A General Target Selection Method for Crystallographic Proteomics

Gautier Robin[1*], Nathan P. Cowieson[1], Gregor Guncar[1,2], Jade K. Forwood[1,2], Pawel Listwan[1,2,3,4], David A. Hume[1,2,3,4], Bostjan Kobe[1,2,4], Jennifer L. Martin[1,2,4], Thomas Huber[2]

[1]Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia

[2]School of Molecular and Microbial Sciences, University of Queensland, Brisbane, Australia

[3]Cooperative Research Centre for Chronic Inflammatory Diseases, University of Queensland, Brisbane, Australia

[4]ARC Special Research Centre for Functional and Applied Genomics, University of Queensland, Brisbane, Australia

*To whom correspondence should be addressed: Institute for Molecular Bioscience, the University of Queensland, St Lucia, Australia. Phone: 61 7 3346 2020. Email: g.robin@imb.uq.edu.au

Running title: Crystallography Target Selection

Abstract

Increasing the success in obtaining structures and maximizing the value of the structures determined are the two major goals of target selection in structural proteomics. This critical process consists of predicting and quantifying target properties to restrict selected candidates to those of particular interest and those that have the greatest chance of being structurally characterized. We present an efficient and flexible target selection procedure supplemented with a web-based resource that is suitable for small- to large-scale structural genomics projects that use crystallography as the major means of structure determination. Based on three criteria, biological significance, structural novelty and “crystallizability”, the approach first removes (filters) targets that do not meet minimal criteria and then ranks the remaining targets based on their “crystallizability” estimates. This novel procedure was designed to maximize selection efficiency, and its prevailing criteria categories make it suitable for a broad range of structural proteomics projects.

Key Words: High-throughput methods, protein crystallization, protein expression, protein properties, protein structure determination, sequence comparison, structural genomics, structural proteomics, target selection.

1. Introduction

The task that initiates all structural proteomics/genomics projects is choosing the targets - proteins, or regions of proteins - that will enter the structure determination process. This target selection step is particularly important because it embodies the goals of each specific project and because it can substantially influence the overall success rate of the process. Although each specific structural proteomics program may have a different biological focus, methodology and throughput (***1-6***), there are nonetheless common criteria that can be used to select targets.

Target selection criteria can be divided into three categories: (a) biological significance/impact, (b) structural novelty and (c) likelihood to crystallize. Clearly, the classification of biological significance/impact will vary from one project to another, but structural novelty and likelihood to crystallize are common criteria for most target selection procedures. Furthermore, selection criteria can be categorized into two types: filters and sorters. Filters remove those targets that possess undesirable features (those having a known structure, for example), while sorters allow the ranking of remaining targets so that the desired number of targets may be selected from the top-scoring targets.

The goal of the method described here is to provide a general approach for selecting targets in a small- to large-scale protein crystallography context.

2. Methods

2.1. *Target Selection Frame Description*

The process can be broken down into three sequential steps: automated filtering, ranking/sorting, and manual filtering (**Fig. 1**). The order of this sequence was designed to maximize efficiency of target selection while allowing the flexibility to cater for a specific focus of a structural proteomics project. This section overviews the strategy and the methods used in the target selection procedure. The following final section presents the application of the procedure.

2.1.1. Automated Filtering (Step 1)

Automated filtering represents the major selection step. It includes criteria from all three of the selection categories mentioned above (biological significance/impact, structural novelty, likelihood to crystallize). It includes all criteria that are amenable to computer automation and that are not components of the ranking parameters (listed in **2.1.2.**). Examples of criteria that can be applied at this step are presented in **Fig. 1**. The outcome of this first step is a list of targets that have passed all the selected criteria (*see* **Note 1**).

2.1.2. Ranking of Targets (Step 2)

The automated filtering will often result in more targets than can be handled in a single iteration of high-throughput protein production and structure determination. It is important therefore to be able to prioritize targets selected in step 1. In this procedure, the targets selected by the automatic filtering procedure are subsequently ranked according to their likelihood to crystallize. Recently, several groups have shown clear

correlations between crystallization success and protein properties predicted from sequence only (***7-10***).

Target selection methods take advantage of the data generated by structural genomics projects to identify correlations between protein attributes (determined by sequence analysis) and its success or failure through the expression-purification-crystallization process. This is then used to extract rules for target selection to optimize the output. One approach to detect such crystallization predictors consisted of generating distributions of various potentially relevant properties from a set of proteins (whole *Thermatoga maritima* proteome) and from the subset of those that crystallized, to analyze trends for crystallization success (***7***). The outcome was a list of crystallization predictors and target filtering strategies. Because we are interested at this stage in ranking rather than filtering, we used a similar approach to re-generate the distributions to quantify the likelihood of target crystallization, which is described below. We chose a larger and more representative initial set of proteins to represent the "whole universe" of proteins by using more than 3 million protein sequences from the non-redundant sequence database, and a non-redundant sample from the Protein Data Bank (PDB) as the subset representing the universe of successfully crystallized proteins. The normalized distributions are similar to those obtained from the *Thermatoga maritima* genome, thus validating the pertinence of both datasets. The predictive power of a given sequence characteristic is inversely proportional to the area of overlap between the global distribution and the crystallized distribution. The protein properties we use for estimating crystallization likelihood are: sequence length, predicted isoelectric point, percentage of charged residues, hydropathy, and a measure of low complexity disorder (*see* **Notes 2-4**). These have already been reported as parameters influencing crystallization success (***7-10***). We are currently in

the process of systematically testing other parameters to further improve efficiency. Finally, the likelihood estimate *p* is calculated according to:

$$p = \prod_i \left(\frac{Y_i^C}{Y_i^U} \right) \quad (1)$$

where Y_i^C represents the frequency of proteins from the crystallized subset that match parameter *i* value (with a sequence length of 250-260 residues for example) of the protein evaluated (253 residues for example); and Y_i^U represents the corresponding value from the whole “universe set of proteins". This formula can be related to a probability calculation where the ratio represents the number of successful events (crystallized proteins subset) over the whole set. The properties (e.g. length, pI, percentage of charged residues) are considered independent, which is reflected by the use of the product in equation 1. The application of the ranking step is achieved through the Web-based “UQSG Target Ranker” (*see* **2.2.2.**)**.**

2.1.3. Manual Filtering (Step 3)

The final step is the manual evaluation of each of the ranked targets. This optional step is used because some criteria may be too difficult to program and because the programed selection procedures may have weaknesses. Typically the protein description, ontology, literature searches and personal knowledge are employed to identify any specific problems in the remaining selected targets. Evaluating each target manually is time-consuming, thus this step is performed last. In truly high-throughput programs manual intervention in the target selection is generally omitted; however, in small- and medium-throughput programs additional quality control can help to focus on targets with higher biological significance and scientific impact.

2.2. Applying the Target Selection

2.2.1. Full-Length Protein or Protein Constructs?

Construct design may be considered for proteins of higher value. The constructs may be designed to pass given selection criteria (e.g. removal of transmembrane domains, deletion of low complexity regions) and to divide the protein into predicted functional domains. In either case the aim is to increase the chance of obtaining a crystal structure at the cost of limiting the structural information to a part of the protein, and lowering the protein flow rate through the pipeline (N constructs of a single protein instead of N different proteins).

Useful information to consider in designing constructs includes alternative splice variations (*see* **Note 5**), domain homology searches in combination with secondary and tertiary structure prediction (to help defining domain boundaries), and "problematic" region prediction such as trans-membrane domains (to omit in construct). **Table 1** provides suggested links for accessing this information.

In any case, the selection procedure that follows is unchanged, as it does not distinguish between full-length or designed protein constructs; the target definition is any input sequence.

2.2.2. Selecting the Appropriate Criteria

Translating the project aims into criteria is a critical and subtle process. The vast amount of information that is readily available on protein targets through bioinformatics can retrieve an enormous amount of "interesting" information that is tempting to include in the selection process. We recommend restricting the criteria at this stage only to those that are able to remove targets (*see* **Note 1**).

Fig. 1 lists example of criteria. They are divided in the three categories mentioned above, namely biological significance/impact, structural novelty, and likelihood to crystallize. Brief comments and tips are also given below.

The biological significance/impact criteria category will segregate targets based on organism origin, amino acid sequence characteristics (e.g. protein family, presence of a particular domain, sequence similarity to human orthologue) and on the results of a specific type of assay (e.g. biophysical, biochemical, clinical test). We recommend using available functional data if possible, as this can be highly valuable for biological significance and impact. For example, we use microarray data to select targets that are likely to have a role in inflammation (*see* **Note 6**). Also, if the target proteins are not from *Homo sapiens*, assessing target relevance in human biology may be assed based on sequence similarity. For example we select those proteins with a minimum of 70% sequence identity with human proteins (probable human orthologues).

The structural novelty criterion is based on sequence alignment with proteins from the PDB. The 30% identity threshold is based on the assumption that a higher sequence identity enables homology modeling, thus reducing the value of structure determination. Although the percentage of sequence identity with known structures is indicative of structural novelty, more powerful tools such as threading (**Table 1**) based on tertiary structure prediction alignment can be used where obtaining new folds is of particular concern.

The “likelihood to crystallize” section is composed of one filter (no transmembrane helix) followed by a series of parameters (sorters) used to estimate the crystallization likelihood (method detailed in **2.1.2.**). Although transmembrane sequences have a large detrimental effect on solubility and crystallization (***7***), this criterion is optional in the Web-based tool to allow for the specific study of membrane proteins.

Finally, changing the order of application of these criteria will not modify the output, however it does affect computer-processing time. Therefore we recommend applying the less time-consuming criteria first.

2.2.3. Using the Web-based Target Ranking Tool

The Web-based resource ("UQSG Target Ranker" available at http://foo.maths.uq.edu.au/~huber/UQSG/ranker.pl), is a tool for automatically selecting and sorting targets based on predicted likelihood of crystallization. The inputs are target sequences and selected criteria; the output is a list of selected targets that are ranked according to their predicted crystallization likelihood estimates. A variety of sequence-based criteria have been proposed to predict the likelihood for a protein to crystallize (***7-10***) and the five most predictive ones have been implemented in our web server (see **2.1.2.** for details). With our tool, we give the user the option to rank target sequences supplied in FASTA format according to all, or individually selected parameters.

3. Notes

1. Why filter rather than rank? The use of a ranked grading system (for example 1: "very bad", 2: "bad", 3: "OK", 4: "good", or 5: "very good") leads to target comparisons based on arbitrary grades and grade combinations, resulting in inconsistent selection of targets. Instead, we use filter-type criteria (0: "discard" or 1: "keep" output per criterion) where each criterion selects independently of the others, allowing unambiguous and precise control over the selection of target proteins. The ranking/sorting is achieved in the following step, on crystallizing likelihood estimates.

2. The length of the protein sequence has a strong influence on protein crystallization. When the size of a protein is too small, generally its thermodynamic stability is marginal and intrinsic thermal motion can inhibit crystallization. Very large proteins are also more difficult to crystallize because they are likely to exhibit higher flexibility and a reduced translational and rotational motion in solution, leading to kinetically inhibited nucleation.

3. The pI of a protein may influence crystallization success. At conditions with pH of the solution equal to the pI of the protein, the net charge on the protein is zero and as a result no overall electrostatic repulsion between protein molecules is present. Standard protein crystallization screens contain conditions optimized to crystallize a "typical" protein with only weakly repulsive (effective) interactions in stock solution. Given a standard (non-optimized) protein buffer (typically pH 7.0-7.5), choosing proteins within the appropriate pI range, and thus appropriate effective interactions, can be beneficial.

4. Another criterion based on similar rationale as the pI is the percentage of charged amino acids in the sequence. Other physical properties of a protein that are known to influence crystallization, and thus can be beneficial when taken into account to rank targets, are the number of residues in regions of low complexity (associated with disordered regions), and the overall hydropathy.

5. It is useful to consider splice isoforms; nature may have already designed the construct for us, although one must keep in mind that splice forms may have different functions or may not be functional at the protein level (***11-13***))

6. Functional data may add great value in terms of biological significance/impact. Microarray technology is particularly suitable for obtaining functional data in high throughput. The "> 3 fold transcriptional regulation upon stimulation" criterion in **Fig. 1**

is an example of a criterion used for such experiment, although the threshold is specific to the data (for more details see a separate chapter in this volume (Meng et al.: Overview of the pipeline for structural and functional characterization of macrophage proteins at the University of Queensland).

7. A note of caution. While target selection can help improving the apparent success rate of structure determination, it simultaneously introduces a strong bias to narrow the diversity of proteins for which the structures are being determined. In an uncontrolled extreme, this can lead to a circular process, in which empirical target selection will be based on previously successful structure determination, but future structure databases will be contain to a large proportion proteins that have been selected with this same bias.

8. Interestingly, despite many improvements in approaches to select new targets for structural genomics pipelines over the last decade, the success of these methods in practice is not well established. Partly, this is due to the difficulty to delineate effects as a result of changes in experimental procedures from improvements as a result of targeted selection itself. With experimental procedures becoming more established over time, it will be interesting to monitor future development of target selection methods.

Acknowledgments

The authors would like to thank Tim Ravasi, Munish Puri, Ian Ross, Tom Alber and all the members of the macrophage protein group for their feedback and advice. This work was supported by an Australian Research Council (ARC) grant to JLM and BK. BK is an ARC Federation Fellow and a National Health and Medical Research Council

Honorary Research Fellow, and NC an Australian Synchrotron Research Program Fellow.

References

1. Berry, I. M., Dym, O., Esnouf, R. M., Harlos, K., Meged, R., Perrakis, A., Sussman, J. L., Walter, T. S., Wilson, J., and Messerschmidt, A. (2006) SPINE high-throughput crystallization, crystal imaging and recognition techniques: current state, performance analysis, new technologies and future aspects. *Acta Crystallogr. D Biol. Crystallogr.* **62,** 1137-1149.
2. Bonanno, J. B., Almo, S. C., Bresnick, A., Chance, M. R., Fiser, A., Swaminathan, S., Jiang, J., Studier, F. W., Shapiro, L., Lima, C. D., Gaasterland, T. M., Sali, A., Bain, K., Feil, I., Gao, X., Lorimer, D., Ramos, A., Sauder, J. M., Wasserman, S. R., Emtage, S., D'Amico, K. L., and Burley, S. K. (2005) New York-Structural GenomiX Research Consortium (NYSGXRC): a large scale center for the protein structure initiative. *J. Struct. Funct. Genomics* **6,** 225-232.
3. Busso, D., Poussin-Courmontagne, P., Rose, D., Ripp, R., Litt, A., Thierry, J. C., and Moras, D. (2005) Structural genomics of eukaryotic targets at a laboratory scale. *J. Struct. Funct. Genomics* **6,** 81-88.
4. Lundstrom, K., Wagner, R., Reinhart, C., Desmyter, A., Cherouati, N., Magnin, T., Zeder-Lutz, G., Courtot, M., Prual, C., Andre, N., Hassaine, G., Michel, H., Cambillau, C., and Pattus, F. (2006) Structural genomics on membrane proteins: comparison of more than 100 GPCRs in 3 expression systems. *J. Struct. Funct. Genomics*.
5. Moreland, N., Ashton, R., Baker, H. M., Ivanovic, I., Patterson, S., Arcus, V. L., Baker, E. N., and Lott, J. S. (2005) A flexible and economical medium-

throughput strategy for protein production and crystallization. *Acta. Crystallogr. D Biol. Crystallogr.* **61,** 1378-1385.

6. Su, X. D., Liang, Y., Li, L., Nan, J., Brostromer, E., Liu, P., Dong, Y., and Xian, D. (2006) A large-scale, high-efficiency and low-cost platform for structural genomics studies. *Acta Crystallogr. D Biol. Crystallogr.* **62,** 843-851.
7. Canaves, J. M., Page, R., Wilson, I. A., and Stevens, R. C. (2004) Protein biophysical properties that correlate with crystallization success in Thermotoga maritima: maximum clustering strategy for structural genomics. *J. Mol. Biol.* **344,** 977-991.
8. Goh, C. S., Lan, N., Douglas, S. M., Wu, B., Echols, N., Smith, A., Milburn, D., Montelione, G. T., Zhao, H., and Gerstein, M. (2004) Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. *J. Mol. Biol.* **336,** 115-130.
9. Rupp, B., and Wang, J. (2004) Predictive models for protein crystallization. *Methods* **34,** 390-407.
10. Smialowski, P., Schmidt, T., Cox, J., Kirschner, A., and Frishman, D. (2006) Will my protein crystallize? A sequence-based predictor. *Proteins* **62,** 343-355.
11. Homma, K., Kikuno, R. F., Nagase, T., Ohara, O., and Nishikawa, K. (2004) Alternative splice variants encoding unstable protein domains exist in the human brain. *J. Mol. Biol.* **343,** 1207-1220.
12. Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T. A., and Soreq, H. (2005) Function of alternative splicing. *Gene* **344,** 1-20.
13. Takeda, J., Suzuki, Y., Nakao, M., Barrero, R. A., Koyanagi, K. O., Jin, L., Motono, C., Hata, H., Isogai, T., Nagai, K., Otsuki, T., Kuryshev, V., Shionyu, M., Yura, K., Go, M., Thierry-Mieg, J., Thierry-Mieg, D., Wiemann, S., Nomura,

N., Sugano, S., Gojobori, T., and Imanishi, T. (2006) Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. *Nucleic Acids Res.* **34,** 3917-3928.

14. Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K. O., Barrero, R. A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M., Yura, K., Miyazaki, S., Ikeo, K., Homma, K., Kasprzyk, A., Nishikawa, T., Hirakawa, M., Thierry-Mieg, J., Thierry-Mieg, D., Ashurst, J., Jia, L., Nakao, M., Thomas, M. A., Mulder, N., Karavidopoulou, Y., Jin, L., Kim, S., Yasuda, T., Lenhard, B., Eveno, E., Suzuki, Y., Yamasaki, C., Takeda, J., Gough, C., Hilton, P., Fujii, Y., Sakai, H., Tanaka, S., Amid, C., Bellgard, M., Bonaldo Mde, F., Bono, H., Bromberg, S. K., Brookes, A. J., Bruford, E., Carninci, P., Chelala, C., Couillault, C., de Souza, S. J., Debily, M. A., Devignes, M. D., Dubchak, I., Endo, T., Estreicher, A., Eyras, E., Fukami-Kobayashi, K., Gopinath, G. R., Graudens, E., Hahn, Y., Han, M., Han, Z. G., Hanada, K., Hanaoka, H., Harada, E., Hashimoto, K., Hinz, U., Hirai, M., Hishiki, T., Hopkinson, I., Imbeaud, S., Inoko, H., Kanapin, A., Kaneko, Y., Kasukawa, T., Kelso, J., Kersey, P., Kikuno, R., Kimura, K., Korn, B., Kuryshev, V., Makalowska, I., Makino, T., Mano, S., Mariage-Samson, R., Mashima, J., Matsuda, H., Mewes, H. W., Minoshima, S., Nagai, K., Nagasaki, H., Nagata, N., Nigam, R., Ogasawara, O., Ohara, O., Ohtsubo, M., Okada, N., Okido, T., Oota, S., Ota, M., Ota, T., Otsuki, T., Piatier-Tonneau, D., Poustka, A., Ren, S. X., Saitou, N., Sakai, K., Sakamoto, S., Sakate, R., Schupp, I., Servant, F., Sherry, S., Shiba, R., Shimizu, N., Shimoyama, M., Simpson, A. J., Soares, B., Steward, C., Suwa, M., Suzuki, M., Takahashi, A., Tamiya, G., Tanaka, H., Taylor, T., Terwilliger, J. D., Unneberg,

P., Veeramachaneni, V., Watanabe, S., Wilming, L., Yasuda, N., Yoo, H. S., Stodolsky, M., Makalowski, W., Go, M., Nakai, K., Takagi, T., Kanehisa, M., Sakaki, Y., Quackenbush, J., Okazaki, Y., Hayashizaki, Y., Hide, W., Chakraborty, R., Nishikawa, K., Sugawara, H., Tateno, Y., Chen, Z., Oishi, M., Tonellato, P., Apweiler, R., Okubo, K., Wagner, L., Wiemann, S., Strausberg, R. L., Isogai, T., Auffray, C., Nomura, N., Gojobori, T., and Sugano, S. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* **2,** e162.

15. Fink, J. L., Aturaliya, R. N., Davis, M. J., Zhang, F., Hanson, K., Teasdale, M. S., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Teasdale, R. D. (2006) LOCATE: a mouse protein subcellular localization database. *Nucleic Acids Res.* **34,** D213-217.

16. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., Weber, R. J., Haussler, D., and Kent, W. J. (2003) The UCSC Genome Browser Database *Nucleic Acids Res.* **31,** 51-54.

17. Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Liebert, C. A., Liu, C., Lu, F., Marchler, G. H., Mullokandov, M., Shoemaker, B. A., Simonyan, V., Song, J. S., Thiessen, P. A., Yamashita, R. A., Yin, J. J., Zhang, D., and Bryant, S. H. (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res.* **33,** D192-196.

18. Geer, L. Y., Domrachev, M., Lipman, D. J., and Bryant, S. H. (2002) CDART: protein homology by domain architecture. *Genome Res.* **12,** 1619-1623.

19. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389-3402.

20. Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol. Biol.* **347,** 827-839.

21. Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21,** 3433-3434.

22. Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure* **11,** 1453-1459.

23. Linding, R., Russell, R. B., Neduva, V., and Gibson, T. J. (2003) GlobPlot: Exploring protein sequences for globularity and disorder *Nucleic Acids Res.* **31,** 3701-3708.

24. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567-580.

25. Bendtsen, J. D., Nielsen, H., von Heijne, G., and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340,** 783-795.

26. Cuff, J. A., and Barton, G. J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* **40,** 502-511.

27. Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M., and Barton, G. J. (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics* **14,** 892-893.

28. Rost, B., and Liu, J. (2003) The PredictProtein server. *Nucleic Acids Res.* **31,** 3300-3304.

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

30. Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299,** 499-520.

31. Shi, J., Blundell, T. L., and Mizuguchi, K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* **310,** 243-257.

32. Torda, A. E., Procter, J. B., and Huber, T. (2004) Wurst: a protein threading server with a structural scoring function, sequence profiles and optimized substitution matrices. *Nucleic Acids Res.* **32,** W532-535.

Table 1.

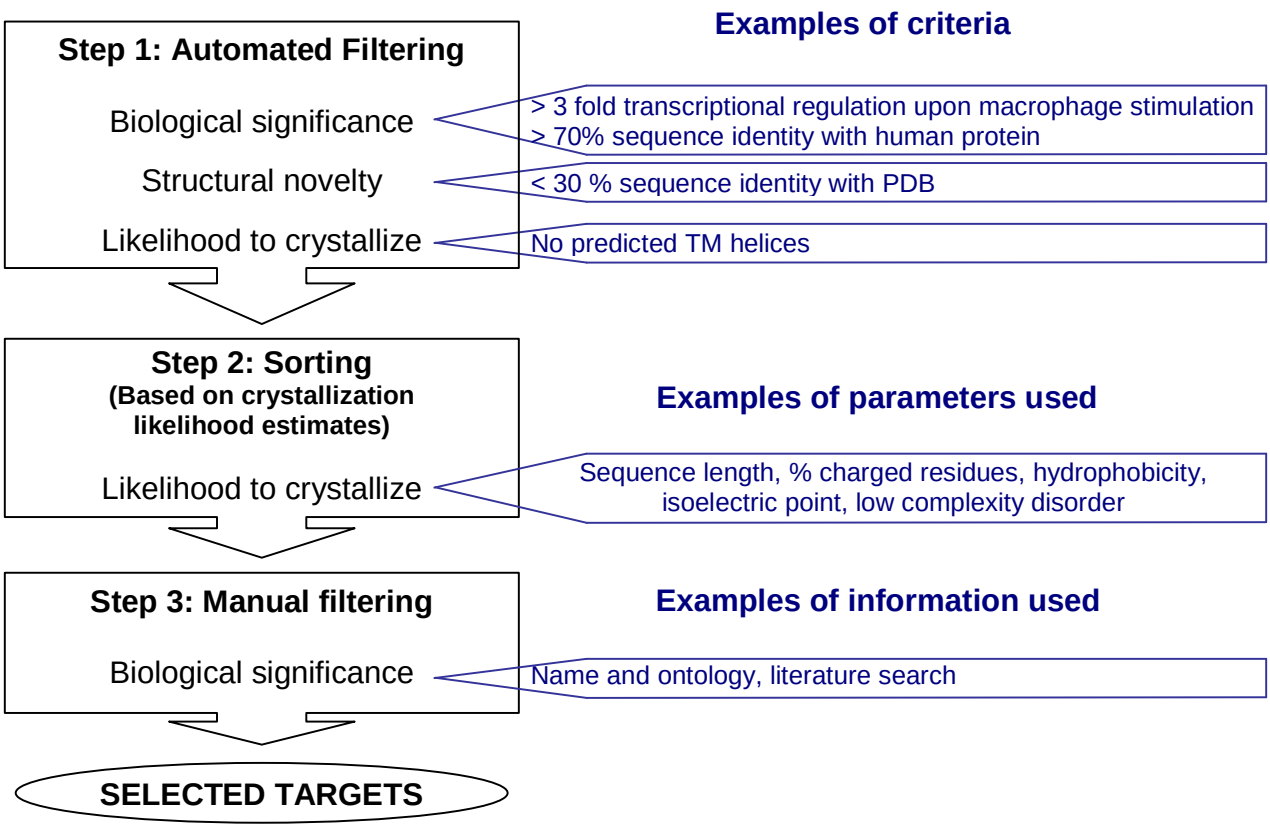
Useful links for protein construct design

	Software	URL
Splice variant database	H-InvDB (***14***)	http://hinvdb.ddbj.nig.ac.jp/ahg-db/index.jsp
	LOCATE (***15***)	http://locate.imb.uq.edu.au/
	Macrophages.com	http://www.macrophages.com/bioinfoweb/
	Genome Browser (***16***)	http://genome.ucsc.edu/cgi-bin/hgGateway
Domain homology search	RPSBLAST (***17***)	http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
	CDART (***18***)	http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?
	BLAST (***19***)	http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?
"Problematic" region prediction	*Unstructured regions*	
	IUPred (***20, 21***)	http://iupred.enzim.hu/index.html
	DisEMBL (***22***)	http://dis.embl.de/
	GlobPlot (***23***)	http://globplot.embl.de/
	(for a more complete list see the chapter by Dosztáni and Tompa in this book)	
	Transmembrane regions	
	TMHMM (***24***)	http://www.cbs.dtu.dk/services/TMHMM-2.0/
	Signal sequence	
	SignalP (***25***)	http://www.cbs.dtu.dk/services/SignalP/
Secondary and tertiary structure prediction	*Secondary structure prediction*	
	JPREP (***26, 27***)	http://www.compbio.dundee.ac.uk/~www-jpred/
	PredictProtein (***28***)	http://www.predictprotein.org/
	PSIPRED (***29***)	http://bioinf.cs.ucl.ac.uk/psipred/
	Tertiary structure prediction	
	Phyre (***30***)	http://www.sbg.bio.ic.ac.uk/~phyre
	FUGUE (***31***)	http://www-cryst.bioc.cam.ac.uk/fugue/
	WURST (***32***)	http://www.zbh.uni-hamburg.de/wurst/

Figure Legend

Fig. 1. A schematic diagram illustrating the target selection procedure.

Figure 1.


Examples of criteria
Step 1: Automated Filtering
Biological significance
Structural novelty
Likelihood to crystallize
> 3 fold transcriptional regulation upon macrophage stimulation
> 70% sequence identity with human protein
< 30 % sequence identity with PDB
No predicted TM helices
Step 2: Sorting
(Based on crystallization likelihood estimates)
Likelihood to crystallize
Examples of parameters used
Sequence length, % charged residues, hydrophobicity, isoelectric point, low complexity disorder
Step 3: Manual filtering
Biological significance
Examples of information used
Name and ontology, literature search
SELECTED TARGETS