Structural coverage using X-ray crystallography for a current snapshot of the protein universe

Marcin J Mizianty[1], Xiao Fan[1], Jing Yan[1], Eric Chalmers[1], Christopher Woloschuk[1], Andrzej Joachimiak[2] & Lukasz Kurgan[1,]*

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada and [2]Midwest Center for Structural Genomics, Biosciences Division, Argonne National Laboratory, USA

[*]lkurgan@ece.ualberta.ca

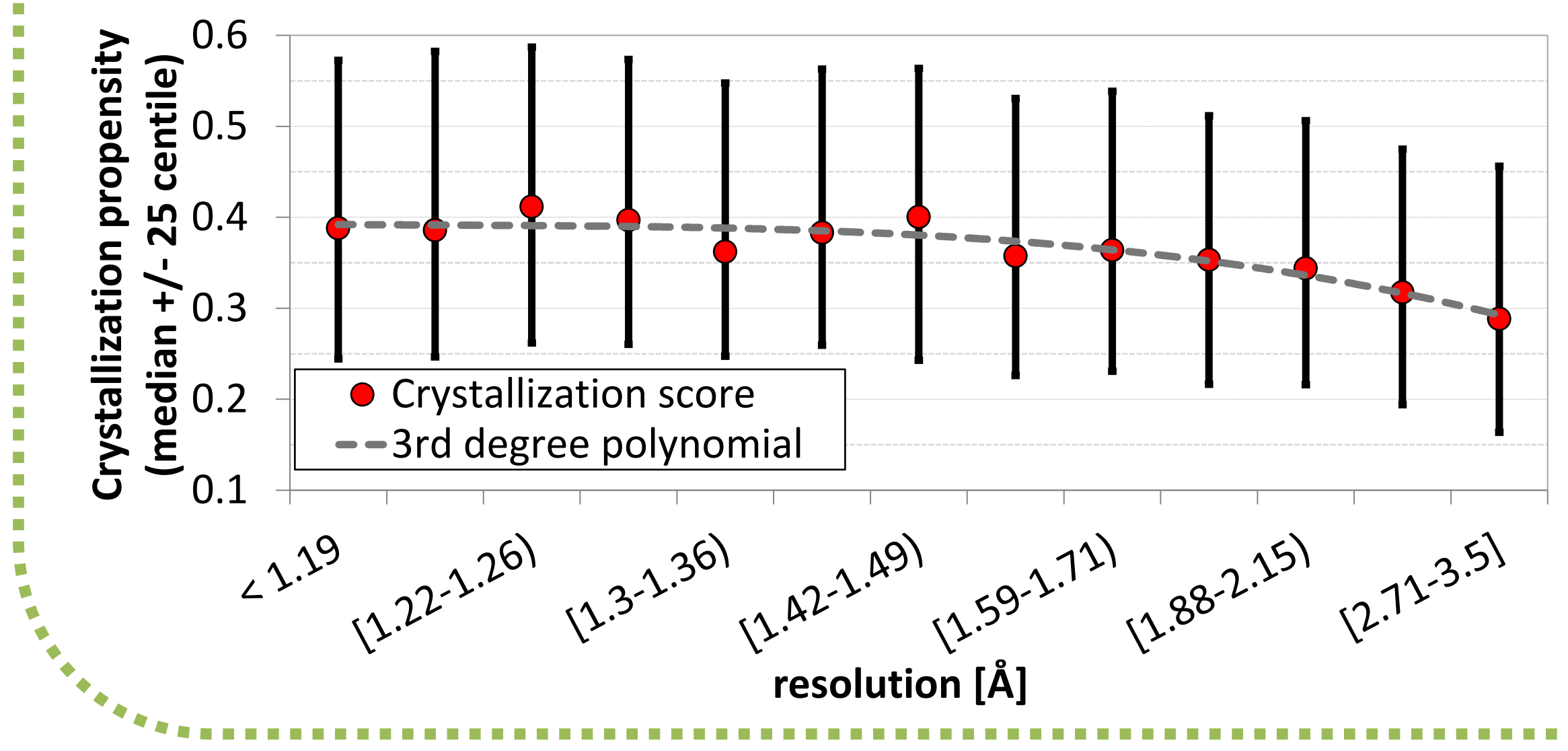
Introduction

*Can three-dimensional structures of all protein families be determined using X-ray crystallography? Utilizing an accurate and time-efficient method for **F**ast **DE**termination of **T**argets' **E**ligibility for **C**rys**T**alization (fDETECT), we performed a first-of-its-kind analysis of crystallization propensity for all proteins encoded in nearly 2,000 fully sequenced genomes. We computed the attainable structural coverage that combines use of current crystallization protocols, and homology modeling, by utilizing crystallization propensity scores for target selection.*

Materials & Methods

Empirical tests show that fDETECT predicts crystallization propensity with competitive predictive performance while being six orders of magnitude faster than similarly accurate methods. Our evaluation also shows an interesting trend between structure resolution and the predicted crystallization propensity score; higher score implies, on average, that the crystal resolution of the structure is better.

Method	Avg. time [ms]	AUC
fDETECT	0.8	**0.754**
PPCpred	152912.9	0.741
XtalPred	70624.4	0.665
CRYSTALP2	**0.3**	0.658
SVMcrys	153.3	0.615
OBScore	64	0.569
ParCrys	N/A	0.557



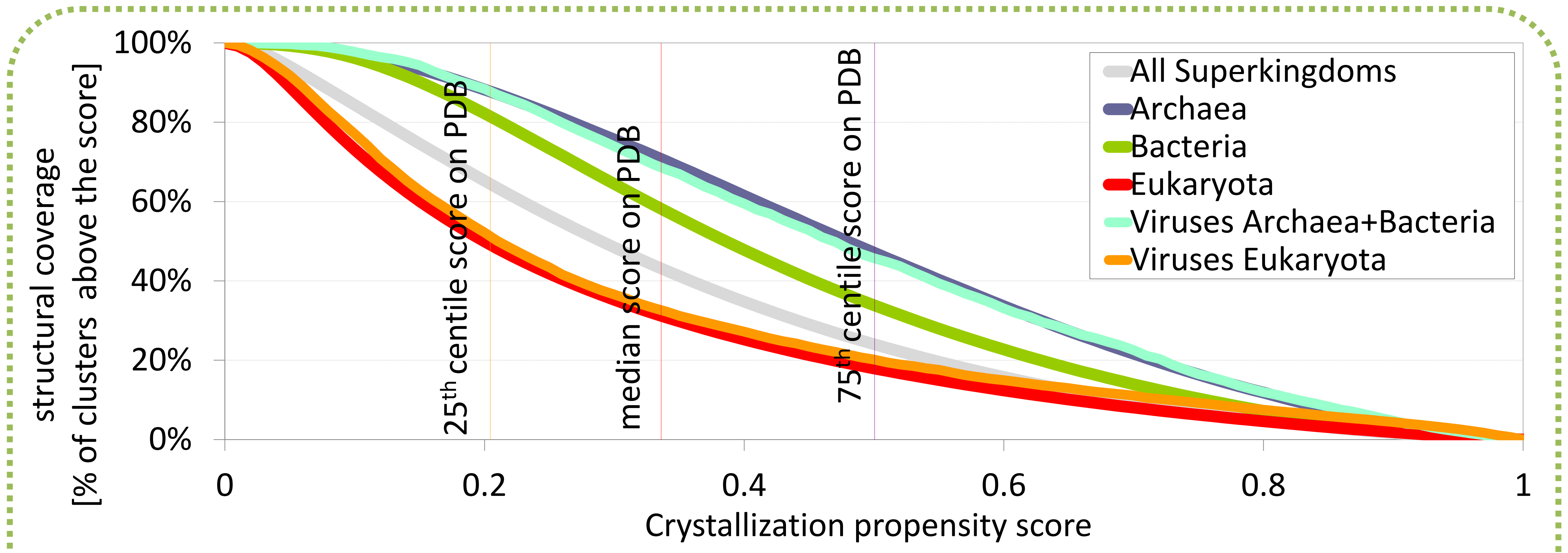
Utilizing fDETECT, we performed a first-of-its-kind analysis of predicted crystallization propensity of 8,652,940 non-redundant proteins from 1,953 fully sequenced proteomes (106 archaea, 1,043 bacteria, 265 eukaryotes and 539 viruses) collected from release 2012_07 of UniProt.

Results

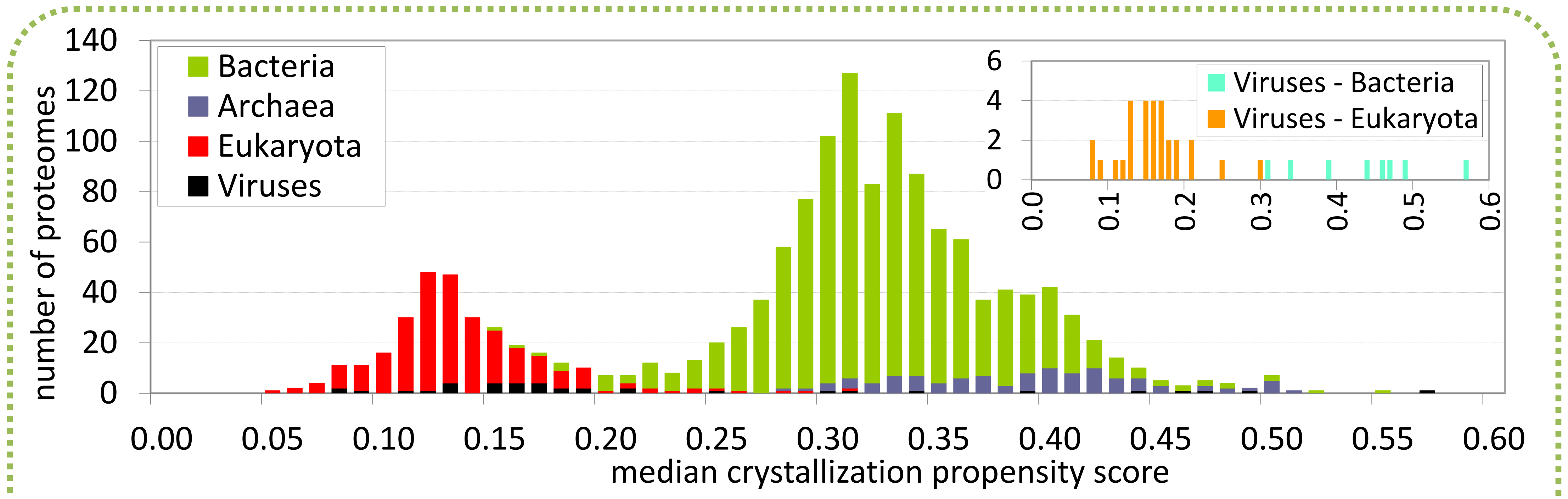
Analysis of the predicted crystallization propensity reveals that three superkingdoms show very different overall propensity for crystallization, with archaea being the easiest to crystallize and eukaryota the hardest. It appears that crystallization propensities of viral proteomes have properties similar to their host proteomes. We also show that current X-ray crystallography know-how combined with homology modeling (using 30% sequence identity cut off) can provide an average structural coverage of 73%. The structural coverage could be increased to 96% by improving homology modeling, if it would produce reliable predictions at 20% sequence identity. Moreover, the use of the knowledge-based target selection strategy significantly increases structural coverage.

Acknowledgments

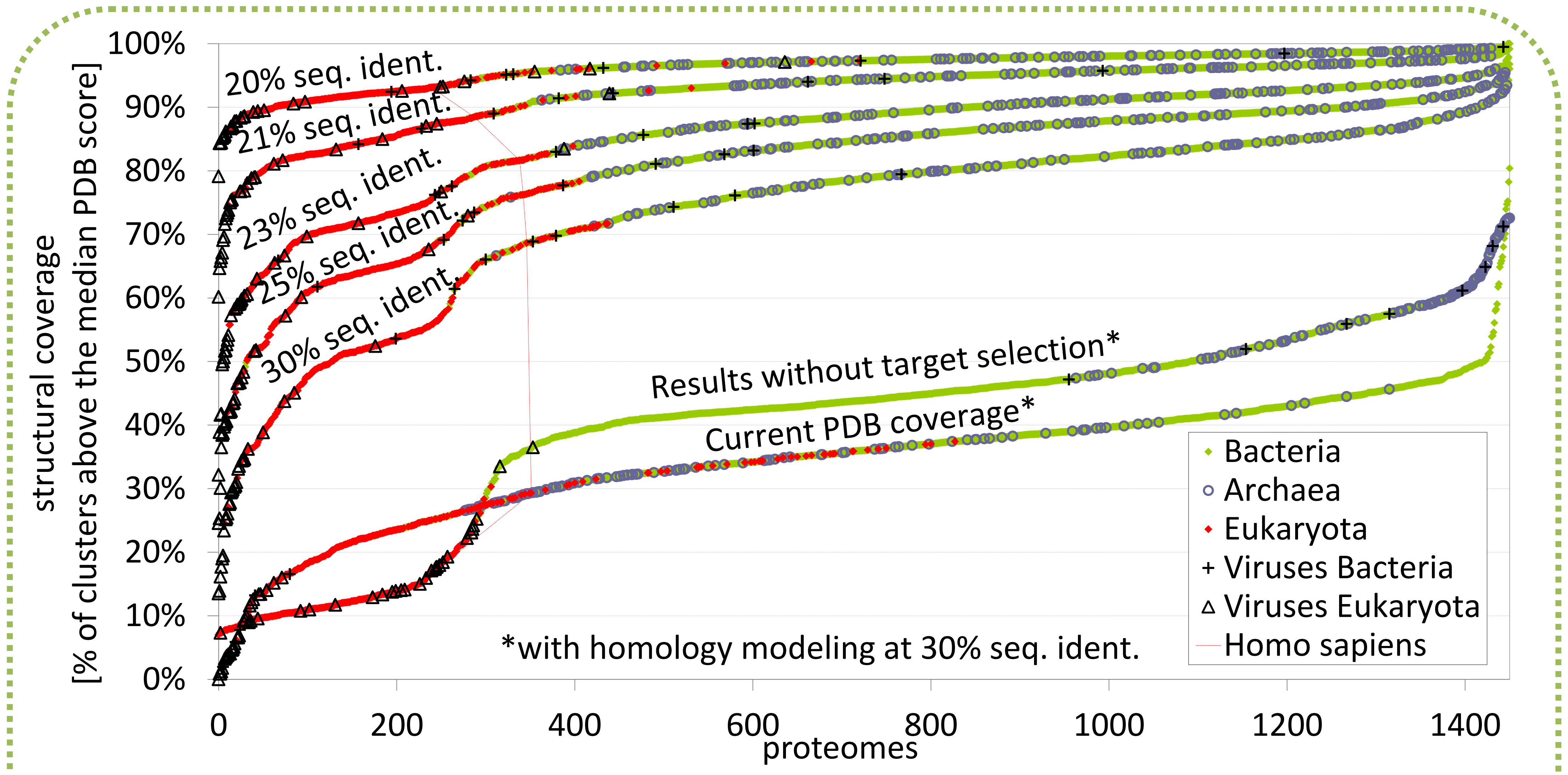
We would like to gratefully acknowledge the travel award from Genome Canada to M.J.M. This work was supported by the Dissertation fellowship awarded by the University of Alberta to M.J.M.; and by the Discovery grant from the NSERC of Canada to L.K.



There are 1,734,048 protein families clustered at 30% sequence identity level (modeling families), in the analyzed 1,953 proteomes. Each superkingdom show very different overall propensity for crystallization.



Bacterial proteomes have the widest distribution of crystallization scores that overlap with scores for archaea proteomes and, to a smaller extend, with eukaryotes. Propensities for archaea and eukaryotes show virtually no overlap.



Using homology modeling at 30% sequence identity cut-off virtually all bacterial and archaeal proteomes as well as bacteriophages can be structurally covered at above 70%. However, majority of the eukaryotes and eukaryotic viruses have coverage well below 70%.