Published in final edited form as:
Nat Methods. 2007 December ; 4(12): 1019–1021. doi:10.1038/nmeth1118.

In situ proteolysis for protein crystallization and structure determination

Aiping Dong[1], **Xiaohui Xu**[2], **Aled M Edwards**[1,2], **Midwest Center for Structural Genomics**[2,3], and **Structural Genomics Consortium**[1,3]

[1]Structural Genomics Consortium, University of Toronto, 100 College Street, Toronto, Ontario M5G 1L5, Canada.

[2]Midwest Center for Structural Genomics, University of Toronto, 112 College Street, Toronto, Ontario M5G 1L6, Canada.

Abstract

We tested the general applicability of *in situ* proteolysis to form protein crystals suitable for structure determination by adding a protease (chymotrypsin or trypsin) digestion step to crystallization trials of 55 bacterial and 14 human proteins that had proven recalcitrant to our best efforts at crystallization or structure determination. This is a work in progress; so far we determined structures of 9 bacterial proteins and the human aminoimidazole ribonucleotide synthetase (AIRS) domain.

Analysis of large-scale structural genomics studies (http://targetdb.pdb.org) shows that, of all proteins that enter crystallization trials, two-thirds will not crystallize and half of those that do crystallize cannot be optimized to form suitable crystals for structure determination —a final success rate of ~15% from purified protein to structure. Given the resources that are invested to generate a purified, concentrated protein and to perform extensive crystallization trials, this level of attrition is of considerable concern.

For decades, scientists have exploited the fact that fragments or domains of proteins often either crystallize better, or form more well-diffracting crystals, compared with the intact protein. In the earliest instances, protein fragments had been prepared in large scale from the purified protein and then crystallized. This method proved successful[1,2], but its widespread use was limited by the need to purify large amounts of the intact protein and the difficulty in purifying protease-derived fragments in a homogeneous form. These limitations had been

Correspondence should be addressed to A.E. (aled.edwards@utoronto.ca)..
[3]Complete lists of authors appear at the end of this paper.

The complete list of authors is as follows:
Midwest Center for Stuctural Genomics. Changsoo Chang, Maksymilian Chruszcz, Marianne Cuff, Marcin Cymborowski, Rosa Di Leo, Olga Egorova, Elena Evdokimova, Ekaterina Filippova, Jun Gu, Jennifer Guthrie, Alexandr Ignatchenko, Andrzej Joachimiak, Natalie Klostermann, Youngchang Kim, Yuri Korniyenko, Wladek Minor, Qiuni Que, Alexei Savchenko, Tatiana Skarina, Kemin Tan, Alexander Yakunin, Adelinda Yee, Veronica Yim, Rongguang Zhang, Hong Zheng
Structural Genomics Consortium. Masato Akutsu, Cheryl Arrowsmith, George V Avvakumov, Alexey Bochkarev, Lars-Göran Dahlgren, Sirano Dhe-Paganon, Slav Dimov, Ludmila Dombrovski, Patrick Finerty Jr., Susanne Flodin, Alex Flores, Susanne Gräslund, Martin Hammerström, Maria Dolores Herman, Bum-Soo Hong, Raymond Hui, Ida Johansson, Yongson Liu, Martina Nilsson, Lyudmila Nedyalkova, Pär Nordlund, Tomas Nyman, Jinrong Min, Hui Ouyang, Hee-won Park, Chao Qi, Wael Rabeh, Limin Shen, Yang Shen, Deepthi Sukumard, Wolfram Tempel, Yufeng Tong, Lionel Tresagues, Masoud Vedadi, John R Walker, Johan Weigelt, Martin Welin, Hong Wu, Ting Xiao, Hong Zeng, Haizhong Zhu

Note: Supplementary information is available on the Nature Methods website.

overcome with the advent of recombinant protein expression and the development of biological mass spectrometry. The protease fragments of the protein of interest could be prepared in small scale by limited proteolysis and their exact masses determined using mass spectrometry[3–6]. By cloning the region(s) of the gene that corresponded to the protease-resistant domain(s), the fragment(s) could be expressed in recombinant form, purified to homogeneity and then crystallized. The same principle is also used to identify well-behaved fragments of proteins with known domain structure; in this case the strategy is to identify the approximate boundaries of the domain of interest using sequence alignment, to screen many recombinant versions of this domain for expression (differing slightly in their N- and C-terminal boundaries), and then to select the fragment(s) that can be produced in soluble form for purification and crystallization (S. Gräslund *et al.*; unpublished data.). These methods are now widely used to generate samples for protein crystallography.

Proteolytic fragments of proteins have also been crystallized serendipitously. In most cases, the purified protein or the crystallization solution had been contaminated with trace amounts of protease and the proteolysis occurred during crystallization. There are many well-characterized cases of this occurring, for example, references 7 and 8, and doubtless there have been many more undocumented examples. Given the historical success of serendipitous proteolysis, recent efforts have explored the possibility of purposely adding trace amounts of a purified protease to the crystallization solution. Some of these experiments have proven successful[9–13], suggesting that this approach may be more generally useful than had been appreciated previously.

Here we explored the efficacy of *in situ* proteolysis by incubating trace amounts of chymotrypsin in crystallization trials of 55 different bacterial proteins whose structures had not been determined previously. Chymotrypsin has a preference for hydrophobic residues, which are likely to be less frequent than the sites for other proteases. All bacterial proteins were appended with a hexahistidine tag and a recognition site for the TEV protease (MGSSHHHHHHSSGRENLYFQG[14] or MGSSHHHHHHSSGRENLYFQGH). This hexahistidine tag contains chymotrypsin cleavage sites (Tyr and Phe). Of the 55 proteins, 20 had previously failed to crystallize either in the presence or absence of the hexahistidine tag in screens of >182 conditions and several rounds of refinement (data not shown). The other 35 proteins had formed crystals that were unsuitable for structure determination either because they diffracted to low resolution or because the diffraction properties were poor (data not shown).

We repurified the 55 different bacterial proteins with hexahistidine tags, concentrated them, incubated them with chymotrypsin (1:100 w/w) and screened them for crystallization at room temperature (20–25 °C) in 96 conditions (Supplementary Table 1 and Supplementary Methods online). For some of the proteins, while reproducing the crystals, we also used alternate concentrations of chymotrypsin (as little as 1:10,000 wt/wt; data not shown). To test whether useful crystals could be obtained by *in situ* proteolysis using other proteases, we also treated some of the recalcitrant proteins with trypsin. To date, of the 20 proteins that never crystallized, 9 formed crystals, and we determined structures for two of these proteins (Table 1, Supplementary Fig. 1a,b and Supplementary Results online). Of the 35 proteins that had formed crystals of poor quality, 30 generated crystals with chymotrypsin or trypsin treatment. We determined six structures using crystals obtained with chymotrypsin treatment (Table 2, Supplementary Fig. 1c–h and Supplementary Results). We also determined one structure using crystals obtained with trypsin treatment (*Agrobacterium tumefaciens* protein ATU0434; Supplementary Fig. 1i and Supplementary Results); this protein had produced poorly diffracting microcrystals in the presence and absence of chymotrypsin.

Within the crystals, we mapped the proteolytic sites using mass spectrometry; in all cases, the N- and/or the C-terminal regions were trimmed (Supplementary Table 2 online). Although it is possible that an internal loop might have been digested[5,12,15], we did not observe this in the small sample set that we studied. The analysis of the efficacy of chymotrypsin and trypsin in promoting crystallization of the 55 sample proteins is a work in progress; we have not adequately investigated all the new crystals.

It is possible, even likely, that the proteins whose crystallization properties were improved using *in situ* proteolysis might also have been successfully crystallized with other methods, such as the use of multiple expression constructs. It is also possible that the fragment of a protein that would have the best crystallization properties might have not been able to be expressed in recombinant form; indeed it is often observed that unstructured N- or C-terminal extensions are required for recombinant expression. In these cases, the use of *in situ* proteolysis would appear advantageous.

To explore whether *in situ* proteolysis may prove efficacious even for those proteins for which many different constructs had been explored, we added chymotrypsin to crystallization trials of 14 human proteins that had resisted crystallization or structure determination, despite attempts to purify and crystallize an average of 16 different constructs per target. We generated crystals for 2 of the 4 targets that had never crystallized and generated new crystal forms for 4 of the 10 proteins that had crystallized previously (data not shown). Although a work in progress, we determined one structure, of the AIRS domain of the human glycinamide ribonucleotide synthetase (GART). GART comprises three domains; only the structure of the middle AIRS domain had not been determined previously. We had designed many expression constructs of the human AIRS domain based on a structure-based sequence alignment with the homologous bacterial protein, but none of these constructs had produced a soluble protein (data not shown). We incubated the full-length GART, which was soluble, with chymotrypsin and by good fortune, the N- and C-terminal domains (two-thirds of the protein) were removed, permitting crystallization and structure determination of the AIRS domain (Protein Data Bank (PDB) code 2V9Y; Supplementary Fig. 1j and Supplementary Results). Thus, it may be strategic to identify the versions of the protein that express, purify or concentrate the best, even though they might possess predicted unstructured regions, and then crystallize this extended form of the protein in the presence and absence of protease.

We used primarily a single concentration of protease. Although this has the clear benefit of simplicity, the enzyme activity is doubtless far from optimal in most of the crystallization conditions tested. It is possible that higher success rates could be achieved by optimizing the amount of the protease in each of the crystallization buffers, or by using other proteases. We also explored only one temperature of crystallization; this parameter will clearly affect the reaction.

As a caveat, we noticed that crystals grown by this method are sometimes difficult to reproduce. On occasion, a new titration of protease was required to generate the crystals, and in two instances (HP0029 and Atu0899), we were unable to repeatedly generate crystals of the quality used to determine the structure. Clearly, the procedure would benefit from standardization of the process and exploration of its versatility. We have not tested tags other than hexahistidine or the potential utility of *in situ* proteolysis for membrane proteins and protein complexes.

Using existing strategies, almost 85% of the proteins entering crystallization trials do not generate crystals suitable for structure determination. Even if our lower estimates of the success rate of *in situ* proteolysis (9 new structures for 55 recalcitrant proteins) is achieved

in larger studies, the method should substantially increase the amount of structures determined.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by US National Institutes of Health grant GM074942, by the US Department of Energy, Office of Biological and Environmental Research, under contract DE-AC02-06CH11357, and by the Structural Genomics Consortium, which is a registered charity (number 1097737) that receives funds from the Canadian Institutes for Health Research, the Canadian Foundation for Innovation, Genome Canada through the Ontario Genomics Institute, GlaxoSmithKline, Karolinska Institutet, the Knut and Alice Wallenberg Foundation, the Ontario Innovation Trust, the Ontario Ministry for Research and Innovation, Merck & Co., Inc., the Novartis Research Foundation, the Swedish Agency for Innovation Systems, the Swedish Foundation for Strategic Research and the Wellcome Trust.

References

1. Huber R, Deisenhofer J, Colman PM, Matsushima M, Palm W. Nature. 1976; 264:415–420. [PubMed: 1004567]
2. Jurnak F, McPherson A, Wang AH, Rich A. J. Biol. Chem. 1980; 255:6751–6757. [PubMed: 6993478]
3. Bochkarev A, et al. Cell. 1995; 83:39–46. [PubMed: 7553871]
4. Cohen SL, Ferre-D'Amare AR, Burley SK, Chait BT. Protein Sci. 1995; 4:1088–1099. [PubMed: 7549873]
5. Dong A, et al. Nucleic Acids Res. 2001; 29:439–448. [PubMed: 11139614]
6. Koth CM, Orlicky SM, Larson SM, Edwards AM. Methods Enzymol. 2003; 368:77–84. [PubMed: 14674269]
7. Campbell EA, et al. Mol. Cell. 2002; 9:527–539. [PubMed: 11931761]
8. Sawaya MR, Pelletier H, Kumar A, Wilson SH, Kraut J. Science. 1994; 264:1930–1935. [PubMed: 7516581]
9. Bai Y, Auperin TC, Tong L. Acta Crystallogr. 2007; 63:135–138.
10. Gaur RK, Kupper MB, Fischer R, Hoffmann KM. Acta Crystallogr. 2004; 60:965–967.
11. Johnson S, et al. Acta Crystallogr. 2006; 62:865–868.
12. Mandel CR, Gebauer D, Zhang H, Tong L. Acta Crystallogr. 2006; 62:1041–1045.
13. Taneja B, Patel A, Slesarev A, Mondragon A. EMBO J. 2006; 25:398–408. [PubMed: 16395333]
14. Zhang RG, et al. Structure. 2001; 9:1095–1106. [PubMed: 11709173]
15. Xiang S, Usunow G, Lange G, Busch M, Tong L. J. Biol. Chem. 2007; 282:2676–2682. [PubMed: 17135236]

Table 1

In situ proteolysis of proteins that had failed to crystallize previously

Expasy ID	Fragment	GI	Protease	Crystals	Diffraction	Structure progress
SCO6256	1–245	gi|21224577	C, T	C	2.9 Å, C	Solved; PDB code 2RA5
SCO4942	1–226	gi|21223315	C	C	2.9 Å, C	Solved; PDB code 2PZ9
SCO1917	1–197	gi|21220404	C, T	T	2.6 Å, T	Solved; inrefinement
SAV5583	1–355	gi|29832126	C, T	C, T	5.0 Å, C	
SCO5046	1–125	gi|21223419	C	C	NA	
SCO2368	1–191	gi|21220836	C	C	NA	
SCO2532	1–359	gi|21220992	C, T	C, T	NA	
SCO7200	1–181	gi|21225478	C	C	NA	
SAV5804	1–191	gi|29832347	C	C	NA	
SCO0485	1–203	gi|21219023	C, T	No		
SCO0641	1–191	gi|32141118	C, T	No		
SCO3261	1–431	gi|21221694	C	No		
SCO4215	1–252	gi|21222611	C, T	No		
SCO5469	1–455	gi|21223826	C	No		
SCO3979	1–196	gi|21222383	C, T	No		
SCO1718	1–234	gi|21220212	C, T	No		
SCO6792	1–198	gi|21225085	C	No		
SCO3367	1–195	gi|21221796	C, T	No		
SCO6778	1–228	gi|21225071	C, T	No		
Atu1785	1–167	gi|17935678	C, T	No		

Structures are shown in Supplementary Figure 1a,b. Proteases used: chymotrypsin (C) or trypsin (T). NA; the crystals were too small to be tested by X-ray diffraction. ID, identifier; GI, genInfo identifier.

Table 2

In situ proteolysis of proteins that had previously formed poor crystals

Expasy ID	Fragment	GI	Protease	Crystals	Diffraction	Structure progress
NE2398	1–146	gi|30250323	C	C	1.8 Å, C	Solved: PDB code 2RC3
Atu0899	1–311	gi|17934807	C	C	2.0 Å, C	Solved: PDB code 2R8W
Atu2452	1–247	gi|17936334	C	C	2.9 Å, C	Solved: PDB code 2R8B
Atu0870	1–256	gi|17934778	C	C	2.1 Å, C	Solved: PDB code 2P35
HP0029	1–218	gi|15644662	C	C	2.0 Å, C	Solved: PDB code 2QMO
Atu0299	1–198	gi|17934215	C	C	2.1 Å, C	Solved: PDB code 2QNI
Atu0434	1–370	gi|17934348	C, T	C, T	2.7 Å, T	Solved: PDB code 2R9Q
Atu0238	1–297	gi|17934154	C, T	C, T	2.0 Å, C	Data collected, crystal twinned
Atu0418	1–293	gi|17934332	C, T	C, T	2.9 Å, C	Data collected, multiple crystals
Atu1003	1–176	gi|17934911	C, T	C, T	2.9 Å, C	Data collected, multiple crystals
Atu0443	1–231	gi|17934356	C, T	C	3.0 Å, C	Data collected, weak anomalous signal
Atu1358	1–169	gi|17935258	C	C	2.4 Å, C	Waiting for data collection
Atu0854	1–164	gi|17934762	C	C	2.4 Å, C	Waiting for data collection
Atu2438	1–215	gi|17936320	C, T	C, T	2.5 Å, C	Waiting for data collection
SCO1463	1–360	gi|21219965	C	C	7.8 Å, C	
rha05074	95–247	gi|111018552	C	C	NA	
Atu2132	1–335	gi|17936016	C	C	NA	
Tbd_0951	98–275	gi|74316969	C, T	C, T	NA	
Atu2821	1–230	gi|17936696	C, T	C	NA	
SCO0155	1–207	gi|21218712	C, T	C	NA	
SCO5636	1–259	gi|21223987	C, T	C	NA	
Atu2572	1–393	gi|17936447	C, T	C	NA	
SCO7694	1–190	gi|21225954	C, T	C	NA	
Atu1730	1–281	gi|17935623	C	C	NA	
SAV6650	1–334	gi|29833192	C, T	C	NA	
SCO3259	1–259	gi|21221692	C	C	NA	
rha02449	2–142	gi|111017506	C, T	C, T	NA	
Atu2110	1–413	gi|17935994	C, T	No		
SCO0775	1–149	gi|21219298	C	No		

Expasy ID	Fragment	GI	Protease	Crystals	Diffraction	Structure progress
SCO7645	1–237	gi|21225906	C, T	No		
Atu2345	1–348	gi|17936228	C, T	No		
SCO5231	1–254	gi|21223599	C	No		
Atu1214	1–229	gi|17935119	C, T	C	NA	
SCO3315	1–230	gi|21221746	C, T	C	NA	
O31761	76–241	gi|2634053	C, T	T	NA	

Structures are shown in Supplementary Figure 1c–i. Proteases used: chymotrypsin (C) or trypsin (T). NA; the crystals were too small to be tested by X-ray diffraction.