PREDICTION OF PROTEIN CRYSTALLIZATION OUTCOME USING A HYBRID METHOD

***PROSPERO*: Prediction of Outcome from Sequence and Experimental Results Online**

Frank H. Zucker, Christine Stewart , Ethan A. Merritt and Wim G. J. Hol

Biomolecular Structure Center, Biochemistry Department, University of Washington, Seattle, WA

Data from structural genomics and large structural biology labs worldwide

Outcomes

Sequences

Experiments

Input Modules
Downloaded scripts convert experimental data into a standard form for upload to the *PROSPERO web* site.

Model Training

Upload data from many samples to web site for model training

External Web Resources

Prediction Model
(on Seattle server)

Prediction

Upload data from one sample to web site for outcome prediction

Any lab can get predictions and suggestions on samples

Sequence

Experiments

Input Modules

Experiments include:

DSF

Differential Scanning Fluorimetry: Fluorescence of SYPRO Orange dye versus temperature. Exposed hydrophobic residues dequench dye. Blue: intensity at T_m. Red: intensity at 30 °C. R_{30}=I_{30} / I_{mc}. Green: I_{30} threshold from the R_{30} criterion for HyXG-1, ie 0.105×I_{mc}.

SEC

Size Exclusion Chromatography: Absorbance at 280 nm fit by gnuplot with a single Gaussian curve and a linear background. **SEC_{fit}**=residual A_{280} after fitting one Gaussian, **SEC_{pp}**=purity of pooled peak after fitting multiple Gaussians.

DLS

Dynamic Light Scattering: For the major peak with R_h of 1 to 10 nm, record polydispersity and percent intensity, excluding small molecule peaks. Calculate **DLS_{MW}** from R_h and **DLS_{MW}**=DLS_{MW}/MW of monomer estimated from sequence.

Yield

Yield: High-throughput expression screening gels, crude lysate (left lane of each pair) and soluble protein bound to nickel resin (right lane) from the equivalent of 48 µL culture. **Yld_x** scores are shown below. Yld_x of 5=~100 mg protein per liter of culture.

Data from limited proteolysis and SDS PAGE were also used in training HyXG-1 but were not predictive in the final model. Other experimental methods to be potentially incorporated include native PAGE, NMR, mass spec, static light scattering, and any results you've got uniformly recorded for many samples.

Estimated MW and **Dis_{max}**=longest stretch of disorder predicted by DisEMBL (dis.embl.de) were used in the final model. Other sequence variables tested include average hydrophobicity, XtalPred score (ffas.burnham.org/XtalPred), P_{XS} and $P_{C\text{-}XS\text{-}Hs}$ (nmr.cabm.rutgers.edu:8080/PXS/).

ABSTRACT

The great power of protein crystallography to reveal biological structure is often limited by the tremendous effort required to produce suitable crystals. A hybrid crystal growth predictive model that combines both experimental and sequence-derived data from target proteins is shown to be more powerful than sequence-based prediction alone – and is likely to be useful for prioritizing and directing the efforts of structural genomics and individual structural biology laboratories.

In addition to predicting outcome, the HyXG-1 decision tree model also suggests which next steps should to be taken when a protein sample fails to crystallize in initial trials: further trials for samples predicted as likely to crystallize; changes to expression, purification or sequence for other samples.

Additional methods of protein characterization and data from additional samples will further improve the model. We are developing a server to predict crystallization based on the current model and to accept additional data to increase the applicability and predictive power of such hybrid models.

COLLABORATE

We are eager to extend our initial training set through collaborations to include data from other large-scale crystallization projects. By adapting the input stages to handle new classes of experimental characterization (e.g. NMR, mass spec, static light scattering), or to score the outcome of standard protocols used elsewhere, we hope to generate customized predictors for individual labs or projects. If you are interested in collaboration or if you can offer access to collections of protein characterizations and corresponding crystallization outcome data, please contact us at

Dr. Ethan Merritt **<merritt @ uw.edu>**

or on the web at

http:// skuld.bmsc.washington.edu / prospero

or leave us your email address below.

Model evaluation

Train models on one set of samples, then test on a separate set with a similar outcome distribution. Optimize for best correlation between observed and predicted DS (DS_O, DS_P), lowest error = sum of squares of DS_O-DS_P and highest area under ROC curve, true positive vs false positive rate. Shown here: HyXG-1, decision tree trained on 77 samples, tested on 30 (Zucker et al. 2010 J. Struct. Bio., in press).

Possible models

Decision Tree

Recursive Regression Partition Tree: find criteria which divide samples by outcome. Gives both prediction and salvage suggestions.

Bayesian

Naïve Bayesian: use observed success rate for bins of sample properties to predict outcome. May give both prediction and suggestions.

SVM

Support Vector Machines: draw boundaries among clusters of samples for best separation of outcome by sample properties. Gives predictions only; suggestions could be derived by trials with artificial samples.

Clustering

K-means Clustering: find a clustering by properties which also clusters outcomes. Gives predictions only.

Predicted Outcome & Salvage Suggestions

The decision tree, predicted outcome and suggestions are based on 77 training MSGPP samples. The percentages of samples with diffraction to at least 10 Å or 2.8 Å is from the combination of the 77 training samples and the 30 test samples.

RESULTS

*(Zucker et al. **J. Struct. Bio**. in press, doi:10.1016/j.jsb.2010.03.016)*

We quantified 21 variables from experimental results and sequence. Several of these variables are novel parameters derived from biophysical characterization experiments. Models were trained on 77 protein samples from the Structural Genomics of Pathogenic Protozoa (SGPP) consortium and the Medical Structural Genomics of Pathogenic Protozoa (MSGPP, www.msgpp.org) project and tested on 30 other MSGPP proteins. This yielded a recursive regression partition tree with more predictive power than models produced by linear regression, naïve Bayesian analysis, SVM or clustering.

The partition tree predicts that low MW proteins with low initial intensity in differential scanning fluorimetry (thermofluor) experiments are likely to produce well-diffracting crystals. Larger proteins with extremely high soluble expression screening yields are also good candidates to produce diffracting crystals. Other samples have lower probability of success, with slightly better outcomes predicted for large proteins with Gaussian SEC curves, or with no long stretches of predicted disorder and with R_h from DLS consistent with oligomerization.

This tree predicted test set outcome with a correlation of 0.56 (p<0.0014). With success defined as better than 10 Å diffraction, 87% were correctly predicted, Matthews correlation coefficient 0.67. For comparison, correlation for the best model with sequence alone was 0.18 (p>0.16); the highest Matthews correlation coefficient on our test set using previously reported sequence-only predictors was 0.48, with an accuracy of 60%.

ACKNOWLEDGEMENTS

We thank the many members of the SGPP consortium and MSGPP program project who contributed to these studies, supported by NIH grants GM64655 and AI067921. UW researchers include Easwara Subramanian, Christophe L.M.J. Verlinde and Frederick S. Buckner for target and domain selection; Chris Mehlin (SGPP), Alberto J. Napuli, Natascha Mueller, Lisa J. Castaneda, Stephen R. Nakazawa Hewitt and Wesley C. Van Voorhis (MSGPP) for protein production and characterization; Jaclyn dela Rosa, Jessica Kim, Li Zhang, Liren Xiao and Jenni Ross for further protein characterization and crystallization; Tracy L. Arakaki and Eric T. Larson for structure determination; and Erkang Fan for project management. We also thank Tina Veatch, Angela Lauricella, Joe Luft and George DeTitta at the Hauptman-Woodward MRI, Buffalo NY for high-throughput crystallization screening; and Thomas E. Kammeyer for assistance in interpreting raw DSF data files. NIH grant GM088518 provides additional funding for development of this predictor and for the PROSPERO crystallization prediction server, soon to be available at

skuld.bmsc.washington.edu/prospero