BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 24 2007, pages 3403–3405
doi:10.1093/bioinformatics/btm477

Structural bioinformatics

XtalPred: a web server for prediction of protein crystallizability

Lukasz Slabinski[1,2], Lukasz Jaroszewski[1], Leszek Rychlewski[2], Ian A. Wilson[1], Scott A. Lesley[1] and Adam Godzik[1,*]

[1]Joint Center for Structural Genomics, La Jolla, CA 92037, USA and [2]BioInfoBank Institute, ul. Limanowskiego 24 A, 60-744 Poznan, Poland

Received on July 17, 2007; revised on August 30, 2007; accepted on September 18, 2007

Advance Access publication October 5, 2007

Associate Editor: Thomas Lengauer

ABSTRACT

Summary: XtalPred is a web server for prediction of protein crystallizability. The prediction is made by comparing several features of the protein with distributions of these features in TargetDB and combining the results into an overall probability of crystallization. XtalPred provides: (1) a detailed comparison of the protein's features to the corresponding distribution from TargetDB; (2) a summary of protein features and predictions that indicate problems that are likely to be encountered during protein crystallization; (3) prediction of ligands; and (4) (optional) lists of close homologs from complete microbial genomes that are more likely to crystallize.

Availability: The XtalPred web server is freely available for academic users on http://ffas.burnham.org/XtalPred

Contact: adam@burnham.org

1 INTRODUCTION

The high failure rate in experimental determination of protein structures is still one of the biggest challenges of structural biology. Data from Structural Genomics (SG) centers show that the overall success rate in a high-throughput (HT) setup has only been around 5% and while no statistics are available for regular structural biology labs, anecdotal evidence suggests that the failure rate is also very high. Bioinformatics tools can aid in recognizing which proteins are more likely to succeed and provide suggestions of possible modifications for all the others. Selection of targets with the highest chance of success is especially useful for SG centers, targeting protein families rather than individual proteins.

The relation between proteins' features and their crystallizability has been investigated by several groups (Bertone *et al.*, 2001; Canaves *et al.*, 2004; Goh *et al.*, 2004; Oldfield *et al.*, 2005). However, traditional labs report only successes in structure determination, making data mining analyses almost impossible due to a lack of appropriately balanced data sets with positive and negative data. This situation changed with establishment of the Protein Structure Initiative (www.nigms.nih.gov/Initiatives/PSI), which requires its member centers to report both successes and failures to a central database, TargetDB (Chen *et al.*, 2004). Learning sets extracted from TargetDB have allowed more advanced analyses (Chandonia *et al.*, 2006; Overton and Barton, 2006; Smialowski *et al.*, 2006), which we expand here using data and insights stemming from work in the Joint Center for Structural Genomics (JCSG).

We have used the logarithmic opinion pool method (Genest *et al.*, 1984) to combine the probability distributions calculated for several individual protein features into a "crystallization feasibility score" (Slabinski *et al.*, 2007), where we demonstrated that our method can significantly improve the overall success rate in structure determination. Analysis of depositions in the PDB (Berman *et al.*, 2000) has confirmed that the same protein features also have substantial impact on success rates in standard, non-HT structure determination, suggesting that the "crystallization feasibility score" would also be of significant interest to a broad structural biology community. Since 2006, our algorithm has been used successfully at the JCSG to select optimal structure determination targets from protein families with no or inadequate structural coverage.

The XtalPred server builds on the statistical knowledge about protein crystallization gathered by the PSI over the past 7 years and makes the insights from the HT structure determination available to a broad community of structural biologists.

2 SERVER FEATURE SUMMARY

Crystallization analyses: the web server compares nine biochemical and biophysical features of the protein being analyzed with corresponding probability distributions from TargetDB. A plot is generated for each protein feature, showing distributions of failures and successes in the sets extracted from TargetDB; interpolated empirical distributions of crystallization probability; and the positions of the protein in those distributions (Fig. 1).

Crystallization prediction: the prediction is made by combining individual crystallization probabilities into a single crystallization score. Based on this score, the protein is assigned to one of five crystallization classes: optimal, suboptimal, average, difficult, and very difficult (Fig. 1).

Summary of information about the protein: the server calculates and predicts protein features that are related to protein crystallizability and summarizes them on one web page.

*To whom correspondence should be addressed.

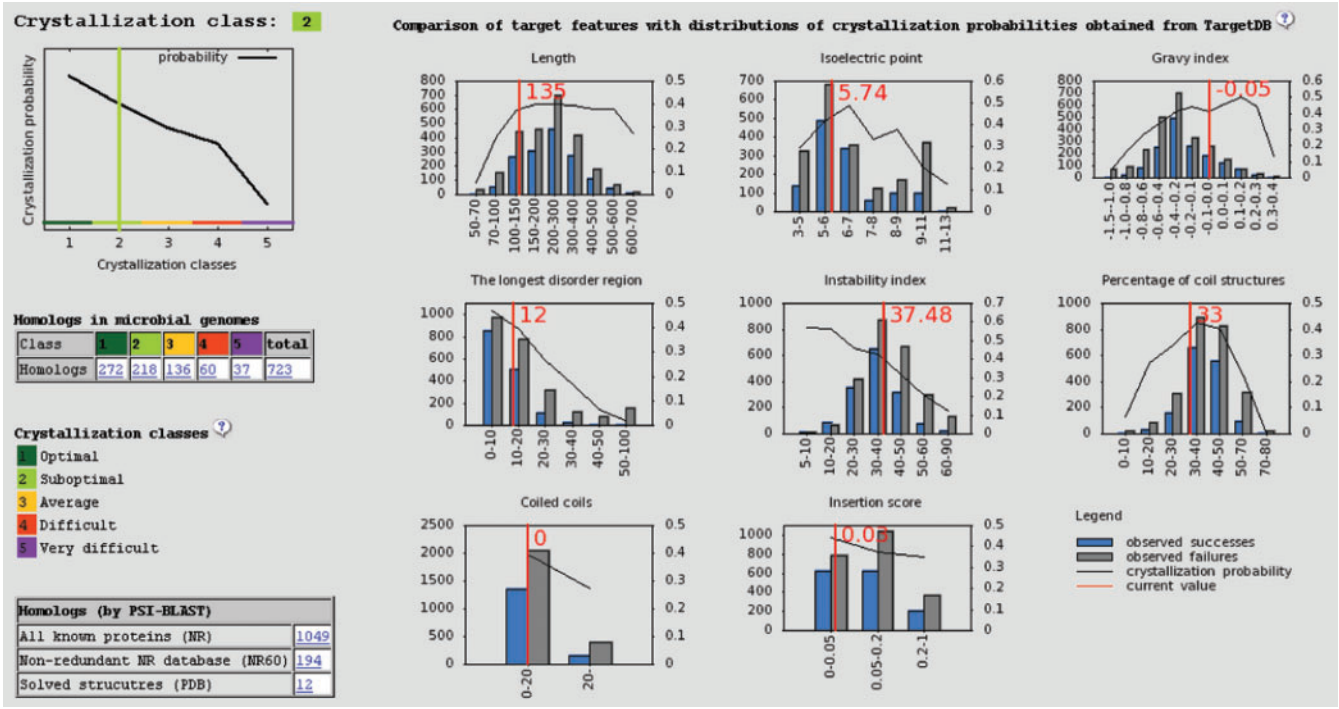

Fig. 1. Example of XtalPred output. The probabilities calculated from histograms obtained for individual protein features (right panel) are used to assign the protein to the appropriate crystallization class (the left upper corner). Links to lists of homologs found in different databases are located in the left lower corner.

Calculated protein features include: protein length; molecular mass; gravy index (Kyte and Doolittle, 1982); instability index (Guruprasad *et al*., 1990); extinction coefficient (Gill and von Hippel, 1989); isoelectric point (Creighton, 1984); content of Cys, Met, Trp, Tyr, and Phe residues; and average number of insertions in the alignment compared to homologs in non-redundant (NR) database of protein sequences. The predicted features include: secondary structure, disordered regions, low-complexity regions, coiled-coil regions, transmembrane helices, and signal peptides. The features that may indicate problems during the crystallization process are highlighted. In the case of predictions made by external software (Section 3), the raw output is available as text files.

Close homologs that are more likely to crystallize: precalculated crystallization class for all complete microbial genomes (currently 487 genomes; 1, 549, 504 proteins) are available from the server. For each submitted protein, the server provides a list of its homologs with the information about their crystallizability class. The list also contains links to detailed information about each homolog.

Fold and ligand prediction: XtalPred provides sequence alignment of the input protein with all homologous proteins in PDB. It also contains a list of ligands co-crystallized with homologous proteins and their secondary structure.

Scalability: the server can process up to 10 sequences in a single submission. Larger submissions should be discussed with a web server administrator.

Homologs: the server provides the alignment with homologs that can be used to propose truncations.

3 SERVER DETAILS

The XtalPred server uses several publicly available programs for calculation and prediction of protein features: PSI-BLAST for homology searches; CD-HIT (Li and Godzik, 2006) for clustering protein sequence databases; COILS (Lupas *et al*., 1991) for prediction of coiled-coil regions, TMHMM (Krogh *et al*., 2001) for prediction of transmembrane helices; RPSP (Plewczynski *et al*., 2007) for prediction of signal peptides, SEG (Wootton, 1994) for calculation of low-complexity regions; PSIPRED (Jones, 1999) for secondary structure prediction; and DISOPRED2 (Ward *et al*., 2004) for prediction of structurally disordered regions.

ACKNOWLEDGEMENTS

This work was supported by the NIH Protein Structure Initiative grants U54 GM074898 (JCSG) and P20 GM076221 (JCMM).

Conflict of Interest: none declared.

REFERENCES

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bertone,P. *et al.* (2001) SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res.*, **29**, 2884–2898.

Canaves,J.M. *et al*. (2004) Protein biophysical properties that correlate with crystallization success in Thermotoga maritima: maximum clustering strategy for structural genomics. *J. Mol. Biol.*, **344**, 977–991.

Chandonia,J.M. *et al*. (2006) Target selection and deselection at the Berkeley Structural Genomics Center. *Proteins*, **62**, 356–370.

Chen,L. *et al*. (2004) TargetDB: a target registration database for structural genomics projects. *Bioinformatics* (*Oxford, England*), **20**, 2860–2862.

Creighton,T.E. (1984) *Proteins: Structure and Molecular Properties*. W. H. Freeman and Co, New York.

Genest,C. *et al*. (1984) Aggregating opinions through logarithmic pooling. *Theor. Decis.*, **17**, 61–70.

Gill,S.C. and von Hippel,P.H. (1989) Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem.*, **182**, 319–326.

Goh,C.S. *et al*. (2004) Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. *J. Mol. Biol.*, **336**, 115–130.

Guruprasad,K. *et al*. (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng.*, **4**, 155–161.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Krogh,A. *et al*. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* (*Oxford, England*), **22**, 1658–1659.

Lupas,A. *et al*. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.

Oldfield,C.J. *et al*. (2005) Addressing the intrinsic disorder bottleneck in structural proteomics. *Proteins*, **59**, 444–453.

Overton,I.M. and Barton,G.J. (2006) A normalised scale for structural genomics target ranking: the OB-Score. *FEBS Lett.*, **580**, 4005–4009.

Plewczynski,D. *et al*. (2007) The RPSP: Web server for prediction of signal peptides. *Polymer*, **48**, 5493–5496.

Slabinski,L. *et al*. (2007) The challenge of protein structure determination – lessons from structural genomics. *Protein Sci.*, **16**, 2472–2482.

Smialowski,P. *et al*. (2006) Will my protein crystallize? A sequence-based predictor. *Proteins*, **62**, 343–355.

Ward,J.J. *et al*. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.

Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.*, **18**, 269–285.