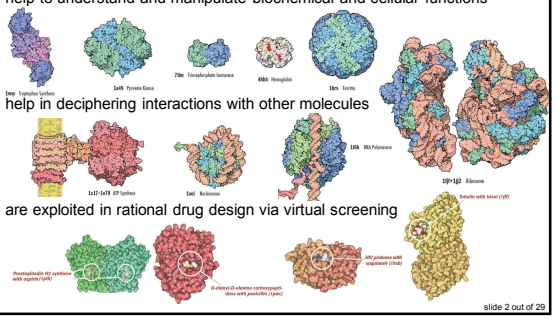
Characterization of propensity for X-ray crystallography of protein chains

Lukasz Kurgan and Marcin Mizianty
University of Alberta, Canada

Andrzej Joachimiak
Midwest Center for Structural Genomics

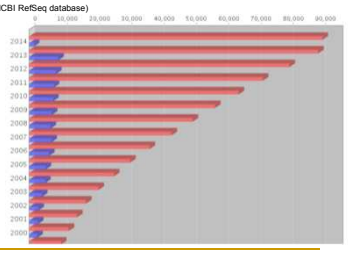
Introduction

- Protein structures
  - help to understand and manipulate biochemical and cellular functions
  - help in deciphering interactions with other molecules
  - are exploited in rational drug design via virtual screening



Introduction

- Protein structures
  - are known for 90,738 proteins (source: PDB database) out of 37,371,278 (source: NCBI RefSeq database) known unique proteins sequences
  - 88.5% of structures were solved using X-ray crystallography



Introduction

- Structural genomics (SG) is a word-wide initiative aimed at mapping of entire protein structure space
  - in 2004/2005 about ½ structures were solved at SG centers rather than in a traditional lab at about 25% of the cost
  - SG shifts the focus from one-by-one determination of individual structures to protein family-directed structure analyses in which a group of proteins is targeted and structure(s) of representative members are determined
    - selection of representative proteins is known as target selection

Brenner SE. *Nature Structural Biology* 2000; 7:967-969
Chandonia JM, Brenner SE. *Science* 2006; 311:347-351
Dessailly BH et al. *Structure* 2009; 17(6):869-881

Introduction

PNAS

Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative

Kamil Khafizov[a,b,c,d,1], Carlos Madrid-Aliste[a,b,c,d], Steven C. Almo[b,c,d,e], and Andras Fiser[a,b,c,d,2]

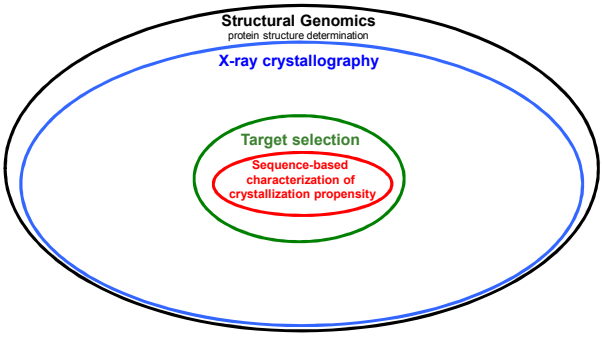
[a]Department of Systems and Computational Biology, [b]Department of Biochemistry, [c]New York Structural Genomics Research Consortium, [d]Immune Function Network, and [e]Department of Physiology and Biophysics, Albert Einstein College of Medicine, Bronx, NY 10461

Edited by Barry Honig, Columbia University and Howard Hughes Medical Institute, New York, NY, and approved January 29, 2014 (received for review November 21, 2013)

The exponential growth of protein sequence data provides an ever-expanding body of unannotated and misannotated proteins. The National Institutes of Health-supported Protein Structure Initiative and related worldwide structural genomics efforts facilitate functional annotation of proteins through structural characterization. Recently there have been profound changes in the taxonomic composition of sequence databases, which are effectively redefining the scope and contribution of these large-scale structure-based efforts. The faster-growing bacterial genomic entries have overtaken the eukaryotic entries over the last 5 y, but also have become more redundant. Despite the enormous increase in the number of sequences, the overall structural coverage of proteins—including proteins for which reliable homology models can be generated—on the residue level has increased from 30% to 40% over the last 10 y. Structural genomics efforts contributed ~50% of this new structural coverage, despite determining only ~10% of all new structures. Based on current trends, it is expected that ~55% to work in unexplored areas owing to a lack of compelling hypotheses. Without discovery-driven efforts, large amounts of biology will continue to remain obscure, and it will not be possible to systemically examine the structural features of full proteomes, which undoubtedly would lead to unexpected functional and biological insights (7–9).

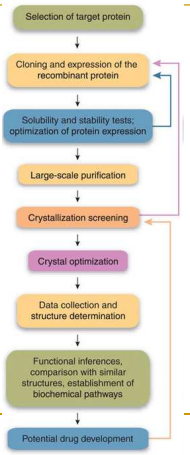
The Protein Structure Initiative (PSI), supported by the US National Institutes of General Medical Sciences, was established in 2000 and is the largest ongoing coordinated effort in the field of structural biology. The PSI has evolved through three phases (10). The first phase (PSI-1; 2000–2005) demonstrated the feasibility of HTP cloning, protein expression, purification, and structure determination. The implementation of this infrastructure was realized and applied during the production phase (PSI-2; 2005–2010) to significantly expand our knowledge of sequence–structure relationships and to complement efforts in computational biology, such as homology modeling (11), as well as to address specific bottlenecks, such as those associated with membrane

Introduction



Introduction

- Pipeline for structure determination using X-ray crystallography



Chayen NE, Saridakis E. *Nature Methods* 2008; 5:147-153

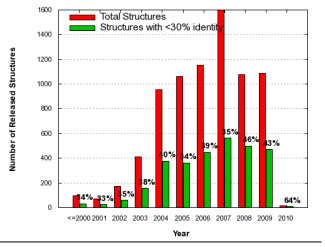
Motivation

- The main challenge of the SG initiative is that only about 2-10% of protein targets pursued yield high-resolution protein structures
  - one of the most important bottlenecks in acquiring the structures is obtaining diffraction-quality crystals
    - crystal should be sufficiently large (> 100 micrometres), pure in composition, regular in structure, and with no significant internal imperfections

Hui R, Edwards A. *J Struct Biol.* 2003; 142:154-61
Rupp B, Wang JW. *Methods* 2004; 34:391–408
Service R. *Science* 2005; 307:1554-1558

Motivation

- TargetDB
  http://targetdb.pdb.org/
  data as of January 2010



Status	# of targets	% of cloned	% of expressed	% of purified	% of crystallized
Cloned	163639	100.0	-	-	-
Expressed	117920	72.1	100.0	-	-
Soluble	45629	27.9	38.7	-	-
Purified	41815	25.6	35.5	100.0	-
Crystallized	14250	8.7	12.1	34.1	100.0
Diffraction-quality crystals	**7504**	**4.6**	**6.4**	**17.9**	**52.7**

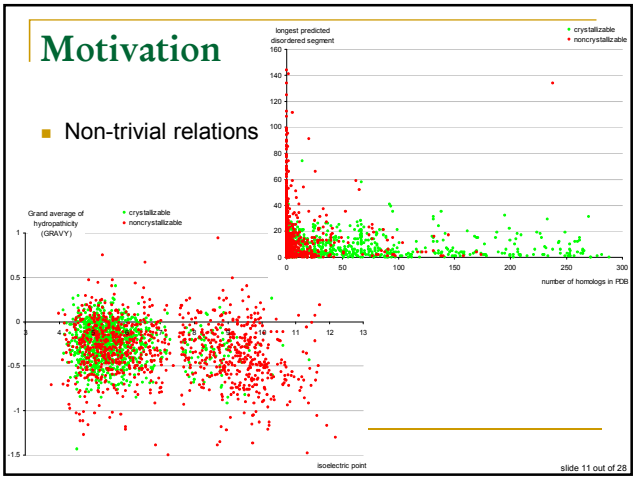
Motivation

- Information derived from protein sequences can be used to predict crystallization propensity
  - conservation of the sequence across organisms
  - inclusion of charged amino acids
  - occurrence of hydrophobic patches
  - presence of transmembrane helices, signal peptides, low-complexity regions, disordered and coiled-coil regions
  - presence of certain amino acids on the protein surface
  - isoelectric point (pI) is used to suggests optimal pH ranges for crystallization screening
  - presence of homologs in PDB

Canaves JM, et al. *J. Mol. Biol.* 2004; 344:977–991
Goh CS, et al. *J. Mol. Biol.* 2004; 336:115–130
Chandonia JM et al. *Proteins* 2006; 62:356-370
Price WN et al. *Nature Biotechnology* 2009; 27(1): 51-57

Motivation

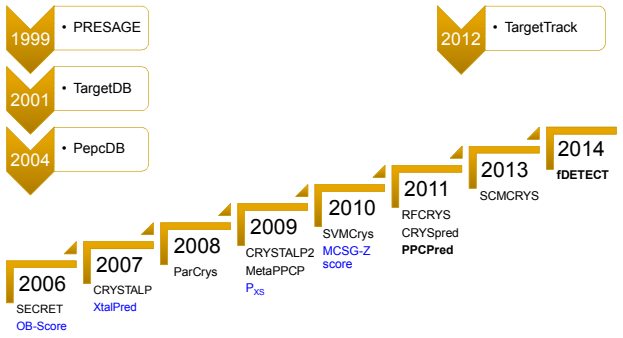
- Non-trivial relations



Problem definition

- To develop an accurate sequence-based in-silico predictor of propensity to yield diffraction quality crystals
  - limitations
    - we take into account only intra-molecular factors that are encoded in the protein chain
    - we may not provide reliable predictions when inter-molecular factors such as protein-protein and/or protein-precipitant interactions, buffer composition, precipitant diffusion method, gravity, etc. must be considered
    - we assume that physical considerations of the crystal growth procedure, purification, expression, etc., will be properly handled

Overview of the field



PPCpred

- Built using a recent and large dataset
- Uses improved annotation protocol
  - in collaboration with curators of PepcDB: Drs Berman & Westbrook
- Predicts success of the entire crystallization process and also which step(s) results in the failed attempts
- Uses a compact and comprehensive set of sequence-derived inputs to generate accurate predictions

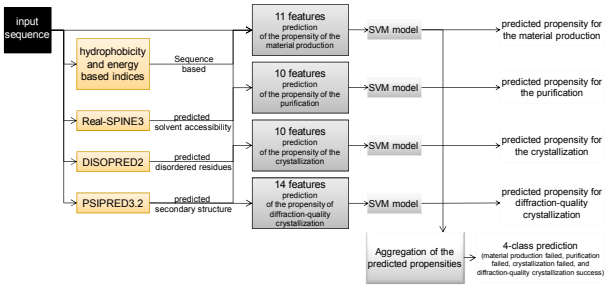
Mizianty MJ, Kurgan L. *Bioinformatics* 2011; 27(13):i24-33

PPCpred

Outcome deduced from PepcDB annotations	Stop status	Current status
Production of protein material failed	sequencing failed cloning failed	cloned
	expression failed	expressed
Purification failed	purification failed	soluble
		purified
Crystallization failed	crystallization failed	crystallized
	poor diffraction	diffraction-quality crystals
		Diffraction (native diffraction-data or phasing diffraction-data)
Crystallizable	structure successful, TargetDB duplicate target found, PDB duplicate found	crystal structure
		in PDB

The current status indicates the current activity, e.g. for the "cloning failed" stop status, the current status "cloned" does not mean that cloning was successful, but if the current status is "expressed" then cloning can be assumed successful.

PPCpred



Mizianty MJ, Kurgan L. *Bioinformatics* 2011; 27(13):i24-33

PPCpred

- Feature types selected for the prediction out of 800+ considered features

Features types	Number of features selected for the prediction of			
	Material production	Purification	Crystallization	Diffraction-quality crystallization
Hydrophobicity-based	2	2	5	5
Energy-based	4	0	2	3
Composition of AAs	1	3	1	1
Isoelectric point	0	1	0	0
Solvent accessibility	3	4	1	3
Disorder	1	0	1	1
Secondary structure	0	0	0	1
Considered AA types	Arginine, Cysteine, Glutamic acid	Asparagine, Cysteine, Serine, Methionine	Histidine	Cysteine, Histidine, Serine

Mizianty MJ, Kurgan L. *Bioinformatics* 2011; 27(13):i24-33

PPCpred

- Prediction of propensity of the diffraction-quality crystallization success

Predictors	MCC		ACC		SPEC	SENS	AUC
	value	sig	value	sig			
ParCrys	0.108	+	47.5	+	31.8	78.6	0.561
OBScore	0.124	+	47.8	+	31.4	**80.3**	0.572
BLAST-based	0.188	+	65.6	+	79.5	38.0	N/A
CRYSTALP2	0.195	+	55.3	+	45.7	74.4	0.648
MetaPPCP	0.195	+	59.9	+	59.0	61.7	0.620
SVMCrys	0.213	+	56.3	+	46.7	75.2	N/A
XtalPred	0.278	+	63.9	+	62.3	67.0	0.683
PPCpred	**0.471**		**76.8**		**84.8**	61.2	**0.789**

Mizianty MJ, Kurgan L. *Bioinformatics* 2011; 27(13):i24-33

PPCpred

- Prediction of the four steps of crystallization pipeline

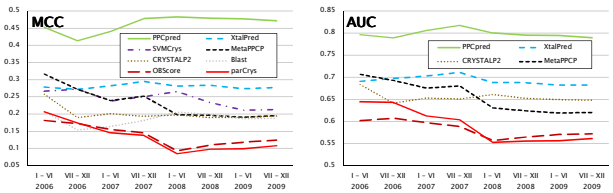
Prediction target	Method	MCC		ACC		SPEC	SENS	AUC
		value	sig	value	sig			
propensity of the diffraction-quality crystallization success	BLAST-based	0.188	+	65.6	+	79.5	38.0	N/A
	PPCpred	**0.471**		**76.8**		**84.8**	**61.2**	0.789
propensity of the material production failure	BLAST-based	0.014	+	55.4	+	35.3	66.0	N/A
	PPCpred	**0.462**		**75.0**		**69.2**	**78.0**	0.755
propensity of the purification failure	BLAST-based	0.102	+	60.0	+	43.2	67.4	N/A
	PPCpred	**0.324**		**72.0**		**50.1**	**81.6**	0.697
propensity of the crystallization failure	BLAST-based	0.060	+	60.9	+	37.0	69.4	N/A
	PPCpred	**0.457**		**76.6**		**70.8**	**78.7**	0.811

Predictor	Mean MCC		ACC	
	Value	sig	Value	sig
BLAST-based	0.041	+	31.1	+
PPCpred	**0.353**		**55.6**	

Mizianty MJ, Kurgan L. *Bioinformatics* 2011; 27(13):i24-33

PPCpred

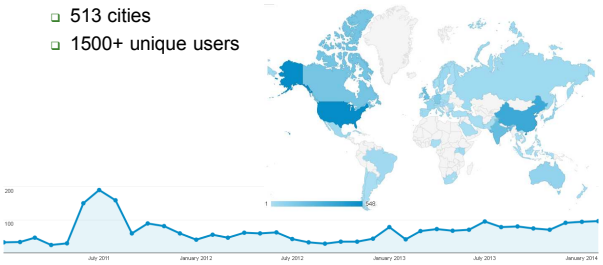
- Prediction performance over dates of trials



Mizianty MJ, Kurgan L. *Bioinformatics* 2011; 27(13):i24-33

PPCpred

- User base
  - 61 countries (top 10: US, China, India, UK, Canada, Garmany, Australia, France, Taiwan, and Japan)
  - 513 cities
  - 1500+ unique users



slide 21 out of 28

http://biomine.ece.ualberta.ca/PPCpred/

PPCPRED WEBSERVER

The server is designed for sequence-based prediction of protein crystallization, purification, and production propensity.

1. Enter protein sequence(s)

Please enter each protein in a new line (FASTA FORMAT) - up to 5 proteins allowed

```
>ttk003002175.1:0:0
MVNPAERLAELDGVLMQYLLEADLLRELPPTYRLVLLPLDEPEVAAQALAWAMEAPNPEGWPSVYALFLQGRPIRLLLLGKE
VEVAPRAA
```

Example | Reset sequence(s)

2. Enter your e-mail address (required)

Please provide your e-mail address: mizianty@ualberta.ca

3. Start

Start | Reload page

References

Upon the usage the users are requested to use the following citations:

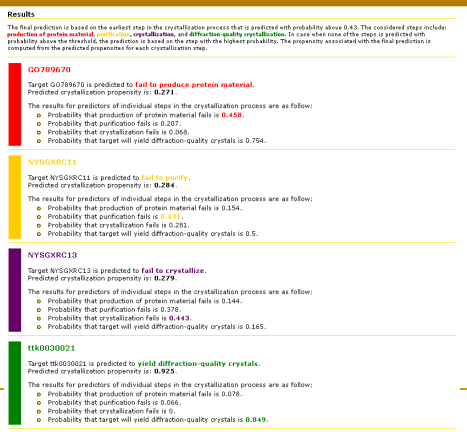
- MIZIANTY MJ, KURGAN L. SEQUENCE-BASED PREDICTION OF PROTEIN CRYSTALLIZATION, PURIFICATION, AND PRODUCTION PROPENSITY. *BIOINFORMATICS*, 27(13):i24-i33

Additional materials

- Training dataset, used to train PPCpred. Dataset was developed in our lab and can be downloaded from HERE
- Test dataset, used to evaluate PPCpred. Dataset was developed in our lab and can be downloaded from HERE
- Supplement can be downloaded from HERE

slide 22 out of 28

http://biomine.ece.ualberta.ca/PPCpred/



Results

The final prediction is based on the earliest step in the crystallization process that is predicted with probability above 0.43. The considered steps include: production of protein material, purification, crystallization, and diffraction-quality crystallization. In case when none of the steps is predicted with probability above the threshold, the prediction is based on the step with the highest probability. The propensity associated with the final prediction is computed from the predicted propensities for each crystallization step.

GO789670

Target GO789670 is predicted to **fail to produce protein material**.
Predicted crystallization propensity is: **0.271**.

The results for predictors of individual steps in the crystallization process are as follow:
- Probability that production of protein material fails is **0.458**.
- Probability that purification fails is 0.207.
- Probability that crystallization fails is 0.068.
- Probability that target will yield diffraction-quality crystals is 0.754.

NYSGXRC11

Target NYSGXRC11 is predicted to **fail to purify**.
Predicted crystallization propensity is: **0.284**.

The results for predictors of individual steps in the crystallization process are as follow:
- Probability that production of protein material fails is 0.154.
- Probability that purification fails is **0.431**.
- Probability that crystallization fails is 0.281.
- Probability that target will yield diffraction-quality crystals is 0.5.

NYSGXRC13

Target NYSGXRC13 is predicted to **fail to crystallize**.
Predicted crystallization propensity is: **0.279**.

The results for predictors of individual steps in the crystallization process are as follow:
- Probability that production of protein material fails is 0.144.
- Probability that purification fails is 0.378.
- Probability that crystallization fails is **0.443**.
- Probability that target will yield diffraction-quality crystals is 0.165.

ttk0030021

Target ttk0030021 is predicted to **yield diffraction-quality crystals**.
Predicted crystallization propensity is: **0.925**.

The results for predictors of individual steps in the crystallization process are as follow:
- Probability that production of protein material fails is 0.078.
- Probability that purification fails is 0.066.
- Probability that crystallization fails is 0.
- Probability that target will yield diffraction-quality crystals is **0.849**.

slide 23 out of 28

Structural coverage using X-ray crystallography

- Aim
  - investigate attainable structural coverage considering current X-ray crystallography combined with homology modeling
- Setup
  - 1,953 fully sequenced proteomes collected from release 2012_07 of UniProt
    - 106 archaea, 1,043 bacterias, 265 eukaryotes and 539 viruses
    - 8,652,940 non-redundant proteins

slide 24 out of 29

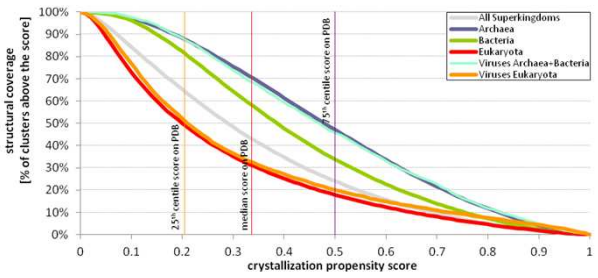
Structural coverage using X-ray crystallography

- fDETECT

Method	Runtime per protein [ms]		ACC		MCC		SPEC	SENS	AUC	
	avg	sig	value	sig	value	sig	value	value	value	sig
fDETECT	0.8		70.6		0.354		75.8	60.3	**0.754**	
PPCpred	152912.9	+	**71.8**	-	**0.361**	-	**79.7**	56.0	0.741	+
XtalPred*	70624.4	+	53.3	+	0.248	+	36.0	**87.6**	0.665	+
CRYSTALP2	**0.3**	-	56.6	+	0.202	+	48.5	72.6	0.658	+
SVMcrys	153.3	+	56.5	+	0.223	+	46.5	76.5	0.615	+
OBScore	64	+	47.2	+	0.130	+	29.3	82.7	0.569	+
ParCrys**	N/A	N/A	48.3	+	0.105	+	34.5	75.9	0.557	+

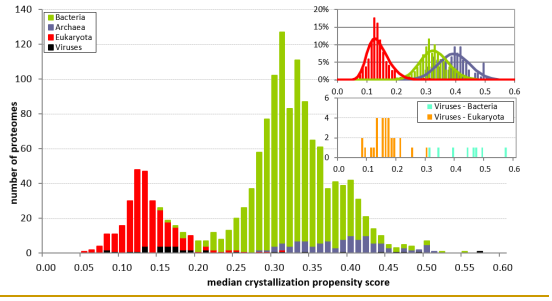
* XtalPred results were obtained from a webserver, the runtime estimates may be inaccurate
**ParCrys is available as webserver and we could not estimate its runtime

Structural coverage using X-ray crystallography

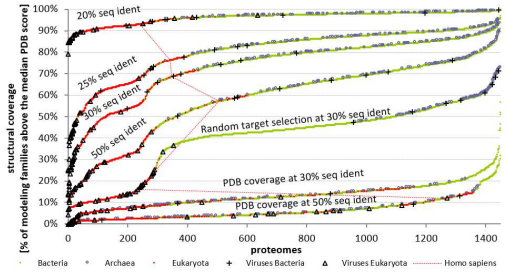


1,734,048 (modeling) protein families clustered at 30% sequence identity level

Structural coverage using X-ray crystallography



Structural coverage using X-ray crystallography



Summary

- Crystallization propensity predictors provide useful input for target selection
  - PPCpred targets several steps in the crystallization pipeline
  - fDETECT offers fast predictions
- Use of the knowledge-based target selection strategy substantially increases structural coverage
- Current X-ray crystallography know-how combined with homology modeling (30% sequence identity cutoff) can provide an average structural coverage of 73%
  - coverage could be increased to 96% by improving homology modeling (assuming 20% sequence identity cutoff)

slide 29 out of 29