Bioinformatics at the Center for Eukaryotic Structural Genomics (CESG)

Craig S. Newman, Zhaohui Sun, Chris Oldfield, Ryan Bannen, Ip Kei Sam, Kaushik Raghunath, Frank C. Vojtik,
Zsolt Zolnai, Brian G. Fox, George N. Phillips Jr., John L. Markley, and Eldon L. Ulrich

University of Wisconsin-Madison, 433 Babcock Drive, Madison, Wisconsin, USA 53706-1549, http://www.uwstructuralgenomics.org

Abstract

CESG is one of nine pilot centers supported by the NIH Protein Structure Initiative (PSI). The aim of CESG is to expand our understanding of fold space and to elucidate novel structure-function relationships by solving 3D structures of a large number of individual proteins in a high throughput manner. Projects such as this generate tremendous amounts of data and require the support of strong bioinformatics and data management efforts. The Bioinformatics Team is charged with evaluating potential targets on a genome-wide basis and with prioritizing the predicted ORFs on the basis a variety of parameters. We present our scoring protocol and illustrate its use in analyzing data. The Bioinformatics Team also harvests data from the Sesame laboratory information management system and other sources to analyze progress, identify bottlenecks, and uncover factors that influence the success of individual steps in the pipeline.

CESG target scoring process

Potential targets are scored on the basis of a variety of quantitative and predicted characteristics identified by software packages developed in house, by collaborators or licensed from academic sources. Characteristics (and the relevant tools) used to score targets include:

- homology to proteins in the PDB and in the TargetDB that are at or nearing structure completion (BLAST)
- presence of target "fragments" which lack homology to proteins in the PDB and in the TargetDB (BLAST)
- coiled coil domains (COILS)
- transmembrane segments (SCOP classification, TMHMM, and HMMTOP)
- signal peptides (SignalP and TargetP)
- number of Cys residues
- percent low complexity (seg)
- microarray analysis of *in vivo* transcription (where available)
- predicted disorder (PONDR)
- similarity to transposable elements (Censor)
- general suitability (on the basis of qualitative annotations, literature searches and knowledge and experience from within the CESG team)

Based on the results of these analyses, each ORF is allocated a 14 digit "target score" (see Figure 1) . In each category, low digits are considered to be desirable. These scores are then used to divide the ORFs into priority tiers. An additional digit, representing the overall tier into which the ORF has been placed, is then added to the beginning of the target score. ORFs in the first, second, or third tiers are considered prime targets.

To date, 69,886 unique targets have been evaluated with the CESG target scoring protocol. This includes the entire *Arabidopsis thaliana* genome (version 5.0), non-redundant clones from the Mammalian Gene Collection (including sequences from human, mouse, rat, zebrafish, and *Xenopus*) and a number of community requested targets. As the new proteome predictions become available, additional species can easily be analyzed and prioritized for entry in to the CESG protein production pipelines.

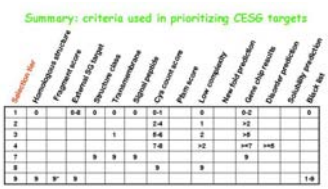

Figure 1. Each scoring criteria (columns 2 through 14) is given a score ranging from 0 to 9. Low scores are considered to be desirable. The overall tier score is limited by the individual criteria scores. The "Pfam Score", "New Fold Prediction" and "Solubility Prediction" categories are not currently in use. Gene chip results are currently only available for targets from *Arabidopsis*. For examples of target scores, see Table 1.

Table 1. Target Scores for the Most Recent *E. coli* Pipeline Produced Structures

Gene ID	Source Organism	Target Score	Selection Tier	Homologous Structure	Fragment Score	Ethered GO Target	Structure Class	Transmembrane Segments	Signal Peptide	Cys Count	Pfam Score	Low Complexity	New Fold Prediction	Gene Chip Results	Disorder Prediction	Solubility Prediction	Block List
BC026994	Mouse	1097000100900000	1	0	9	7	0	0	1	0	0	9	0	0	0	0	
At5g56450	Arabidopsis	2090000100930000	2	0	9	0	0	0	1	0	0	9	3	0	(0)	0	
At5g02240	Arabidopsis	2090000100930000	2	0	9	0	0	0	1	0	0	9	3	0	(0)	0	
At3g03773	Arabidopsis	2090000300930000	2	0	9	0	0	0	3	0	(0)	9	3	0	(0)	0	
At3g22680	Arabidopsis	3090000500960000	3	0	9	0	0	0	5	0	(0)	9	3	0	0	0	
At5g08170	Arabidopsis	3090000600930000	3	0	9	0	0	0	6	0	(0)	9	3	0	0	0	
At1g05000	Arabidopsis	3090000600940000	3	0	9	0	0	0	6	0	(0)	9	4	0	0	0	
At2g34160	Arabidopsis	3090010001954000	3	0	9	0	1	0	(0)	0	1	9	3	4	0	0	
At1g01470	Arabidopsis	7090090100930000	7	0	9	0	0	9	0	(1)	0	0	9	3	0	0	
At1g77540	Arabidopsis	9099000209930000	9	0	9	9	0	(0)	0	2	0	9	9	3	0	0	
At2g17340	Arabidopsis	9099010500940000	9	0	9	9	0	(0)	0	5	0	0	9	4	0	0	
At1g24000	Arabidopsis	9990000100970000	9	9	9	(9)	0	0	0	1	0	0	9	7	0	0	
At5g18200	Arabidopsis	9990000600940000	9	9	9	(9)	0	0	0	6	0	1	9	4	0	0	
At5g56660	Arabidopsis	9990000900919000	9	9	9	(9)	0	0	9	2	0	1	9	3	0	0	
At3g04780	Arabidopsis	9990010900934000	9	9	9	(9)	0	1	9	9	0	0	9	3	4	0	0

Table 2 uses the final target scores for the most recent CESG structures to illustrate our target scoring protocol. For each target in a lower tier (2 through 9), the criterion that determined the overall Selection Tier is highlighted, in some cases, two different criteria relegate a target to the same lower tier. Note that scores are updated routinely and that the target score at the time of selection may not be the same as the score at the time of structure completion.

Current scoring criteria

Some of the key CESG scoring criteria are as follows:

Selection Tier represents the overall ranking of the target as determined by the values in the remaining 14 categories

Homologous Structure reflects homology to a previously solved structure deposited in the Protein Data Bank (PDB)

- **0** - expectation value greater than E-6
- **9** - expectation value less than E-6 and percent identity >29%

Fragment Score identifies targets with significant stretches of sequence not homologous to a known structure

- **0** - percent identity <29% over at least 100 residues against the PDB and an expectation value greater than E-3 -- with this score, the Homologous Structure score is ignored in determining the overall Selection Tier

Transmembrane reflects the number of predicted transmembrane segments

- **0** - no predicted segments or no data
- **1** - one predicted segment
- **9** - more than one predicted segment

Cysteine Count reflects the number of cysteine residues present in a protein

- **0** - 0 cysteine residues
- **1** - 1 cysteine residue
- **2** - 2 cysteine residues
- **3** - 3 cysteine residues
- **4** - 4 cysteine residues
- **5** - 5 or 6 cysteine residues
- **6** - 7 or 8 cysteine residues
- **7** - 9 or 10 cysteine residues
- **9** - greater than 10 cysteine residues

Low Complexity reflects the amino acid complexity of the target

- **0** - less than 5% of the target is determined to be of low complexity
- **1** - between 5% and 10% of the target is determined to be of low complexity
- **2** - between 10% and 15% of the target is determined to be of low complexity
- **3** - between 15% and 20% of the target is determined to be of low complexity
- **9** - greater than 20% of the target is determined to be of low complexity

Intrinsic Protein Disorder Prediction reflects the predicted intrinsic disorder for a target

- **0** - less than 30% disorder predicted
- **4** - between 30% and 35% disorder predicted
- **5** - between 35% and 40% disorder predicted
- **6** - between 40% and 45% disorder predicted
- **7** - between 45% and 50% disorder predicted
- **9** - greater than 50% disorder predicted or greater than 30% disorder predicted and the longest disordered segment greater than or equal to 40 residues

Analysis of experimental results

To evaluate the efficacy of the CESG score scheme, results from the CESG *E. coli* protein production pipeline have been examined. Presented below is a comparison of data from the protein production and structure determination pipeline versus key characteristics used in generating target scores. Our analysis shows that the CESG target score can predict the potential success of a target and that the success of the protein purification step seems to be most reflective of the target score. Table 2 shows that targets from the top three tiers yield 10 mg or more of purified protein at almost twice the frequency of targets in tiers 4, 7, or 8. A number of the specific criteria used to assign the Selection Tier seem to impact the purification success rate (Table 3). The cysteine count of a protein, number of transmembrane segments, amino acid complexity and intrinsic disorder all seem to have an effect on the ability to recover sufficient amounts of purified protein. Success at other steps of the protein production and structure determination pipeline (*i.e.*, cloning, protein expression, NMR spectroscopy, and X-ray crystallography) show less correlation with target score. However, some key characteristics do appear to predict success at specific stages. The intrinsic protein disorder score of a target correlates well with the production of a folded protein for NMR spectroscopy (see below for a more detailed discussion) while an increasing number of transmembrane domains in a target adversely affects cloning efficiency and protein expression (Tables 4 and 5).

Table 2. Selection Tier Analysis for the Protein Purification Step

selection tier	total	purified >10mg	% purified >10mg	purified between 0 and 10 mg	% purified between 0 and 10 mg	failed purification	% failed purification
1	26	14	53.8%	4	15.4%	8	30.8%
2	181	78	43.1%	20	11.0%	83	45.9%
3	189	67	35.4%	31	16.4%	91	48.1%
4	112	23	20.5%	10	8.9%	79	70.5%
7	95	19	20.0%	9	9.5%	67	70.5%
8	23	5	21.7%	2	8.7%	16	69.6%
9	103	41	39.8%	14	13.6%	48	46.6%
total	729	247	33.9%	90	12.3%	392	53.8%

Table 3. Protein Purification Data

Cys score	total	purified >10 mg	% purified >10 mg	purified between 0 and 10 mg	% purified between 0 and 10 mg	failed purification	% failed purification
0	80	28	35.0%	4	5.0%	48	60.0%
1	90	33	36.7%	11	12.2%	46	51.1%
2	98	36	36.7%	17	17.3%	45	45.9%
3	97	36	37.1%	11	11.3%	50	51.5%
4	91	33	36.3%	5	5.5%	53	58.2%
5	155	46	29.7%	27	17.4%	82	52.9%
6	93	28	30.1%	12	12.9%	53	57.0%
7	13	4	30.8%	1	7.7%	8	61.5%
9	12	3	25.0%	2	16.7%	7	58.3%
total	729	247	33.9%	90	12.3%	392	53.8%

# of transmembrane segments	total	purified >10 mg	% purified >10 mg	purified between 0 and 10 mg	% purified between 0 and 10 mg	failed purification	% failed purification
0	603	213	35.3%	77	12.8%	313	51.9%
1	98	27	27.6%	10	10.2%	61	62.2%
>1	28	7	25.0%	3	10.7%	18	64.3%
total	729	247	33.9%	90	12.3%	392	53.8%

low complexity score	total	purified >10 mg	% purified >10 mg	purified between 0 and 10 mg	% purified between 0 and 10 mg	failed purification	% failed purification
0	498	179	35.9%	66	13.3%	253	50.8%
1	139	43	30.9%	13	9.4%	83	59.7%
2	55	14	25.5%	9	16.4%	32	58.2%
3	17	7	41.2%	1	5.9%	9	52.9%
9	20	4	20.0%	1	5.0%	15	75.0%
total	729	247	33.9%	90	12.3%	392	53.8%

disorder score	total	purified >10 mg	% purified >10 mg	purified between 0 and 10 mg	% purified between 0 and 10 mg	failed purification	% failed purification
0	575	194	33.7%	76	13.2%	305	53.0%
4	81	32	39.5%	7	8.6%	42	51.9%
5	8	1	12.5%	1	12.5%	6	75.0%
6	8	4	50.0%	0	0.0%	4	50.0%
7	4	1	25.0%	0	0.0%	3	75.0%
9	53	15	28.3%	6	11.3%	32	60.4%
total	729	247	33.9%	90	12.3%	392	53.8%

Table 4. Cloning Data

# of transmembrane segments	cloning total	cloning +	% cloning +	cloning -	% cloning -
0	2198	1908	87.7%	269	12.2%
1	324	272	84.0%	52	16.0%
>1	137	99	72.3%	38	27.7%
total	2659	2299	86.5%	359	13.5%

Table 5. Protein Expression Data

# of transmembrane segments	total	expression +	% expression +	expression -	% expression -
0	1802	1231	67.6%	591	32.4%
1	274	182	66.4%	92	33.6%
>1	105	59	56.2%	46	43.8%
total	2201	1472	66.9%	729	33.1%

Protein disorder predicts HSQC results

In solution, intrinsically disordered proteins lack a stable three-dimensional conformation or contain a combination of well-structured (*i.e.,* ordered) and disordered regions. Conformational heterogeneity makes intrinsically disordered proteins inherently antagonistic to high throughput solution-based structure determination (for example, 44 of 71 proteins screened were found to be disordered and thus not amenable to NMR-based structure determination). The avoidance of sequences with such properties should increase the efficiency of efforts aimed at high-throughput structure determination. CESG uses the Predictor of Naturally Disordered Regions (PONDR® - developed by Dr. A.K. Dunker and Molecular Kinetics, Inc.) VL-XT algorithm (a collection of neural networks derived from known ordered and disordered residues characterized by X-ray crystallography, NMR, or circular dichroism) to evaluate targets as part of its target scoring scheme. A detailed investigation of the correlation of PONDR analysis with experimental HSQC data has been undertaken (see Oldfield *et. al.*, Proteins, in press).

This study (Table 6) looked at a total of 71 proteins experimentally determined to be either ordered (and thus suitable for structure determination) or disordered (and thus unsuitable for continued experimentation) by HSQC NMR spectroscopy. Of those 71 targets, 42 were selected before PONDR disorder predictions were incorporated into the CESG scoring scheme and 29 were selected after disorder predictions began being used.

Table 6. Disorder Predictions and HSQC Results

disorder selection criteria	number of targets	% proteins predicted to be ordered	% proteins predicted to be disordered	% predictions that proved to be correct by HSQC		% folded by HSQC
				ordered	disordered	
Total	71	56%	44%	55%	84%	38%
None	42	40%	60%	65%	84%	36%
Weighted	29	79%	21%	48%	83%	41%

CESG has found a strong correlation between proteins predicted to be disordered and those found to be experimentally disordered (84% success rate). Incorporation of the disorder score into the CESG scoring scheme has resulted in an increase in the fraction of targets found to be folded (36% prior to incorporation versus 41% after incorporation). While the current target scoring scheme successfully uses a disorder prediction to weight the overall tier score, our data suggests that we may be able to set a hard threshold for disorder score, eliminating a significant fraction of targets, most which are unlikely to yield samples suitable for NMR structure determination. For this data set, such a filtering of targets would have resulted in a 12% increase in success.

Comparing structures solved by structural genomics groups vs. PDB

An analysis of the structures deposited in PDB by structural genomics group reveals several interesting trends. Figure 2 shows the percentage of proteins containing various degrees of low complexity regions in eukaryotic proteins solved by X-ray crystallography from the entire Protein Data Bank (PDB) and from just the structural genomics centers (SG). The differences between the PDB and SG for the various categories indicate that the proteins solved by structural genomics centers contain more regions of low complexity. This shows that the structural genomics centers are solving structures that were previously under-represented in the PDB in terms of presence of low-complexity (see Huntley and Golding, Proteins, 48:134 (2002)). Thus, the SG centers seeming to be fulfilling an important role in structural biology.

When the PDB entries are categorized species (Figure 3), there is a clear difference between the structures in the PDB and the structures solved by the structural genomics groups (SG). The majority of the structural genomics groups have clearly been focusing on prokaryotic proteins (~78% of all SG structures have come from prokaryotes), however, the prokaryotes account for less than 40% of all structures in the PDB. Additional focus on eukaryotic structures must be a priority for the future. The vast majority of CESG structures are eukaryotic and provide a necessary balance with the focus of the structural genomics initiatives. The data in this graph comes from all structures solved by both X-ray crystallography and NMR. Archaea and viruses were included as additional categories and factor into the displayed percentages. However, they are not shown on the graph because the percentages were small compared to the prokaryotes and eukaryotes.

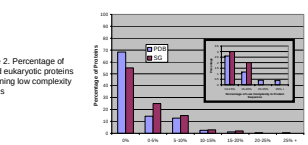

Figure 2. Percentage of solved eukaryotic proteins containing low complexity regions

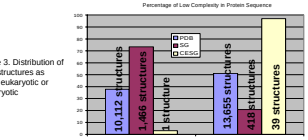

Figure 3. Distribution of PDB structures as either eukaryotic or prokaryotic

CESG is supported by the National Institute of General Medical Sciences through the Protein Structure Initiative (P50 GM 64598).
C.O. and R.B. are supported by the National Library of Medicine through the Computation and Informatics in Biology and Medicine Training Grant (T15 LM 07359).