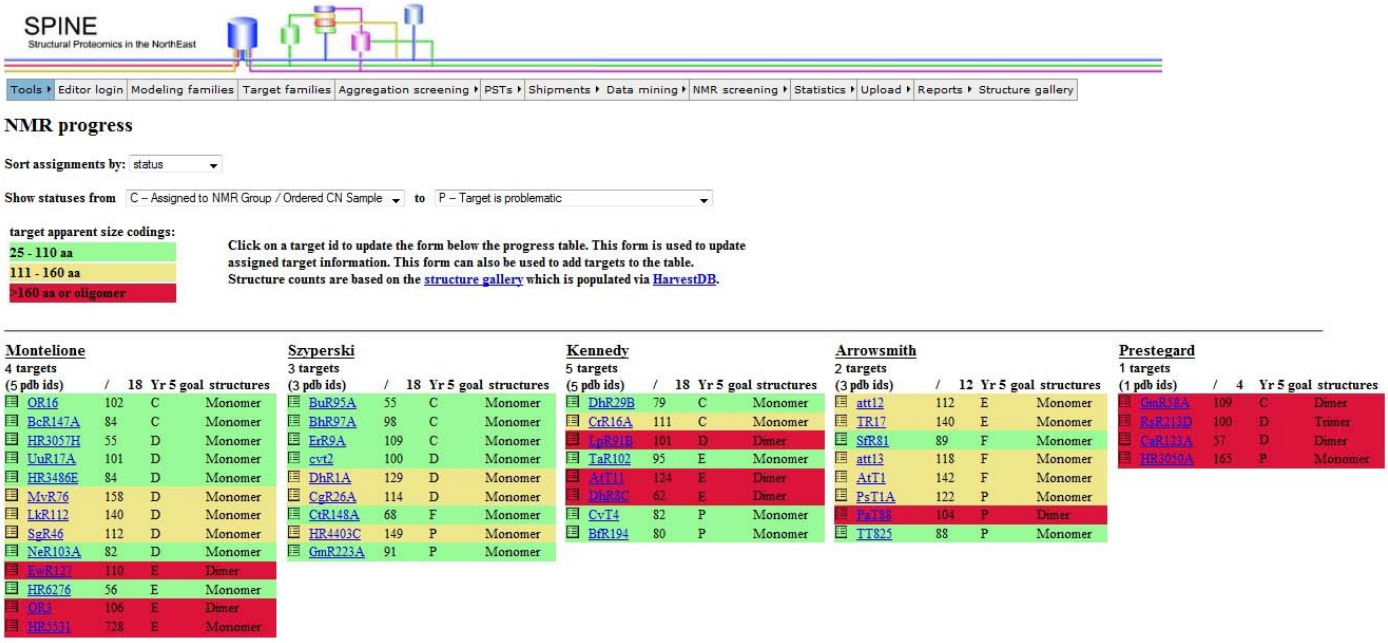

Figure 2. The NMR Target Progress Table is one of SPINE's project coordination tools, providing details of all active NESG NMR targets. SPINE's NMR progress table tracks the NMR targets assigned to the NESG NMR groups at several institutions, and their progress in the NMR structure determination process (preparation, data collection, refinement, problematic, and In PDB).

Bertone, P.; Kluger, Y.; Zheng, D.; Edwards, A.M.; Arrowsmith, C.H.; Montelione, G.T.; Gerstein, M. Nucleic Acids Res. 2001, 29: 2884 - 2898. SPINE: An integrated tracking database and data mining approach for prioritizing feasible targets in high-throughput structural proteomics.

Goh, C.-S.; Lan, N.; Echols, N.; Douglas, S.; Milburn, D.; Bertone, P.; Xiao, R.; Ma, L.-C.; Zheng, D.; Wunderlich, Z.; Acton, T. ; Montelione, G.T.; Gerstein, M. Nucleic Acids Res. 2003, 31: 2833 - 2838. SPINE 2: A system for collaborative structural proteomics within a federated database framework.

Wunderlich, Z.; Acton, T.B.; Liu, J.; Kornhaber, G.; Everett, J.; Carter, P.; Lan, N.; Echols, N.; Gerstein, M.; Rost, B.; Montelione, G.T. PROTEINS: Struct. Funct. Bioinformatics 2004, 56: 181-187. The protein target list of the Northeast Structural Genomics Consortium.

Acton, T.B.; Gunsalus, K.C.; Xiao, R.; Ma, L-C.; Aramini, J.M.; Baran, M.C.; Chiang, Y-W.; Climent, T.; Cooper, B.; Denissova, N.; Douglas, S.M; Everett, J.K.; Ho, C.K.; Macapagal, D.; Paranji, R.K.; Shastry, R.; Shih, L-Y.; Swapna, G.V.T.; Wilson, M.; Wu, M.; Gerstein, M.; Inouye, M.; Hunt, J.F.; Montelione, G.T. Meth. Enzymology 2005, 394:210-243. Robotic v cloning and protein production platform of the Northeast Structural Genomics Consortium.

Goh, C.-S.; Lan, N.; Douglas, S.; Wu, B.; Bertone, P.; Echols, N.; Milburn, D.; Montelione, G.T.; Zhao, H.; Gerstein, M. J. Mol. Biol. 2004, 336: 115-130. Mining the structural genomics pipeline: Identification and analysis of protein properties that affect high-throughput experimental analysis.

Price, W.N.; Chen, Y.; Handelman, S.K.; Neely, H.; Manor, P.; Karlin, R.; Nair, R.; Liu, R.; Baran, M.; Everett, J.; Tong, S.N.; Forouhar, F.; Swaminathan, S.S.; Acton, T.; Xiao, R.; Luft, J.R.; Lauricella, A.; DeTitta, G.T.; Rost, B.; Montelione, G.T.; Hunt, J.F. Nature Biotechnology 2009, 27: 51 - 57. Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data