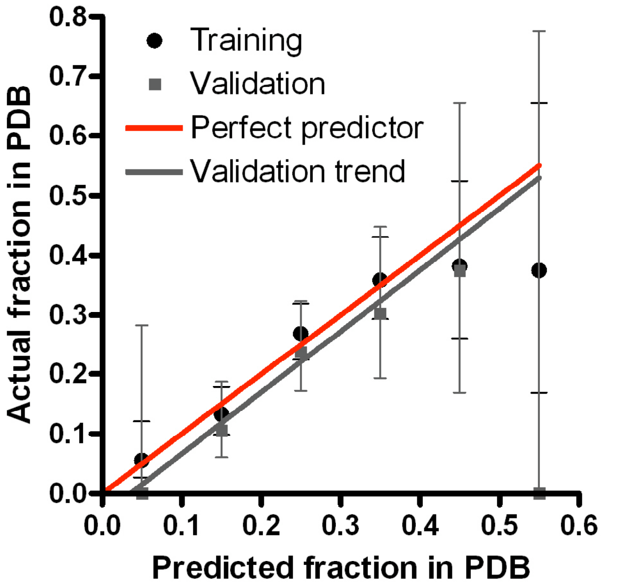
The Northeast Structural Genomics Consortium
The Pxs Server: Probability of Crystal Structure

Protein crystal structures are crucial to elucidating biological function, but the process of obtaining crystals suitable for structure determination is often arduous and unpredictable. Both structural genomics and individual structural biology laboratory groups are benefited by better understanding the molecular mechanism of protein crystallization and the intrinsic protein features that favor or disfavor crystallization. In particular, knowing whether a specific protein construct is more or less likely, other factors being equal, to provide an eventual crystal structure can be useful for target selection, homolog screening, and target triage within a structure determination effort. Data from structural genomics efforts is uniquely suited to answer these questions. Since proteins are taken through a uniform, controlled, and carefully observed pipeline, external factors are held constant and only intrinsic protein features influence structure determination outcome. Just as important, the large number of proteins taken through the same pipeline allows the identification of statistically significant, rather than anecdotal, effects.

P_{XS} (probability of crystal structure solution, publicly available at www.nesg.org/PXS/) is the probability of a monodisperse and biochemically well-behaved protein preparation giving a crystal structure, assuming thorough but not exhaustive crystallization efforts. Using structural genomics data, we determined that the formation of high quality protein crystals depends primarily on the prevalence of well-ordered surface epitopes suitable for crystal packing contacts. Several individual sequence features were identified as statistically significantly contributing to this. The fraction of glycines in short surface loops and the fraction of phenylalanine were positively correlated with successful structure determination. The fraction of residues predicted to be disordered by the program DISOPRED2 (Ward *et al.*, 2004) and the mean side chain entropy (Creamer *et al.*, 2000) of predicted surface exposed residues (Rost, et al., 2004) were negatively correlated with successful structure determination. P_{XS} combines and weights these four factors to predict overall protein crystallizability. The predictive value of P_{XS} for development and test data for the P_{XS} metric are shown in Fig. 1.



Bin Center	0.05	0.15	0.25	0.35	0.45	0.55	Total
N (Train.)	72	196	238	131	34	8	679
N (Valid.)	8	66	84	33	8	1	200

Fig. 1. P_{XS} performance in predicting the probability of successful crystal structure determination. 679 development proteins were all well-behaved monodisperse proteins taken through the NESG pipeline, and 200 test proteins were similarly taken through the pipeline after the metric was developed. Proteins were binned in intervals based on predicted likelihood of yielding a structure (P_{XS}) and the actual fraction of proteins in each bin which yielded structures suitable for PDB deposition was calculated. The metric accurately predicts the probability of structure solution up to probabilities around 0.35, after which the metric loses accuracy due to the small numbers of proteins in those bins.

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

Ward *et al.*, 2004 Bioinformatics *20*, 2138.

Creamer *et al.*, 2000 Proteins *40*, 443.

Rost, Yachdav, & Liu, 2004 Nucleic Acids Res *32*, W321.