Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Shotgun crystallization strategy for structural genomics: an optimized two-tiered crystallization screen against the *Thermotoga maritima* proteome

Rebecca Page,[a] Slawomir K. Grzechnik,[b,c] Jaume M. Canaves,[b,c] Glen Spraggon,[c,d] Andreas Kreusch,[c,d] Peter Kuhn,[c,e] Raymond C. Stevens[a,c]* and Scott A. Lesley[c,d]

[a]Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA, [b]University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA, [c]Joint Center for Structural Genomics, USA, [d]Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego CA 92121, USA, and [e]Department of Cell Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA

Correspondence e-mail: stevens@scripps.edu

Received 13 February 2003
Accepted 7 April 2003

As the field of structural genomics continues to grow and new technologies are developed, novel strategies are needed to efficiently crystallize large numbers of protein targets, thus increasing output, not just throughput [Chayen & Saridakis (2002). *Acta Cryst.* D**58**, 921–927]. One strategy, developed for the high-throughput structure determination of the *Thermotoga maritima* proteome, is to quickly determine which proteins have a propensity for crystal formation followed by focused SeMet-incorporated protein crystallization attempts. This experimental effort has resulted in over 320 000 individual crystallization experiments. As such, it has provided one of the most extensive systematic data sets of commonly used crystallization conditions against a wide range of proteins to date. Analysis of this data shows that many of the original screening conditions are redundant, as all of the *T. maritima* proteins that crystallize readily could be identified using just 23% of the original conditions. It also shows that proteins that contain selenomethionine and are more extensively purified often crystallize in distinctly different conditions from those of their native less pure counterparts. Most importantly, it shows that the two-tiered strategy employed here is extremely successful for predicting which proteins will readily crystallize, as greater than 99% of the proteins identified as having a propensity to crystallize under non-optimal native conditions did so again as selenomethionine derivatives during the focused crystallization trials. This crystallization strategy can be adopted for both large-scale genomics programs and individual protein studies with multiple constructs and has the potential to significantly accelerate future crystallographic efforts.

1. Introduction

Protein structure determination is a powerful means for elucidating biological function. The last decade has seen an exponential increase in the number of protein structures solved. However, the total number of structures is still a small fraction of the number of sequenced genes and representative structures of many protein families are still unknown. By implementing high-throughput (HT) and parallel technologies at each step of the structure-determination processes, structural genomics (SG) efforts are positioned to significantly accelerate the rate of structure determination and, in turn, our understanding of protein function and human disease.

The Joint Center for Structural Genomics (JCSG) has developed an HT structural determination pipeline for genome-scale processing (Lesley *et al.*, 2002). The first genome screened by the JCSG pipeline was that of the thermophilic

Table 1
Definition and relationship between the native crystallization screen (tier 1) and the selenomethionine targeted structure-determination screen (tier 2) of the *T. maritima* proteome.

	Tier 1	Tier 2
Project stage		
Protein target selection	*T. maritima* open reading frame Overexpressed Purified	Crystallized in tier 1 Satisfies bioinformatic prioritization of tier 1 hits
Expression	Native conditions	Selenomethionine conditions
Purification	Metal-affinity	Metal-affinity Ion exchange Sometimes size exclusion
Crystallization	480 conditions at 293 K	480 conditions at 293 K 480 conditions at 277 K
Status	Completed	Ongoing Current results reported
Results		
No. of crystallization experiments	258720	66240
Unique proteins set up:unique proteins crystallized	539:465 (86%)	69:68 (99%)
Unique conditions tested:unique conditions that produced crystals	480:472 (98%)	480:416 (87%)
Total crystal hits:harvestable crystal hits	5546:1259 (23%)	1893:779 (41%)

bacterium *T. maritima*. It was selected because its proteome is small [1877 predicted open reading frames (ORFs); Nelson *et al.*, 1999], its proteins have historically overexpressed more readily in *Escherichia coli* than those from non-bacterial organisms and its proteins are expected to be more stable at room temperature than their non-thermophilic homologs (Das & Gerstein, 2000; Vieille & Zeikus, 2001). The JCSG strategy was to process every predicted *T. maritima* ORF (1877) through the SG pipeline (Lesley *et al.*, 2002), rather than restrict the initial target list to a smaller subset of the proteome, such as just those proteins whose structures are novel or just those implicated in human disease. This is a powerful strategy because it not only produces hundreds of new crystals for structure determination, but also enables analysis of the deficiencies in target-selection processing methods for different types of proteins to be identified and modified for future efforts. Additionally, since a number of *T. maritima* (44) structures have been determined by other groups, these targets can be used as controls to indicate the effectiveness of the SG strategy and pipeline process. Finally, this approach produces a less biased and more complete proteome crystallization data set.

Crystallization is often a time-limiting step in protein structure determination owing to the extensive number of components that can be systematically altered for optimal crystal formation. Over the last 20 y, efforts to make the search for optimal crystallization conditions more tractable have led to a number of novel crystallization screens (Carter & Carter, 1979; Jancarik & Kim, 1991; McPherson, 1990). The most widely used screens to date, sparse-matrix screens, sample diverse ranges of buffers and precipitants and are heavily biased towards conditions which have previously produced diffraction-quality crystals (Cudney *et al.*, 1994; Jancarik & Kim, 1991; McPherson, 1990). Initial crystallization trials are typically performed using sparse-matrix screens (coarse screening). Once an initial crystallization condition is found, the condition is then systematically optimized until the protein of interest forms diffraction-quality crystals (fine screening; McPherson, 1990). This approach is not easily implemented nor feasible for large-scale SG projects. Instead, novel strategies and technologies must be employed which streamline the crystallization process.

The JCSG has adopted a two-tiered shotgun strategy for the crystallization of the *T. maritima* proteome in order to identify and focus the majority of crystallization efforts on those proteins with a demonstrated propensity to crystallize (Lesley *et al.*, 2002; Table 1). This strategy is founded on the hypothesis that proteins which crystallize readily, even under suboptimal conditions, will do so again during focused crystallization attempts. In tier 1, the goal is to identify those targets which have a propensity to crystallize under the conditions tested; the quality of the crystals produced is not significantly important. To maximize throughput, the protein samples are purified with only one round of affinity purification and screened for crystal formation against a limited number of crystallization conditions; it is expected that some of the proteins will not be sufficiently pure or in the optimal state to crystallize. In tier 2, the objective is to obtain diffraction-quality crystals suitable for structure determination. In this stage, the targets that crystallized in tier 1 are reprocessed to contain selenomethionine, purified extensively and screened against an expanded set of crystallization conditions.

The processing of the *T. maritima* proteome through the JCSG HT structure-determination pipeline has resulted in over 320 000 individual crystallization experiments (Table 1). Both the positive (crystals) and negative (precipitation/clear drops) results have been recorded, making it one of the most systematic and extensive data sets available for the evaluation of sparse-matrix conditions for crystal formation to date (data available at http://www.jcsg.org). Analysis of these experiments indicates that the two-tiered crystallization strategy effectively limits the majority of crystallization efforts to those proteins with a demonstrated ability to crystallize. In addition, it also shows that this strategy can be optimized to further increase the efficiency of genome-scale crystallization efforts. In particular, a number of crystallization conditions are either redundant or ineffective and thus can be eliminated in future tier 1 screens. These conditions should not, however, be eliminated in tier 2 screens. Rather, tier 2 screens should be expanded further in order to maximize the likelihood that a diffraction-quality crystal will be obtained during screening.

The conclusions drawn from these studies will be used to optimize the efficiency of future SG efforts and should also be

Table 2
Proteins set up and crystallized in tier 1 of the *T. maritima* proteome screen.

Protein function	No. of *T. maritima* proteins	Proteins passed to tier 1			
		Set up	%	Crystallized	%
Amino-acid biosynthesis	73	34	47	29	86
Biosynthesis of cofactors, prosthetic groups and carriers	31	14	45	10	71
Cell envelope	73	15	21	14	93
Cellular processes	49	12	24	9	75
Central intermediary metabolism	44	23	52	21	91
DNA metabolism	54	15	28	11	73
Energy metabolism	195	88	45	84	95
Fatty acid/phospholipid metabolism	15	6	40	5	83
Hypothetical proteins	774	188	24	159	85
Other categories	13	4	31	3	75
Protein fate	48	6	13	5	83
Protein synthesis	106	30	28	27	90
Purines, pyrimidines, nucleosides and nucleotides	45	20	44	17	85
Regulatory functions	70	18	26	16	89
Transcription	16	8	50	7	88
Transport and binding proteins	188	30	16	25	83
Unknown function	83	28	34	23	82
Total	1877	539	29	465	86

applicable to single-protein studies. For single targets, multiple protein constructs can be rapidly purified and screened in parallel to determine which protein constructs have the greatest propensity to crystallize (tier 1). Subsequent crystallization efforts (tier 2) can then focus on only those clones that crystallize readily.

2. Experimental

Processing of the *T. maritima* proteome has been described previously (Lesley *et al.*, 2002). Briefly, every *T. maritima* ORF was processed through the JCSG high-throughput structural genomics pipeline using automated technologies to maximize efficiency. In tier 1, proteins were expressed under native conditions and purified by a single pass of affinity chromatography using HT parallel-processing methods (Table 1). Each of the proteins included an N-terminal tag to facilitate expression and purification (MGSDKIHHHHHH), which was not removed prior to crystal screening. Once purified, these proteins were immediately screened against 480 different commonly used sparse-matrix crystallization conditions at a single temperature, 293 K, for crystal formation (Table 2). None of the samples were evaluated for structural content (CD) or aggregation (DLS) prior to crystal trials, since the purpose of the screen was to identify those proteins with a natural propensity to crystallize under minimal purification conditions. Up to 96 protein samples were screened for crystal formation against 480 crystallization conditions per week using this pipeline strategy.

In tier 2, proteins that crystallized in tier 1 were prioritized for reprocessing based on the likelihood that they had novel folds. When reprocessed, these targets were expressed in the presence of selenomethionine and purified extensively, using a combination of affinity chromatography, ion-exchange and occasionally size-exclusion chromatography. Purified tier 2 proteins were then screened for crystal formation using the same 480 conditions as tier 1, but now at two distinct temperatures, 277 and 293 K. As in tier 1, the N-terminal expression/purification tag was not removed prior to crystal trials, nor were the protein samples examined by CD or DLS for structural integrity.

The 480 conditions used for crystal screening sampled a wide range of precipitant, buffer, additive and pH variables. These conditions were compiled from ten commercially available kits [Crystal Screen, Crystal Screen 2, Crystal Screen Cryo, PEG/Ion Screen, Grid Screen Ammonium Sulfate, Grid Screen PEG 6000, Grid Screen PEG/LiCl, Grid Screen MPD (Hampton Research, Riverside, CA, USA), Wizard I/II, Cryo I/II (Emerald Biostructures, Bainbridge Island, WA, USA)] and the collection is best described as a broad sparse-matrix screen with more finely sampled grid screens about some of the sparse-matrix conditions. The conditions, many of which have previously produced diffraction-quality protein crystals, can be divided into five classes based on the primary precipitant type: high-MW PEGs (1000–20 000 Da), low-MW PEGs (200–1000 Da), ammonium sulfate/salts, polyalcohols (such as MPD, EG and 1,2-butanediol) and other organics (such as ethanol, Jeffamine and 2-propanol). The majority of conditions contain high-MW PEGs as the primary precipitant (171), with the fewest containing other organics (55). While some of these conditions were quite similar to one another, all were included in the initial screen so that those conditions that were most effective at promoting protein crystallization could be identified.

Crystallization was carried out using the vapor-diffusion method. Sitting drops composed of only 50 nl of protein and 50 nl of mother liquor were equilibrated against 100 µl of well solution (Santarsiero *et al.*, 2002). Images of the drops were taken 1, 7 and 28 d following setup and then examined manually for crystal formation. Crystal hits (Fig. 1) were classified as either 'hits for fine screening' (not suitable for mounting, but requiring further fine screening) or 'harvestable' (suitable for mounting and data collection). Harvestable crystals were either immediately flash-cooled using liquid nitrogen (those grown in cryoprotectant conditions) or else passed through cryoprotectant solution just prior to freezing. Mounted crystals were then screened for diffraction at 100 K using the beamlines at the Stanford Synchrotron Radiation Laboratory (SSRL). Data sets from those crystals which diffracted to at least 3.0 Å were collected and ultimately processed using the protocols established by the JCSG for structure determination.

Minimal screens (subsets of conditions that would produce crystals for every tier 1 target crystallized) were calculated using the *Min_Cov* algorithm (S. Grzechnik, in preparation). Initially, the algorithm is seeded with a single condition that produces a crystal for one or more targets; this is the first member of a set of conditions called the Current Selection (CSel). The algorithm then identifies which of the remaining conditions would maximize the number of novel targets crystallized by the conditions in the CSel. Once identified, the

condition is added to the CSel and the calculation is repeated with the remaining conditions. Conditions are added to the CSel in this way until it contains a set of conditions which produce crystals for every crystallized target. This set of conditions is then referred to as a Minimal Screen. This algorithm, like most optimization algorithms, can only be used to find local solutions, not global ones. Thus, repeated runs of the program using a different conditions for the initial seed of the CSel often results in the identification of minimal screens which are composed of different conditions. In this study, each condition that produced a crystal for the tier 1 crystallized targets was used as an initial seed. By identifying the conditions present in every minimal screen, a Core Screen (CS) for the set of crystallized targets was identified. Since these conditions are present in every calculated screen, they are considered to be those which are most critical for the crystallization of the tier 1 proteins targets.

3. Results and discussion

3.1. Two-tiered crystallization strategy for the *T. maritima* proteome

The strategy for screening the *T. maritima* genome was to process every predicted ORF (1877) through the two-tiered SG pipeline (Lesley *et al.*, 2002) (Table 1). In tier 1, proteins were cloned, expressed under native conditions and rapidly purified in parallel using single-step affinity chromatography, followed by buffer exchange and concentration. Protein purity was not optimized, but only those samples that were significantly enriched with the protein of interest (90% or greater, as determined by gel electrophoresis) were then subjected to 480 commercially available coarse-screen conditions at 293 K for crystallization. While some of these conditions were similar to one another (owing to overlap in the commercially available screens), all 480 conditions were still used, so that those conditions which most effectively promote protein-crystal formation could be identified (see Fig. 2). Proteins that crystallized in tier 1 were then prioritized for fold novelty and passed to tier 2, where they were expressed in the presence of selenomethionine and this time extensively purified using affinity, ion-exchange and sometimes size-exclusion chromatography. After buffer exchange and concentration, the proteins were then subjected to the same coarse-screen conditions as in tier 1, but now at 277 and 293 K. Tier 1 screening has been completed, with 539 targets for a total of 260 160 crystal trials; tier 2 screening is still ongoing, with 69 targets for a total of 66 240 trials completed to date (Table 1).

3.1.1. Tier 1: identification of *T. maritima* proteins with a high propensity to crystallize. In tier 1, 539 (29%) of the predicted 1877 *T. maritima* ORFs were successfully cloned, solubly expressed, purified and sent to crystallization trials. Of these, 465 crystallized, a success rate of 86% (Fig. 2*a* and Table 2), and over half of these crystallized in five or more conditions. Based on their computed isoelectric points (pI), molecular weights (MW) and biological functions, the proteins that crystallized comprised a broad subset of the entire proteome. They tended to be smaller (average *T. maritima* protein = 35 161 Da; average crystallized *T. maritima* protein = 31 969 Da) and slightly more acidic (average *T. maritima* protein pI = 7.16; average crystallized *T. maritima* protein pI = 6.55) than those of the proteome as a whole, yet they still spanned an extensive range of pI and MW values. In fact, crystals were obtained from proteins as small as 4821 Da and as large as 93 500 Da, with pIs between 4.1 and 11.4. In addition, between 10 and 48% of the proteins from each

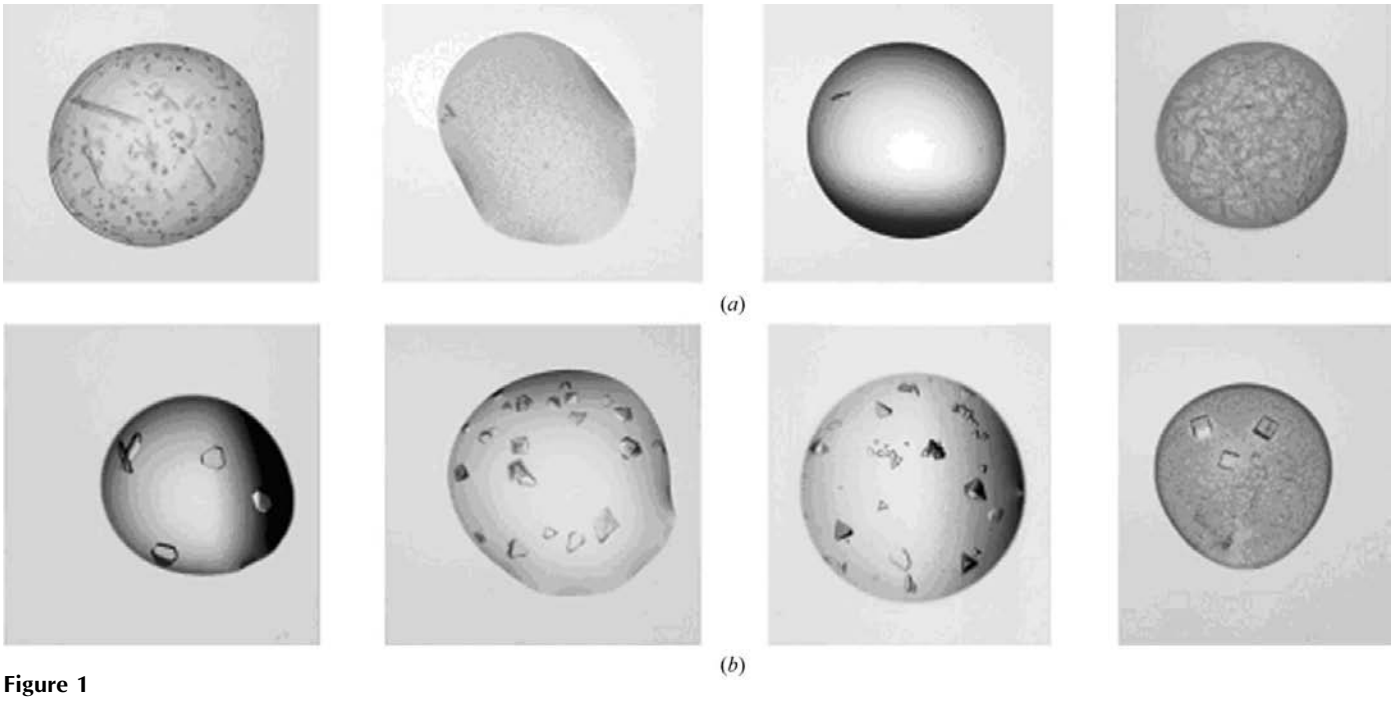

Figure 1
Examples of protein crystals described as hits for fine screening (*a*) and harvestable (*b*).

T. maritima functional class (Nelson *et al.*, 1999) were represented in this set of crystallized targets, from proteins involved in amino-acid biosynthesis to those required for transcription

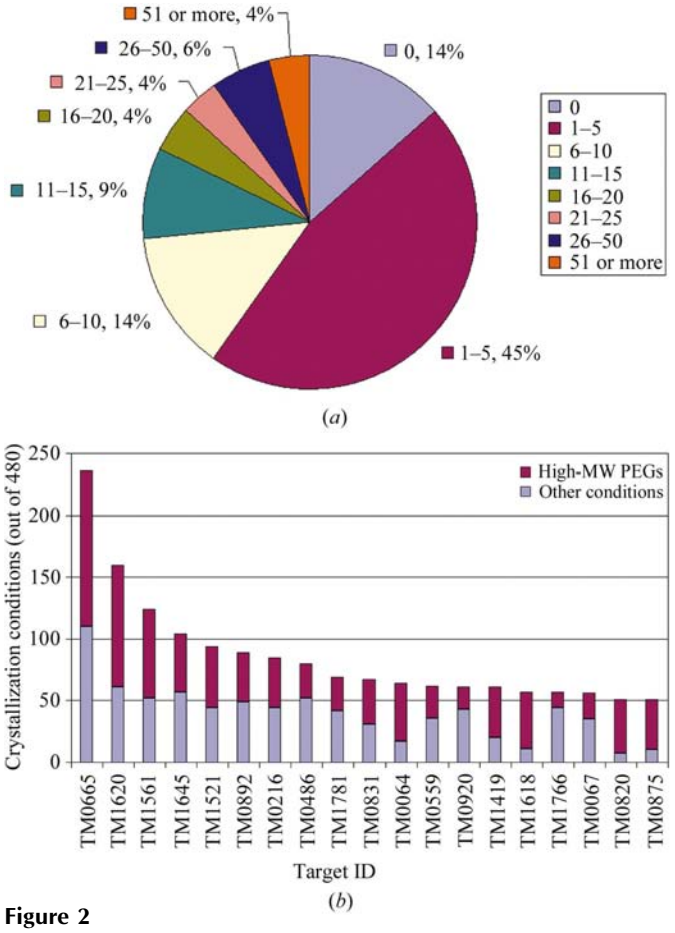

Figure 2
Number of crystal hits per protein in tier 1. (*a*) The number of crystal hits per protein obtained in tier 1, the *T. maritima* native proteome screen. (*b*) Proteins which crystallized in 50 or more conditions in the tier 1 screen. Four proteins, TM0665, TM1620, TM1561 and TM1645, crystallized in over 100 distinct conditions.

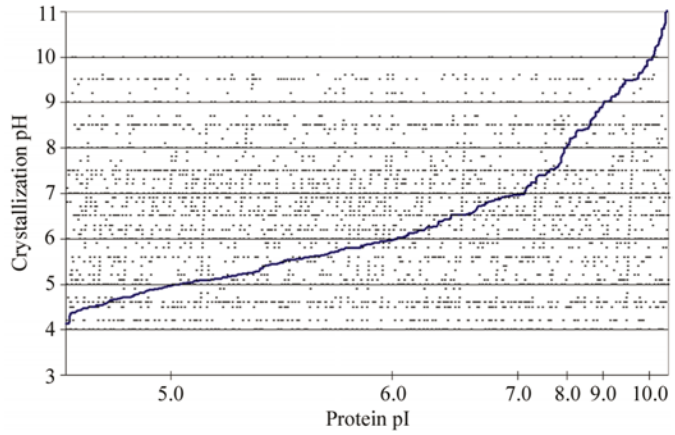

Figure 3
Protein pI *versus* crystallization pH of tier 1 crystallization hits. Scatter plot of protein pI *versus* crystallization condition pH. The blue line indicates protein pI. There is no correlation between protein pI and crystallization condition pH.

(Table 2). As expected, the proteins predicted to be membrane or membrane-associated proteins, such as cell envelope or transport and carrier proteins, were less well represented than those expected to be expressed in the cytosol, such as metabolic and biosynthetic enzymes. Finally, only three of the 1800 protein crystals screened for diffraction to date have been salt, indicating that the high crystallization rate does not arise from artifactual results. The low percentage of salt-crystal formation is probably largely a consequence of the absence of components in the protein buffer (20 m*M* Tris pH 7.4, 150 m*M* NaCl, 0.25 m*M* TCEP) that interact with those in the various crystallization conditions to promote salt-crystal formation.

Nearly every condition (473; Table 1) of the 480 used for screening produced a crystal of at least one of the tier 1 targets that crystallized and many produced crystals for five targets or more. This resulted in a total of 5546 unique crystal hits (individual protein/condition combinations) for the entire 465-protein set. While the proteins might have been expected to preferentially crystallize at pH values near their isoelectric points, those pH values at which proteins have no net charge and are often minimally soluble, the results from this study show that for successful crystallization protein pI and condition pH are not necessarily correlated (Fig. 3). Many of the proteins crystallized over a wide range of pH values, often over 6 pH units, and usually more than 1 pH unit away from the protein pI. The correlation coefficients between protein pI and crystallization condition pH were only 0.01 for all the crystal hits and still only 0.07 when proteins which crystallized in more than eight conditions (181) were excluded from the calculation. This lack of correlation between protein pI and crystallization pH has been observed previously (Gilliland & Ladner, 1996; Gilliland *et al.*, 1996; McPherson, 1994).

At the time of manuscript submission, there were structures for 54 different *T. maritima* proteins determined crystallographically and deposited in the PDB; ten of these have been solved by the JCSG.[1] Of the remaining 44 proteins, 19 have successfully passed through the JCSG SG pipeline to purification, of which 18 have crystallized (the one exception, ribosomal protein L12, TM0047, was expressed and purified very differently from the methods used here; Wahl, Bourenkov *et al.*, 2000; Wahl, Huber *et al.*, 2000). These results show that the methods used for tier 1 screening successfully identify proteins with a high propensity for crystal formation. However, the inability of TM0047 to crystallize in tier 1 and the failure of 25 targets to reach crystal trials also suggests that additional processing methods need to be developed to enable more difficult targets to be expressed, purified and crystallized.

3.1.2. Many commonly used crystallization conditions are redundant. Tier 1 screening against this set of 480 conditions resulted in the crystallization of 465 different *T. maritima* proteins. The ten most effective conditions produced crystals for 196 (42%) different proteins, while the best 108 (23%)

[1] *Note added in proof:* there are now 95 different *T. maritima* structures in the PDB (four from NMR), 40 of these are from the JCSG.

Table 3
Top unique conditions based on the tier 1 native proteome screen (Core Screen).

	Condition	Screen†
1	50%(*w*/*v*) PEG 400, 0.2 *M* Li_2SO_4, 0.1 *M* acetate pH 5.1	W1cryo #47
2	20%(*w*/*v*) PEG 3000, 0.1 *M* citrate pH 5.5	W1 #06
3	20%(*w*/*v*) PEG 3350, 0.2 *M* diammonium hydrogen citrate pH 5.0	PEG/ion #48
4	30%(*v*/*v*) MPD, 0.02 *M* $CaCl_2$, 0.1 *M* NaOAc pH 4.6	H1 #01
5	20%(*w*/*v*) PEG 3350, 0.2 *M* magnesium formate pH 5.9	PEG/ion #20
6	20%(*w*/*v*) PEG 1000, 0.2 *M* Li_2SO_4, phosphate–citrate pH 4.2	W1 #39
7	20%(*w*/*v*) PEG 8000, 0.1 *M* CHES pH 9.5	W1 #01
8	20%(*w*/*v*) PEG 3350, 0.2 *M* ammonium formate pH 6.6	PEG/ion #23
9	20%(*w*/*v*) PEG 3350, 0.2 *M* ammonium chloride pH 6.3	PEG/ion #09
10	20%(*w*/*v*) PEG 3350, 0.2 *M* potassium formate pH 7.3	PEG/ion #22
11	50% MPD, 0.2 *M* $(NH_4)H_2PO_4$, 0.1 *M* Tris pH 8.5	H2 #43
12	20%(*w*/*v*) PEG 3350, 0.2 *M* potassium nitrate pH 6.9	PEG/ion #18
13	0.8 *M* $(NH_4)_2SO_4$, 0.1 *M* citric acid pH 4.0	$AmSO_4$ #01
14	20%(*w*/*v*) PEG 3350, 0.2 *M* sodium thiocyanate pH 6.9	PEG/ion #13
15	20%(*w*/*v*) PEG 6000, 0.1 *M* bicine pH 9.0	P6K #18
16	10%(*w*/*v*) PEG 8000, 8% ethylene glycol, 0.1 *M* HEPES pH 7.5	H2 #37
17	40%(*v*/*v*) MPD, 5%(*w*/*v*) PEG 8000, 0.1 *M* cacodylate pH 7.0	W2cryo #01
18	40% ethanol, 5%(*w*/*v*) PEG 1000, 0.1 *M* phosphate–citrate pH 5.2	W1cryo #40
19	8%(*w*/*v*) PEG 4000, 0.1 *M* NaOAc pH 4.6	H1 #37
20	10%(*w*/*v*) PEG 8000, 0.2 *M* $MgCl_2$, 0.1 *M* Tris pH 7.0	W2 #43
21	20%(*w*/*v*) PEG 6000, 0.1 *M* citric acid pH 5.0	P6K #14
22	50%(*v*/*v*) PEG 200, 0.2 *M* $MgCl_2$, 0.1 *M* cacodylate pH 6.6	W2cryo #36
23	1.6 *M* sodium citrate pH 6.5	H2 #28
24	20%(*w*/*v*) PEG 3350, 0.2 *M* tripotassium citrate monohydrate pH 8.3	PEG/ion #47
25	30% MPD, 0.02 *M* $CaCl_2$, 0.1 *M* NaOAc pH 4.6	H1cryo #01
26	20%(*w*/*v*) PEG 8000, 0.2 *M* NaCl, 0.1 *M* phosphate–citrate pH 4.2	W1 #31
27	20%(*w*/*v*) PEG 6000, 1.0 *M* LiCl, 0.1 *M* citric acid pH 4.0	P6K/LiCl #13
28	20%(*w*/*v*) PEG 3350, 0.2 *M* ammonium nitrate pH 6.3	PEG/ion #19
29	10%(*w*/*v*) PEG 6000, 0.1 *M* HEPES pH 7.0	P6K #10
30	0.8 *M* NaH_2PO_4/0.8 *M* KH_2PO_4, 0.1 *M* HEPES pH 7.5	H1 #35
31	40%(*v*/*v*) PEG 300, 0.1 *M* phosphate–citrate pH 5.2	W2cryo #18
32	10%(*w*/*v*) PEG 3000, 0.2 *M* $Zn(OAc)_2$, 0.1 *M* acetate pH 4.5	W2 #01
33	20% ethanol, 0.1 *M* Tris pH 8.5	H2 #44
34	25%(*v*/*v*) 1,2-propanediol, 0.1 *M* Na/K phosphate, 10%(*v*/*v*) glycerol pH 6.8	W2cryo #11
35	10%(*w*/*v*) PEG 20 000, 2% dioxane, 0.1 *M* bicine pH 9.0	H2 #48
36	2.0 *M* $(NH_4)_2SO_4$, 0.1 *M* acetate pH 4.6	H1 #47
37	10%(*w*/*v*) PEG 1000, 10%(*w*/*v*) PEG 8000	H2 #07
38	24%(*w*/*v*) PEG 1500, 20% glycerol	H1cryo #43
39	30%(*v*/*v*) PEG 400, 0.2 *M* $MgCl_2$, 0.1 *M* HEPES pH 7.5	H1cryo #23
40	50%(*v*/*v*) PEG 200, 0.2 *M* NaCl, 0.1 *M* Na/K phosphate pH 7.2	W2cryo #15
41	30%(*w*/*v*) PEG 8000, 0.2 *M* Li_2SO_4, 0.1 *M* acetate pH 4.5	W1 #17
42	70%(*v*/*v*) MPD, 0.2 *M* $MgCl_2$, 0.1 *M* HEPES pH 7.5	H2 #35
43	20%(*w*/*v*) PEG 8000, 0.1 *M* Tris pH 8.5	W2 #03
44	40%(*v*/*v*) PEG 400, 0.2 *M* Li_2SO_4, 0.1 *M* Tris pH 8.4	W1cryo #38
45	40%(*v*/*v*) MPD, 0.1 *M* Tris pH 8.0	MPD #17
46	25.5%(*w*/*v*) PEG 4000, 0.17 *M* $(NH_4)_2SO_4$, 15% glycerol	H1cryo #31
47	40%(*v*/*v*) PEG 300, 0.2 *M* $Ca(OAc)_2$, 0.1 *M* cacodylate pH 7.0	W1cryo #37
48	14% 2-propanol, 0.14 *M* $CaCl_2$, 0.07 *M* acetate pH 4.6, 30% glycerol	H1cryo #24
49	16%(*w*/*v*) PEG 8000, 0.04 *M* KH_2PO_4, 20% glycerol	H1cryo #42
50	1.0 *M* sodium citrate, 0.1 *M* cacodylate pH 6.5	W1 #14
51	2.0 *M* $(NH_4)_2SO_4$, 0.2 *M* NaCl, 0.1 *M* cacodylate pH 6.5	W2 #04
52	10% 2-propanol, 0.2 *M* NaCl, 0.1 *M* HEPES pH 7.5	W1 #02
53	1.26 *M* $(NH_4)_2SO_4$, 0.2 *M* Li_2SO_4, 0.1 *M* Tris pH 8.5	W1 #47
54	40%(*v*/*v*) MPD, 0.1 *M* CAPS pH 10.1	W2cryo #25
55	20%(*w*/*v*) PEG 3000, 0.2 *M* $Zn(OAc)_2$, 0.1 *M* imidazole pH 8.0	W2 #40
56	10% 2-propanol, 0.2 *M* $Zn(OAc)_2$, 0.1 *M* cacodylate pH 6.5	W2 #11
57	1.0 *M* $(NH_4)_2HPO_4$, 0.1 *M* acetate pH 4.5	W1 #09
58	1.6 *M* $MgSO_4$, 0.1 *M* MES pH 6.5	H2 #20
59	10%(*w*/*v*) PEG 6000, 0.1 *M* bicine pH 9.0	P6K #12
60	14.4%(*w*/*v*) PEG 8000, 0.16 *M* $Ca(OAc)_2$, 0.08 *M* cacodylate pH 6.5, 20% glycerol	H1cryo #46
61	10%(*w*/*v*) PEG 8000, 0.1 *M* imidazole pH 8.0	W2 #34
62	30% Jeffamine M-600, 0.05 *M* CsCl, 0.1 *M* MES pH 6.5	H2 #24
63	3.2 *M* $(NH_4)_2SO_4$, 0.1 M citric acid pH 5.0	AmSO4 #20
64	20% MPD, 0.1 *M* Tris pH 8.0	MPD #11
65	20% Jeffamine M-600, 0.1 *M* HEPES pH 6.5	H2 #31
66	50%(*v*/*v*) ethylene glycol, 0.2 *M* $MgCl_2$, 0.1 *M* Tris pH 8.5	W1cryo #43
67	10% MPD, 0.1 *M* bicine pH 9.0	MPD #06

† H1, H2, H1cryo, PEG/Ion, $AmSO_4$, P6K, P6K/LiCl, MPD: Crystal Screen, Crystal Screen 2, Crystal Screen Cryo, PEG/Ion Screen, Grid Screen Ammonium Sulfate, Grid Screen PEG 6000, Grid Screen PEG/LiCl, Grid Screen MPD, respectively (Hampton Research). W1, W2, W1cryo, W2cryo: Wizard I and II and Cryo I and II, respectively (Emerald Biostructures).

produced crystals for all 465. This indicates that many of the conditions used here are unnecessary for the purpose of initial screening and can be eliminated in future tier 1 screens. Since some of the initial conditions were quite similar to one another, this substantial decrease in the number of required conditions was not entirely unexpected; however, by including all of the conditions in the initial screen it did enable the most effective crystallization conditions to be determined.

To identify the redundant conditions, the five major precipitants [high-MW PEGs (1000–20 000 Da), low-MW PEGs (200–1000 Da), ammonium sulfate/salts, polyalcohols (such as MPD, EG and 1,2-butanediol) and other organics (such as ethanol, Jeffamine and 2-propanol)] were examined for their ability to promote crystallization. It was found that the greatest number of distinct proteins (358) crystallized in conditions containing high-MW PEGs, while the fewest (210) crystallized in conditions containing other organics (Fig. 4). This difference was a consequence in part of the difference in the number of high-MW PEG (171) *versus* other organic (55) conditions available for screening. When the number of distinct proteins was normalized by the number of conditions, organic precipitants crystallized the largest number of distinct proteins per condition tested, whereas high-MW PEG precipitants produced the fewest. This indicates that many of the high-MW PEG conditions are redundant in the tier 1 crystal screen.

The oversampling of PEG conditions in the tier 1 screen was especially apparent when analyzing the conditions of those proteins with a spectacular ability to crystallize (Fig. 2*b*). TM0665, TM1620, TM1561 and TM1645 crystallized in 236 (49% of the 480

screened), 160 (33%), 124 (26%) and 104 (29%) distinct conditions, respectively, while 15 other proteins crystallized in over 48 distinct conditions (10%; Fig. 2*b*). The majority of these conditions contained high-MW PEGs (55% of the conditions for the top four crystallizing proteins contained high-MW PEGs, even though high-MW PEGs only accounted for 35% of the total conditions tested). TM0665, TM1620, TM1561 and TM1645 crystallized in 126, 99, 72 and 47 different high-MW PEG conditions, respectively, with PEG components that ranged in size from 1000 to 8000 Da and had concentrations between 5 and 30%(*w*/*v*). PEGs are precipitants which have been shown to promote crystallization over a broad range of sizes and concentrations (McPherson, 1976), consistent with these results.

Since many of the proteins crystallize in multiple conditions, many of the conditions can be eliminated without impacting the number of distinct proteins crystallized. These redundant conditions were identified using the *Min_Cov* algorithm (S. Grzechnik, in preparation). The program uses an iterative selection algorithm to identify subsets of the 480 screening conditions which would have produced crystals for every protein crystallized in tier 1 (for details, see *Experimental*). In this study, each condition that produced a crystal for the tier 1 targets was used as an initial seed for *Min_Cov*. This resulted in 473 different runs of the program, which produced 415 distinct minimal screens (sometimes the same minimal screen was identified even though the condition used for the initial seed was different). By identifying the conditions present in each of the 415 minimal screens, a Core Screen (CS) for the set of targets crystallized in tier 1 was identified (Table 3). Since these conditions were present in every calculated minimal screen, they were considered to be those which were most essential for crystallizing the tier 1 targets.

Only 108 conditions (23%; the number of conditions in the smallest minimal screen) were needed to crystallize all 465 proteins in tier 1 and the Core Screen contained only 67 conditions. Significantly, restricting the initial screening to only these 67 conditions (14% of the original 480) would have still produced crystals for 392 of the 465 crystallized proteins, or 84% of the entire set crystallized. All five primary precipitant classes [high-MW PEGs (31 of the original 171 conditions), low-MW PEGs (8 of 67), ammonium sulfate/salts (10 of 106), polyalcohols (11 of 83) and remaining organics (7 of 54)] were represented in this Core Screen, although high-MW PEG conditions were the most prevalent (47%). As expected, however, based on the analysis of those proteins with a spectacular ability to crystallize, more high-MW PEG conditions were eliminated (140) from the tier 1 screen than any other precipitant class. Remarkably, the ten best conditions of the Core Screen still produced crystals for 192 (41%) different proteins (Table 3). As observed for the entire Core Screen, the majority (eight) of the top ten conditions also contained PEGs as their primary precipitant.

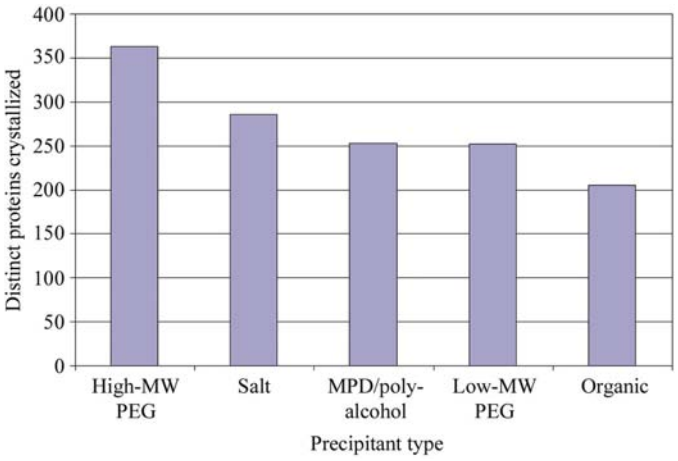

Figure 4
Number of distinct proteins crystallized by primary precipitant type.

3.1.3. Initial screens are still incomplete. While most proteins readily crystallized in one or more of the original 480 screening conditions, 74 failed to crystallize in any of them (Fig. 2*a*). These proteins may have failed to crystallize for any number of reasons: they may not have been sufficiently pure for crystal formation after just one round of affinity purification, they may have aggregated prior to screening, the N-terminal expression/purification tag may have inhibited crystallization, they may have required binding partners or small-molecule cofactors for folding and stability or they may not have been screened against their optimal crystallization conditions. In fact, recent results show that the minimal purification protocol used here may have inhibited crystallization of some of the proteins: four out of six proteins which did not crystallize after only one round of affinity chromatography did so after affinity, ion-exchange and size-exclusion chromatography. Thus, by extension, two-thirds of the proteins which did not crystallize in tier 1 may do so if reprocessed with additional purification measures. In addition, it is expected that since most sparse-matrix screen conditions are heavily biased towards those which have previously produced protein crystals, some protein families which have never crystallized before may require totally novel conditions for crystal formation. An active search for novel precipitants and conditions is under way for these recalcitrant targets. Detailed sequence and functional analysis of these proteins is also under way in an effort to identify characteristics which might suggest their limited crystallization potential prior to screening. Research in these areas is ongoing and the results will be presented elsewhere.

3.1.4. Glycerol inhibits crystal formation. Finally, the tier 1 results also show that the presence of cryoprotectants in screening conditions, especially glycerol, generally inhibit crystal formation. When the crystal hits of tier 1 were grouped by commercially available screen, it was found that while most screens produced crystals for nearly equivalent numbers of different proteins (each of the top six screens produced crystals for between 213 and 224 different proteins; data not shown), the screens identified as cryoscreens generally produced crystals for fewer proteins (160–208). In fact, 44 identical conditions containing glycerol resulted in only 70% (154/219) of the proteins crystallizing. More importantly, all but five (97%) proteins that crystallized in conditions containing glycerol also crystallized in conditions without it,

indicating that conditions containing glycerol can be eliminated in future tier 1 screens without significantly impacting the final number of distinct proteins crystallized.

3.2. Tier 2: screening for diffraction-quality crystals suitable for structure determination

Proteins that crystallized in tier 1 were prioritized and reprocessed to produce diffraction-quality crystals suitable for structure determination. Specifically, the targets that crystallized in tier 1 were reprocessed to contain selenomethionine, purified extensively and screened against the set of 480 crystallization conditions at two distinct temperatures: 277 and 293 K. Moreover, it was hoped many of the crystals produced during the automated screening procedure in tier 2 would be immediately harvestable, reducing the need for time-consuming protein-specific fine screens.

Of the 69 proteins processed to date in tier 2, 68 (99%) have been successfully crystallized (Fig. 5 and Table 1). This nearly perfect success rate validates the two-tiered approach as a very efficient method for producing diffraction-quality crystals. The percentage of the total crystals that were harvestable nearly doubled (41%) in tier 2, with 63 (89%) of the tier 2 proteins producing at least one harvestable crystal directly from the coarse-screen nanodrops. This increase was probably a consequence of the more extensive purification of these samples compared with their tier 1 counterparts. To date, nearly all of the tier 2 crystals which have been mounted for diffraction screening have been harvested directly from the nanodrops that were set up using automated robotic systems (Santarsiero *et al.*, 2002).

3.2.1. More extensively purified selenomethionine-containing proteins crystallize differently from their native counterparts. While the purpose of tier 1 screening was to identify those proteins which have a propensity to crystallize, it was expected that many of the conditions that produced crystals for the native proteins would also produce crystals for their more extensively purified selenomethionine-containing counterparts. Surprisingly, this was not generally the case. Instead, on average only 30% of the original tier 1 conditions that successfully produced crystals for a given native protein produced crystals for its selenomethionine-containing counterpart (Figs. 6*a* and 6*b*). In fact, only one of the ten most effective crystallization conditions for tier 2 was also identified as one of the ten most effective conditions for tier 1 (data not shown). This clearly indicates that the tier 2 samples should be considered to be different proteins with unique crystallization properties. The differences in crystallization conditions between the two sets of proteins could be because of the more extensive purification protocols used in tier 2 (single-pass affinity *versus* affinity, ion-exchange and sometimes size-exclusion chromatography) and/or the different expression conditions in tier 1 *versus* tier 2 (native *versus* selenomethionine). Regardless, it is clear that for the two-tiered crystallization strategy employed here each tier 2 selenomethionine extensively purified protein must be considered to be distinct from its native less pure counterpart and must be rescreened against all 480 crystallization conditions in order to maximize the likelihood that diffraction-quality crystals will be obtained.

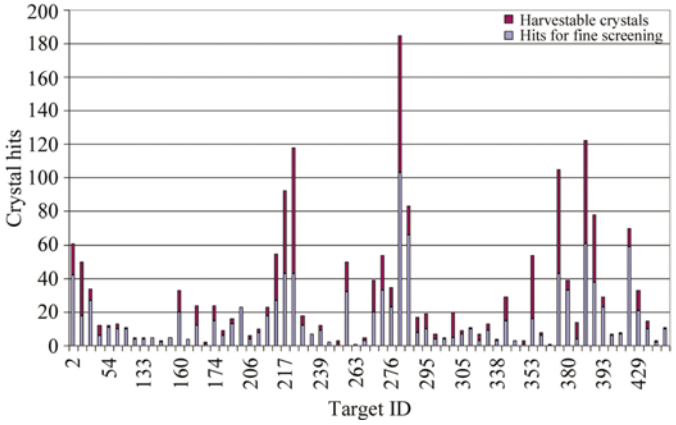

Figure 5
Number of crystal hits per protein in tier 2. Blue bars indicate the number of crystals identified as hits which need fine screening and magenta bars indicate the crystals classified as harvestable and ready for data collection.

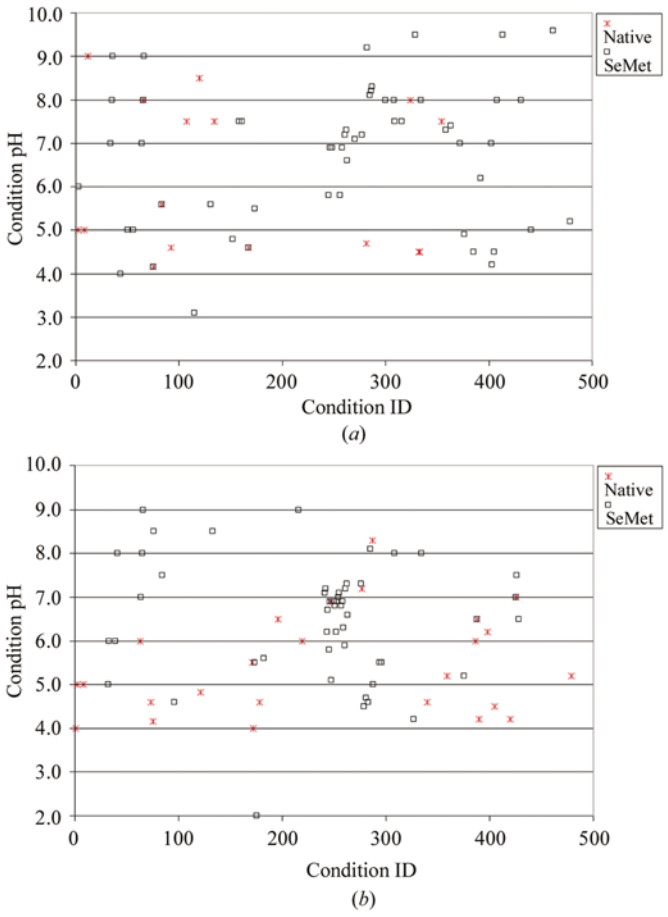

Figure 6
Native crystallization conditions do not predict selenomethionine crystallization conditions. Scatter plots of conditions that produced crystals in tier 1 (red asterisks) and tier 2 (black squares) for two targets: (*a*) TM0008 and (*b*) TM0828.

3.2.2. Crystals grown in cryoconditions are preferentially used for structure determination. The crystallization of protein in the presence of cryoprotectants significantly streamlines the crystal freezing process, since crystals grown in such conditions can be frozen without first being passed through cryoprotectant solution. Even though cryoprotectants generally inhibit crystal formation, they are still essential for the success of any HT SG efforts. To date, over half of the structures solved by the JCSG have been obtained from crystals grown in cryoprotectant conditions.[2] One-third of the screening conditions used in this study contain cryoprotectants at concentrations sufficient to prevent ice formation upon freezing. Crystals grown in one of the remaining conditions, however, must be passed through a cryoprotectant solution prior to freezing, a process that exposes the crystal to a new chemical environment (the cryoprotectant solution) and requires additional crystal handling. This process may degrade the quality of the crystal, resulting in poor X-ray diffraction. Harvestable crystals for a given protein which grew in the presence of cryoprotectants were therefore preferentially selected for diffraction screening, resulting in a bias of the number of solved structures towards these types of crystals. Increasing the number of effective cryoprotectant crystallization conditions in future tier 2 screens may improve the efficiency of tier 2 screening even further.

3.2.3. The effect of temperature. Tier 2 proteins were screened against all 480 sparse-matrix screening conditions at two distinct temperatures, 277 and 293 K, in order to maximize the probability that more crystals, particularly harvestable crystals, would be obtained during the automated screening procedure. This indeed was the case and, as expected, many of the proteins which crystallized in a given condition did so at only one of the two temperatures (data not shown). Interestingly, there were 25% more crystal hits at 293 than 277 K. This difference could be because of an increased tendency of *T. maritima* proteins to crystallize at room temperature, but is more likely to reflect the fact that the initial tier 1 screening was carried out at 293 K and thus may have biased the protein set which was passed to tier 2 towards those proteins which crystallize more readily at this temperature. Temperature, however, had no influence on the production of harvestable crystals, since 41% of the crystals obtained at both 293 and 277 K were described as harvestable.

4. Conclusions

The two-tiered strategy implemented for the crystallization of the *T. maritima* proteome successfully identified those proteins with a high propensity for crystal formation, confirming the hypothesis that proteins which crystallize readily, even under suboptimal conditions, will do so under a variety of conditions. This enabled the majority of crystallization efforts to be carried out only on those proteins most likely to form diffraction-quality crystals, thereby maximizing the efficiency of SG efforts. Over 28% of proteins of the *T. maritima* proteome were passed through the JCSG HT pipeline to crystal trials and, of these, 86% successfully produced protein crystals in tier 1 screening. This protein crystallization rate is considerable and is partially a consequence of the fact that only those targets that expressed at very high levels were attempted, but may also be a consequence of the increased stability of thermophilic enzymes at room temperature (Das & Gerstein, 2000; Vieille & Zeikus, 2001); the likelihood that proteins from non-thermophilic organisms will match or surpass this success rate is unknown.

The results of the native proteome screen, tier 1, show that over 75% of the commonly used crystal screening conditions are redundant and can be eliminated from future tier 1 screens without a substantial impact upon the number of distinct proteins crystallized. The tier 1 Core Screen, which contains the subset of conditions that most effectively crystallized the proteins in tier 1, contained just 67 conditions, yet still produced crystals for 86% of the tier 1 crystallized proteins. All five primary precipitant classes (high-MW PEGs, low-MW PEGs, ammonium sulfate/salts, polyalcohols and remaining organics) were represented in this Core Screen, although high-MW PEG conditions were the most prevalent. However, as expected based on the analysis of those proteins with a spectacular ability to crystallize, more high-MW PEG conditions were eliminated from the Core Screen than any other precipitant class. The results from the tier 1 screening also show that glycerol clearly inhibits crystal formation, since the addition of glycerol to 44 conditions reduced their crystallization potential by 30%. Finally, these results also show that tier 1 screening is incomplete. While this strategy does successfully identify the proteins with a high propensity to crystallize, it misses those which may require additional attention. Purity of the tier 1 samples was not optimized prior to screening and it is likely that copurified contaminants may have inhibited crystal formation in some of the tier 1 samples. In fact, four out of six samples which failed to crystallize in tier 1 did so once they were reprocessed and more extensively purified. Thus, modifications to the pipeline which enable samples that fail to crystallize in tier 1 to be reprocessed such that they are more extensively purified and/or screened against novel precipitants will result in a higher success rate.

The results of tier 2 show that the rapid screening protocol used in tier 1 was extremely successful at identifying those targets with a high propensity to crystallize, since nearly 99% of the proteins passed to tier 2 produced crystals. They also show, however, that the tier 1 screening protocol cannot be used to predict which conditions will crystallize tier 2 proteins, since the conditions which produced crystals for a given protein in tier 1 were often different from those in tier 2. Therefore, while the results from tier 1 can be used to predict native protein crystallizability, they can not be used to predict the optimal crystallization conditions for their more extensively purified selenomethionine-containing counterparts. Finally, the conditions used for screening tier 2 proteins should

[2] Supplementary material has been deposited in the IUCr electronic archive (Reference: he0318). Details for accessing these data are described at the back of the journal.

include those which contain cryoprotectant agents and perhaps should be expanded to include more of them. Over half of the JCSG structures determined to date have been obtained from crystals grown in cryoprotectant solution because harvestable crystals from such solutions are easier to mount and screen.

The conclusions drawn from these studies are now being used to develop additional procedures in the SG pipeline to maximize the number of targets that can be processed and crystallized successfully using the minimum amount of resources and time. First, the tier 1 crystal screen is being reduced from 480 conditions to 96 (since this is the smallest number of conditions which can be processed in a single pass by the crystallization robot) and will include all of the Core Screen conditions. Second, new cryoprotectant conditions are being tested for crystallization potential and added to tier 2 crystal screens. Finally, supplementary channels are being added to the pipeline to enable proteins that fail to crystallize during tier 1 to be rescreened using different processing methods. For example, constructs that do not crystallize in tier 1 will be reprocessed so they are more extensively purified prior to screening. If that fails, additional truncations, mutations and domain screens of those targets will be created and rescreened in tier 1 to identify alternate domains which may more readily crystallize (Huang *et al.*, 2002; Mateja *et al.*, 2002). Rather than make extensive modifications to purification and crystallization protocols, the approach taken here is to revisit the construct and produce one which has a greater propensity to crystallize. The strategy for deciding which types of changes should be made in various constructs for improved crystal formation are currently under development. While the two-tiered approach is clearly effective for HT SG efforts, we also believe that it can be used in smaller scale crystallographic studies. Rather than work with one or a few constructs at a time, a large number of constructs can be made and rapidly screened for crystal formation (tier 1 screening) and more extensive crystallization efforts can then be focused on only those constructs which crystallize readily under suboptimal conditions (tier 2 screening). By applying these results to both large-scale SG programs and more focused single protein studies, the rate of protein structure determination will be substantially accelerated and in turn so will our understanding of protein function and human disease.

The authors thank Duncan McRee and Dan Scheibe from Syrrx Inc. for arranging and setting up the initial crystallization experiments, respectively. These studies were supported by the NIH grant for the JCSG (GM62411); RP was supported by an NIH post-doctoral fellowship (NS11146).

References

Carter, C. W. Jr & Carter, C. W. (1979). *J. Biol. Chem.* **254**, 12219–12223.
Chayen, N. E. & Saridakis, E. (2002). *Acta Cryst.* D**58**, 921–927.
Cudney, B., Patel, S., Weisgraber, K., Newhouse, Y. & McPherson, A. (1994). *Acta Cryst.* D**50**, 414–423.
Das, R. & Gerstein, M. (2000). *Funct. Integr. Genomics*, **1**, 76–88.
Gilliland, G. L. & Ladner, J. E. (1996). *Curr. Opin. Struct. Biol.* **6**, 595–603.
Gilliland, G. L., Tung, M. & Ladner, J. (1996). *J. Res. Natl Inst. Stand. Technol.* **101**, 309–320.
Huang, M., Weissman, J. T., Wang, C., Balch, W. E. & Wilson, I. A. (2002). *Acta Cryst.* D**58**, 700–703.
Jancarik, J. & Kim, S.-H. (1991). *J. Appl. Cryst.* **24**, 409–411.
Lesley, S. A. *et al.* (2002). *Proc. Natl Acad. Sci. USA*, **99**, 11664–11669.
McPherson, A. (1990). *Eur. J. Biochem.* **189**, 1–23.
McPherson, A. (1994). *Crystallization of Biological Macromolecules.* Cold Spring Harbor Laboratory Press.
McPherson, A. Jr (1976). *J. Biol. Chem.* **251**, 6300–6303.
Mateja, A., Devedjiev, Y., Krowarsch, D., Longenecker, K., Dauter, Z., Otlewski, J. & Derewenda, Z. S. (2002). *Acta Cryst.* D**58**, 1983–1991.
Nelson, K. E. *et al.* (1999). *Nature (London)*, **399**, 323–329.
Santarsiero, B. D., Yegian, D. T., Lee, C. C., Spraggon, G., Gu, J., Scheibe, D., Uber, D. C., Cornell, E. W., Nordmeyer, R. A., Kolbe, W. F., Jin, J., Jones, A. L., Jaklevic, J. M., Schultz, P. G. & Stevens, R. C. (2002). *J. Appl. Cryst.* **35**, 278–281.
Vieille, C. & Zeikus, G. J. (2001). *Microbiol. Mol. Biol. Rev.* **65**, 1–43.
Wahl, M. C., Bourenkov, G. P., Bartunik, H. D. & Huber, R. (2000). *EMBO J.* **19**, 174–186.
Wahl, M. C., Huber, R., Marinkovic, S., Weyher-Stingl, E., Ehlert, S., Bourenkov, G. P. & Bartunik, H. D. (2000). *Biol. Chem.* **381**, 221–229.