Chikayama et al. BMC Bioinformatics 2010, 11:113
http://www.biomedcentral.com/1471-2105/11/113

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Mathematical model for empirically optimizing large scale production of soluble protein domains

Eisuke Chikayama[1*], Atsushi Kurotani[1], Takanori Tanaka[1], Takashi Yabuki[1], Satoshi Miyazaki[1,2], Shigeyuki Yokoyama[1,2], Yutaka Kuroda[3*]

Abstract

Background: Efficient dissection of large proteins into their structural domains is critical for high throughput proteome analysis. So far, no study has focused on mathematically modeling a protein dissection protocol in terms of a production system. Here, we report a mathematical model for empirically optimizing the cost of large-scale domain production in proteomics research.

Results: The model computes the expected number of successfully producing soluble domains, using a conditional probability between domain and boundary identification. Typical values for the model's parameters were estimated using the experimental results for identifying soluble domains from the 2,032 Kazusa HUGE protein sequences. Among the 215 fragments corresponding to the 24 domains that were expressed correctly, 111, corresponding to 18 domains, were soluble. Our model indicates that, under the conditions used in our pilot experiment, the probability of correctly predicting the existence of a domain was 81% (175/215) and that of predicting its boundary was 63% (111/175). Under these conditions, the most cost/effort-effective production of soluble domains was to prepare one to seven fragments per predicted domain.

Conclusions: Our mathematical modeling of protein dissection protocols indicates that the optimum number of fragments tested per domain is actually much smaller than expected *a priori*. The application range of our model is not limited to protein dissection, and it can be utilized for designing various large-scale mutational analyses or screening libraries.

Background

Comprehensive elucidation of the functional and structural units present in the proteome is the ultimate goal in proteomics research, and it is expected to provide basic data for a rational understanding of complex biological systems. As proteomics studies are being pursued [1-5], the development of efficient methodologies for dissecting long protein sequences into their domains is becoming critical. This is because biologically important proteins are often large and are thus difficult to express, purify and characterize in a high throughput manner [6].

Experimental approaches for dissecting proteins are usually based on limited proteolysis, which has been used to explore protein domain boundaries [7]. Although experimental protein dissection methods have been extended to high throughput protocols [8-10], they remain essentially expensive and time-consuming.

Computer-aided protein dissection approaches are relatively inexpensive, and thus represent promising methodologies that have practical values in high throughput proteomics research. The strategies for predicting novel domain regions, without sequence similarity to domain databases, can be categorized into two classes. The first strategy aims at directly predicting domain regions by analyzing various sequence properties of the foldable region (e.g., see Refs [11,12]). The second strategy is to first predict the location of the domain boundaries and then use this information to infer the domain's position (e.g., see Refs [13,14]). Both strategies are essential to efficiently identify novel protein domains.

* Correspondence: chika@psc.riken.jp; ykuroda@cc.tuat.ac.jp
[1]Genomic Sciences Center RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan
[3]Department of Biotechnology and Life Science, Faculty of Technology, Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei, Tokyo 184-0012, Japan

Proteomics projects require the identification of soluble and well behaved proteins enabling a fast structural/functional analysis [15]. Solubility is an important criterion strongly reflecting a protein's suitability for biophysical characterization. It can be readily monitored, and solubility assays are thus applied to large-scale studies [16]. Furthermore, when solubility is used to assess domain dissection experiments, it appears that a large fraction of soluble fragments are indeed well folded as assessed by NMR [17,18].

So far, reports of high throughput protein domain production protocols have mainly focused on their development and on optimizing individual experimental steps of the protocols. No study has mathematically modelled a protein domain production protocol in terms of a production system, and thus substantial room for cost-optimization through mathematical modeling remains available. In this report, we present a mathematical model for empirically optimizing large-scale protein domain production. Our model conceptually divides domain predictions into the prediction of the domain region and its boundary, and it computes the expected number of successfully produced soluble domains, using a conditional probability between these two events. We estimated the model parameters using the experimental results from a computer-aided identification of novel soluble protein domains from Kazusa HUGE protein sequences, in which 436 fragments, encoding 36 novel putative domains with slightly different domain boundaries, were expressed by using an E-coli-based cell-free system, and their solubilities were assessed with SDS-PAGE gels.

Results and Discussion

Mathematical model of protein dissection

In our mathematical model, the prediction of a protein domain is conceptually divided into two steps: A first step that predicts the existence of a domain, and a second step that predicts its boundaries (or termini). "Fragments" are domain fragments with specific termini residues, and each fragment is either soluble or insoluble. Soluble fragments and insoluble fragments belong to the soluble (S) and insoluble set (S^c), respectively (Figure 1a). S and S^c are mutually exclusive sets, and S^c is the complement of S. We define a "soluble domain" as a domain that encodes at least one soluble fragment. A fragment that is associated with a "soluble domain" is an element of the set D (Figure 1b). According to this definition, some fragments encoding a soluble domain may be insoluble. The fragments encoding non-soluble domains, i.e., predicted domains for which all fragments are insoluble, are elements of the set D^c (Figure 1b and

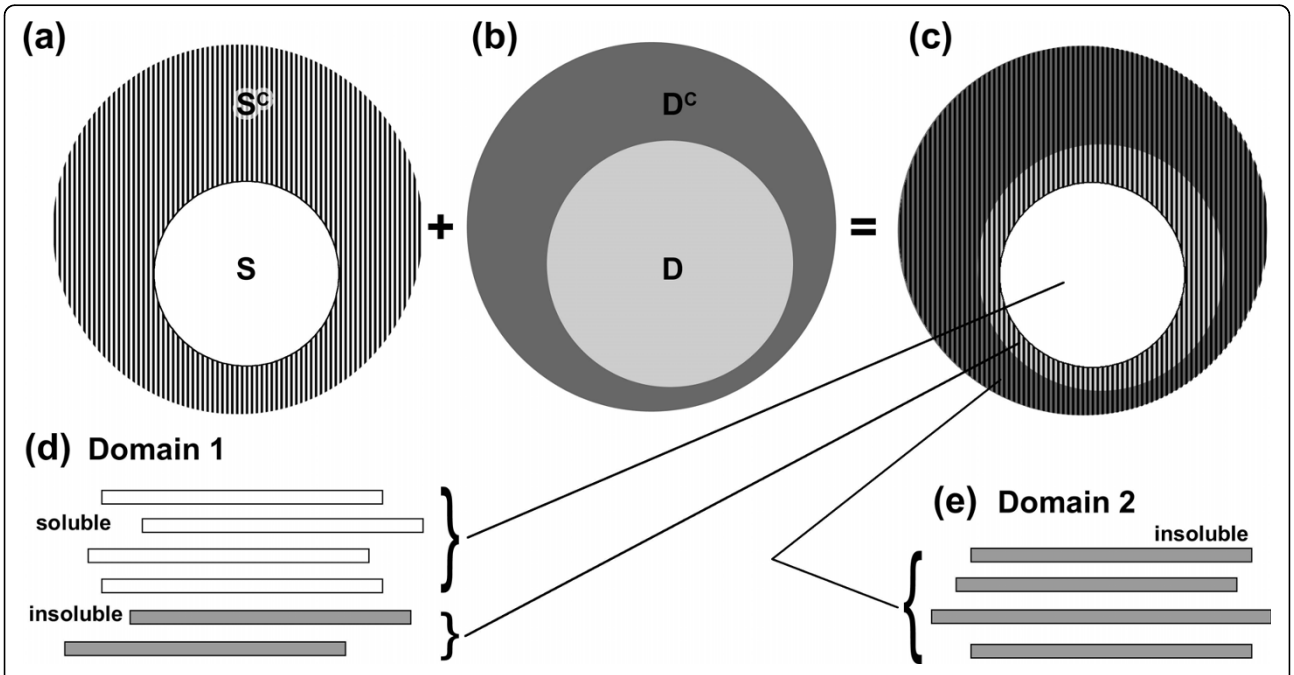

Figure 1 Sets for our model. An element of the sets in our model is an experimentally expressed protein domain fragment. (a) All elements are either soluble (S) or insoluble (S^C) fragments. (b) Alternatively, each fragment belongs to either a soluble domain (D) or a non soluble domain (D^C). (c) The whole set can be decomposed into four exclusive subsets: D ∩ S (white area); D ∩ S^C (lines on light grey area); D^C ∩ S^C (lines on dark grey area); and D^C ∩ S (not shown, since it is empty). (d) An example of the elements in the D subset. Fragments from a soluble domain (domain1) are shown. They can be classified into either the D ∩ S or D ∩ S^C subset (e) Fragments from a non soluble domain (domain 2) are shown. All elements belong to D^C ∩ S.

1e). Practically, a domain is defined as non-soluble if all of the tested fragments associated with a given domain (in our experiment, 9 per domain, on average) are insoluble. D and D^c are also mutually exclusive sets, and their elements are fragments (not domains). The above classification yields four fragment categories (Figure 1c): S ∩ D (Soluble domain fragments with correct N and C termini), S^c ∩ D (Insoluble fragments encoding a soluble domain, presumably because of incorrect termini), S ∩ D^c (Soluble fragment encoding a non-domain or a failed domain prediction; this set is obviously empty), and S^c ∩ D^c (Insoluble fragment encoding a non-domain; all elements of D^c obviously belong to S^c).

The probability of successfully predicting a soluble fragment, $p(S)$, is expressed as

$$p(S) = p\left(\{S \cap D\} \cup \left\{S \cap D^c\right\}\right) = p(S \cap D), \quad (1)$$

since S ∩ D^c is empty and all sets are exclusive. $p(D)$ is thus related to $p(S)$ as

$$p(S) = p(S \cap D) = p(D)\,p(S \mid D), \quad (2)$$

where $p(S|D)$ is the conditional probability of obtaining a soluble fragment of a soluble domain. When $p(S)$ and $p(S|D)$ are given, the probability of successfully predicting the existence of a soluble domain is calculated as

$$p(D) = \frac{p(S \cap D)}{p(S|D)} = \frac{p(S)}{p(S|D)}. \quad (3)$$

Note that Eqs. 1-3 are direct consequences from probability rules for independent sets, without any approximations or assumptions.

In a large-scale experiment aimed at obtaining as many soluble domains as possible, cost-optimization is achieved by maximizing the number of soluble domains for a fixed number of tests. As an approximation, our model computes the expected number of producing soluble domains, E_{domain}, by assuming average probability values over all of the protein domains examined in the experiment. Our model examines M protein domains and generates N fragments per domain. According to this model, the expected number of soluble domains, E_{domain}, is given by:

$$E_{\mathrm{domain}} = Mp(D)P_N = \frac{constant}{N}\,p(D)P_N(N, p(S \mid D)), \quad (4)$$

where M and N are, respectively, the number of domains and the number of fragments per domain that are assessed. When the total number of fragments (MN) is held *constant*, the probability of obtaining one or more soluble fragments, P_N, is a function of N and $p(S|D)$. The explicit form of P_N depends on the experimental setting, and is derived in the next section for two specific cases.

We can further modify Eq. 4 to add a set-up cost related to the analysis of a new domain. The set-up cost is taken into account by expressing the total cost as $MN + Mr$, where r is the ratio between the supplemental cost of analyzing a fragment from a new domain and that of analyzing a new fragment from the current domain. Keeping the total cost constant [$M(N + r) = constant$] yields:

$$E_{\mathrm{domain}} = \frac{constant}{N+r}\,p(D)P_N(N, p(S \mid D)). \quad (5)$$

Derivation of P_N for two basic experimental settings

Let us derive P_N for two basic types of experimental settings. In the first one, the generation of N fragments occurs by independent events (multiple copy case). This situation occurs in genetic screening experiments, where N fragments per domain are randomly selected and tested, allowing multiple copies of the same fragment to be tested. In this case, the mathematical expression for P_N is simply calculated as:

$$P_N = 1 - p\left(S^c|D\right)^N = 1 - \left\{1 - p(S|D)\right\}^N = 1 - \left(1 - \frac{f}{F}\right)^N \quad (6)$$

where P_N is the probability of obtaining one or more soluble fragments of a soluble domain when N fragments are simultaneously tested. F is the (average) number of all of the testable (assessable) fragments, and f is the (average) number of soluble fragments associated to a domain.

The second situation occurs when N fragments are generated, but each fragment is selected only once (single copy case). This situation occurs when the fragments are identifiable, such as in our pilot experiment with the Kazusa sequences. P_N is derived by using a hypergeometric distribution ($P_N{}^C$), which describes the probability of obtaining no soluble elements when N elements are drawn without replacement from a finite population of ${}_FC_N$ elements (${}_mC_n$ indicates the binomial coefficient for choosing n elements from m elements). P_N is given by

$$P_N = 1 - P_N^C = 1 - \frac{{}_{F-f}C_N}{{}_FC_N} = 1 - \frac{(F-f)!(F-N)!}{(F-f-N)!F!}, \quad (7)$$

where $P_N{}^C$ is the probability of obtaining no soluble fragments when N fragments are tested (see Additional file 1 for detail.) A JavaScript program, implementing Eqs. 5-7, is available in Additional file 2.

Parameter estimation from a pilot experiment with Kazusa HUGE domains

The application range of our mathematical model is not restricted to describe or analyze a specific domain prediction method (such as Armadillo[19], or PRODOM [20]), or experimental procedure (E-coli strains, or cell free systems). The specific settings/protocols are taken into account by adjusting or optimizing the values of the model's parameters. Here, we will estimate typical values for the parameters using the solubility of protein domains predicted in the Kazusa HUGE protein sequences [21].

We first identified 36 putative domains using ProteoMix [22] (Figure 2a), and for each domain, we expressed several fragments with different N and/or C-terminal residues distributed over the predicted termini window (Figure 2b). This yielded a total of 436 fragments; namely, 12 fragments per domain were generated, on average, to probe the domain termini. Among the 436 fragments, 215 (corresponding to 24 domains) were expressed correctly and were eventually assessed (Additional file 3); 111 fragments (encoding 18 domains) were soluble, and 104 fragments (encoding 6 domains) were insoluble (Figure 3).

For the purpose of discussion, let us estimate the parameters using the 215 fragments corresponding to the 24 domains that expressed correctly. Among these

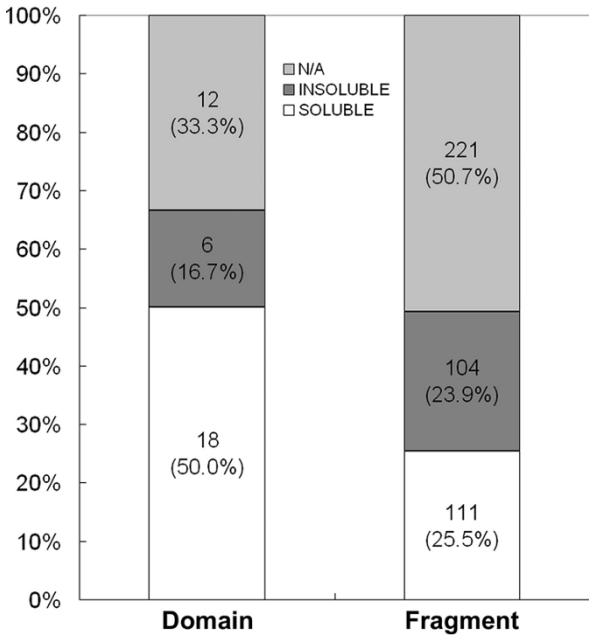

Figure 3 Summary of the solubility experiment. Soluble (SOLUBLE), insoluble (INSOLUBLE), and not applicable (N/A) domains (left bar) and the corresponding fragments (right bar) are shown. N/A includes all kinds of experimental incompleteness, such as insufficient PCR, wrong cDNA sequences, no expression or ambiguous SDS-PAGE results.

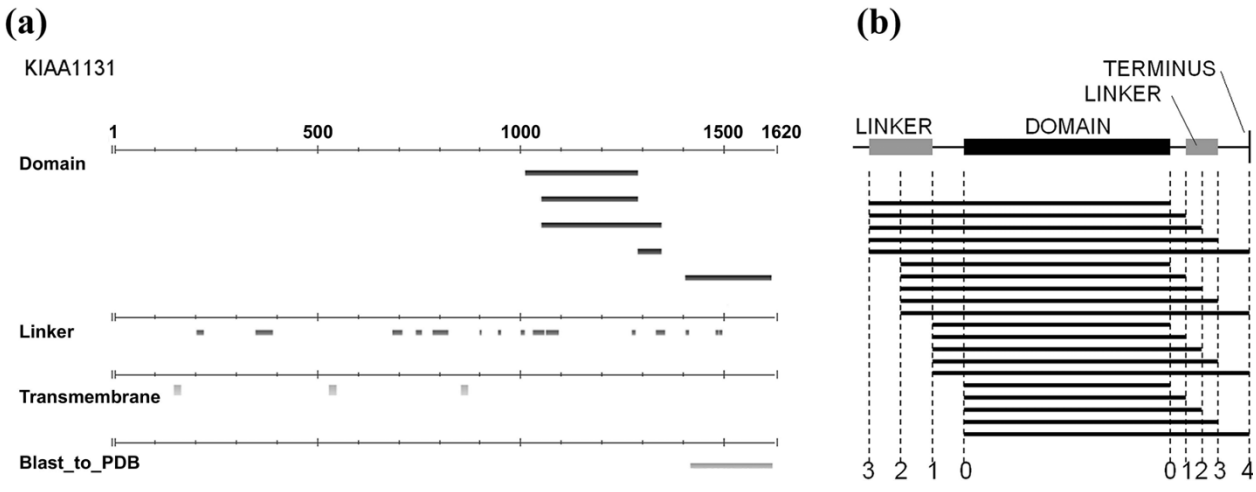

Figure 2 Design of domain fragments. (a) Snap shot of the ProteoMix for the Kazusa HUGE KIAA1131 sequence. The results of PASS (domain predictions, Domain), DLiP and DLP (domain linker predictions, Linker), HMMTOP and MEMSAT (transmembrane predictions, Transmembrane), and BLAST to PDB sequences (homology to structurally known sequences, Blast_to_PDB) are shown. (b) Domain fragment design scheme. For each dissected domain, the corresponding linker or the N- or C-terminus of the full-length sequence that is in the N- or C-terminal end of the domain is used to determine the termini of the fragments. Three cleavage sites are defined along the linker region. The first one is at the end of the predicted linker on the PASS-predicted domain side (cleavage site 1), the second one is located in the middle of the linker region (cleavage site 2), and the third one is at the end of the linker region, opposite from the domain side (cleavage site 3). In addition, we designed a cleavage site at the terminal ends of the PASS-predicted domain (cleavage site 0) and at either the N or C terminus of the protein (cleavage site 4, in case it was close to the C- or N- terminus of the domain). Some cleavage sites were not defined: for example, when the length of a linker was too short to determine its center, the central cleavage site was omitted. The number of generated fragments was regarded as sufficient for assessing the presence of a cleavable site and for estimating the model's parameters.

Table 1 Estimated model parameters

	D	D^C	ALL
S	0.52 ± 0.03	0.00 ± 0.00	0.52 ± 0.03[a]
S^C	0.29 ± 0.03	0.19 ± 0.03	0.48 ± 0.03[b]
ALL	0.81 ± 0.03[c]	0.19 ± 0.03[d]	1.00

a: p(S); b: $p(S^C)$; c: p(D); d: $p(D^C)$; S, S^C, D, D^C: see the main text; ALL: All the considered set; p(S ∩ D) (upper left cell): The probability of predicting a soluble fragment of a soluble domain, similarly for the S^C ∩ D (middle left), S ∩ D^C (upper center), and S^C ∩ D^C (middle center); Standard deviations were computed by bootstrapping, in which 1,000 sets of 215 randomly selected fragments were sampled.

domains, 75% (18/24), and 52% (111/215; which corresponds to p(S) and $p(S^C)$ = 0.48; Table 1) of the fragments were soluble when 9 fragments were tested on average (instead of 12 when all 436 of the fragments are considered). p(S|D), which is the conditional probability of correctly predicting a soluble fragment of a soluble domain, is 0.63 (111 fragments/175 fragments). The calculation of p(D) from Eq. 3, using the obtained values of p(S) and p(S|D), generates 0.81. Substituting 0.81 and 9 for p(D) and N, respectively, in Eq. 4, yields E_{domain}/M as 0.81, which is equal to p(D), within rounding error, since P_N is almost 1 when N = 9. The discrepancy from the above experimental E_{domain}/M value of 0.75 (Figure 3), is due to the use of average values. Indeed, for 6 domains, all of the fragments were soluble, and 12 domains yielded a mixture of soluble and insoluble fragments (Figure 4a). However, the corrections appear to be modest (less than 10%) and will not significantly affect the following discussion (the calculation with non-averaged f and F values is discussed in Additional file 1).

Let us perform a simple mathematical simulation to assess the discrepancy between the above experimental and calculated values of 0.75 and 0.81. The discrepancy can be resolved by setting the number of tested fragments per domain to, for example, 10, and by uniformly using for all domains the above experimentally determined average values of 0.75 and 0.52 for p(D) and p(S), respectively. As a result, 240 fragments are produced, corresponding to 24 domains, of which 6 do not yield any soluble fragment, since p(D) = 0.75. Since p(S) = 0.52, 124.8 fragments are soluble. These soluble fragments are uniformly distributed among the 18 domains, and thus 6.93 (7) fragments per domain out of 10 fragments are soluble, which yields 0.693 for p(S|D). Using Eq. 3, p(D) = 0.52/0.693 = 0.75, and thus p(D) P_N = 0.75 * {1 - (1 - 0.693)^{10} = 0.75, which is equal, within rounding error, to the experimentally determined value of 0.75 (18 domains out of 24).

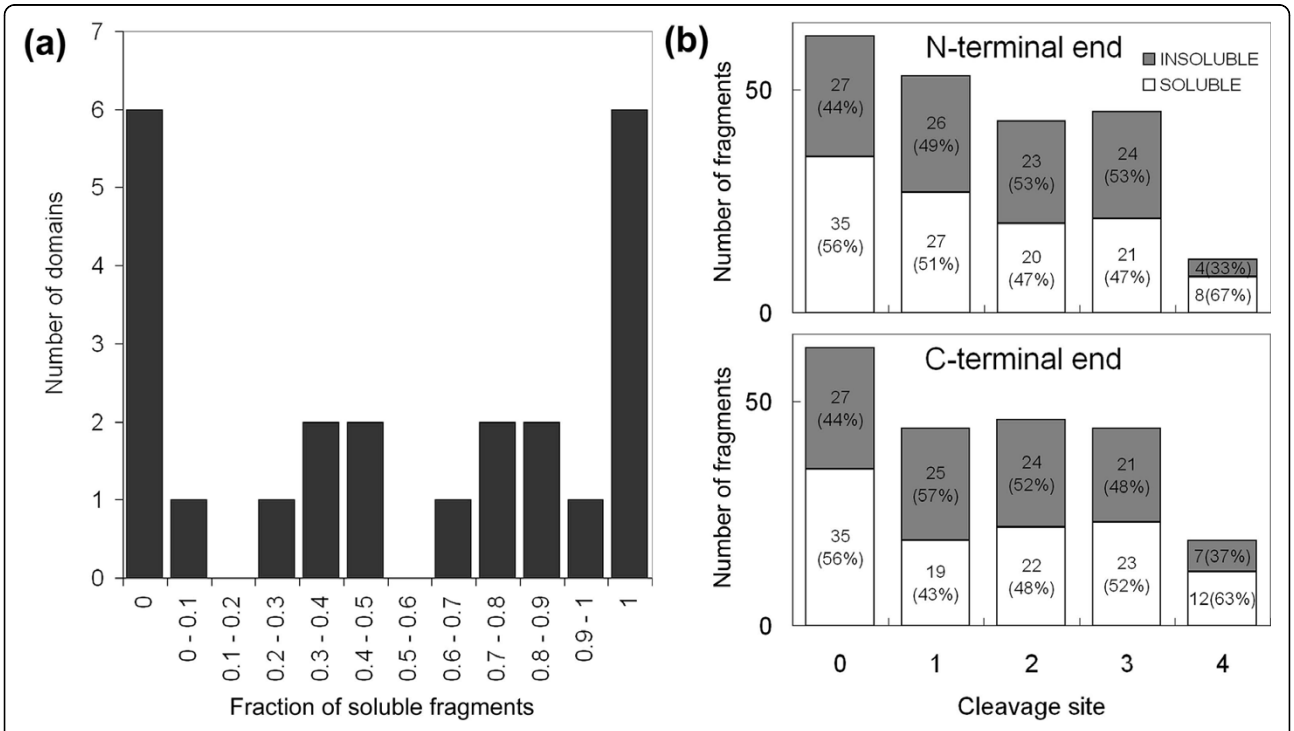

Figure 4 Soluble fragments for individual domains and cleavage sites. (a) Fractions of soluble fragments for individual domains. The horizontal axis represents the ratio of soluble fragments to assessed fragments for each of the 24 analyzable domains. (b) Distribution of the cleavage sites yielding soluble/insoluble fragments. The numbers of soluble and insoluble fragments are shown by white and dark gray bars, respectively. The N- (upper panel) and C-terminal (bottom panel) cleavage sites of the fragments are classified into sites 0-4, as defined in Figure 2b.

In the present experiment, 12 domains (221 fragments) could not be analyzed because of failure to express a fragment with a correct molecular weight, or to assess the solubility, which obviously indicates that the efficiency of our automated protein expression system has room for improvement. These 12 domains can be included in the analysis and will give lower (or upper) limits by considering them as insoluble (or soluble) domains. The inclusion yields 50% (18/36) and 25% (111/436) for, respectively, the lower probability limit of predicting the domain existence and that of predicting a soluble fragment for 12 tested fragments, on average. Similarly, some or all of the 6 domains that were considered insoluble in the above discussion might turn out to be soluble if more fragments were tested.

Effect of the cleavage site on the solubility of the dissected domain

The effect of the cleavage site on the solubility of the dissected domain was also examined (Additional file 3). All of the fragments corresponding to 6 domains (KIAA0190.21-227, KIAA0309.483-717, KIAA1142.35-97, KIAA1256.13-117, KIAA1416.55-202, and KIAA1459.745-848) were soluble (all-soluble class); while KIAA0067.1143-1295, KIAA0175.480-651, KIAA0190.436-653, KIAA0277.120-349, KIAA0641.76-352, and KIAA1338.180-406 were insoluble (all-insoluble class), indicating a failed domain prediction. On the other hand, the solubility of the remaining 12 domains was dependent on the position of the cleavage site, and both soluble and insoluble fragments were produced, depending on slight shifts of the cleavage residue at each of the N- and C- domain terminal ends (soluble +insoluble class). The effects of the cleavage site within the window were also examined. Except for site 4, which corresponded to either the N or C protein terminus, no differences in the yields of soluble fragments were observed (Figure 4b). Note that the small number of cleavage at site 4 simply results from the fact that not all domains were located at the N or C protein terminus.

Optimum fragment number per domain

Let us use our model to analyze protein dissection experiments with different settings. Since $p(\mathrm{D})$ is independent of N, we can simplify Eq. 5, and examine the normalized expected number of soluble domains, e_{domain}, instead of E_{domain}:

$$e_{\mathrm{domain}} = \frac{E_{\mathrm{domain}}}{constant \cdot p(\mathrm{D})} = \frac{P_N}{N+r}, \quad (8)$$

where P_N can be computed with either Eq. 6 or 7, since no noticeable difference will occur for practical purposes (for most cases, within a few percent error; see Figure 5). When no set-up cost is present ($r = 0$), e_{domain} is a monotone decreasing function of N (Figure 5a). Thus, one fragment per predicted domain will optimize the number of soluble domains ($N_{\mathrm{optimum}} = 1$). This is analytically demonstrated by showing that N_{optimum} is the solution of a transcendental equation with a unique solution (A mere exception occurs for $f = 1$ when the single copy per fragment model is used; Figure 5a and Additional file 1).

Set-up costs are typically generated by the purchase of new chemicals associated with the examination of a new domain, and may, for example include the clone's cost, from which the domain fragments are prepared by PCR. Large r values may occur in a genetic screening-type experiment, where the cost of assessing a new fragment is small as compared to that of starting a new experiment with a new domain. For $r > 0$, e_{domain} is not a monotone decreasing function of N, but reaches a maximum at $N = N_{\mathrm{optimum}}$ (Figure 5b). The value of N_{optimum} increases for increasing values of r (Figure 5b).

As an example, using the values derived from our pilot experiment [$p(\mathrm{S}|\mathrm{D}) = 0.63$ ($= f/F$)] and assuming that $r = 10$, we find that testing three fragments per predicted domain would yield more soluble domains than just one, for the same total cost (Figure 5b). Note that, as r increases, the peak broadens and the maximum e_{domain} value, e_{domain} (N_{optimum}), becomes smaller (Figure 5b). Thus, in practical terms, the e_{domain} dependence on N is small for large r and it becomes less important to accurately determine N_{optimum}. Finally, we note that in Eq. 6 (multiple copy model), e_{domain} depends on the ratio of f and F, which is the probability of finding the correct cleavage sites, but it does not depend directly on F alone, which is the total number of possible cleavage sites. The direct dependency on F alone is also minimal when the single copy model (Eq. 7) is used (Figure 5b).

Insight into a wide range of experimental settings can be obtained by analyzing the behavior of N_{optimum} for several values of $p(\mathrm{S}|\mathrm{D})$ and r (Figure 5c). For example, we find that N_{optimum} increases with decreasing values of $p(\mathrm{S}|\mathrm{D})$, which is intuitively sensible, since a smaller $p(\mathrm{S}|\mathrm{D})$ requires a larger number of trials for finding the termini residues that yield a soluble domain fragment. N_{optimum} is between 1 and 4 for $r = 1$ and a broad range of $p(\mathrm{S}|\mathrm{D}) > 0.1$, which covers most experimental settings including our pilot experiment using Kazusa sequences. A $p(\mathrm{S}|\mathrm{D}) > 0.1$ would also cover typical domain prediction tools such as Armadillo prediction, which has a $p(\mathrm{S}|\mathrm{D})$ value estimated by cross validation method between 0.3 and 0.5 [19]. Finally, for a typical value of $r = 10$ and $p(\mathrm{S}|\mathrm{D}) > 0.5$, N_{optimum} is between 1 and 3; and for $p(\mathrm{S}|\mathrm{D}) = 0.63$, N_{optimum} is 7 fragments, even for $r = 1{,}000$ (i.e., a very large set-up cost).

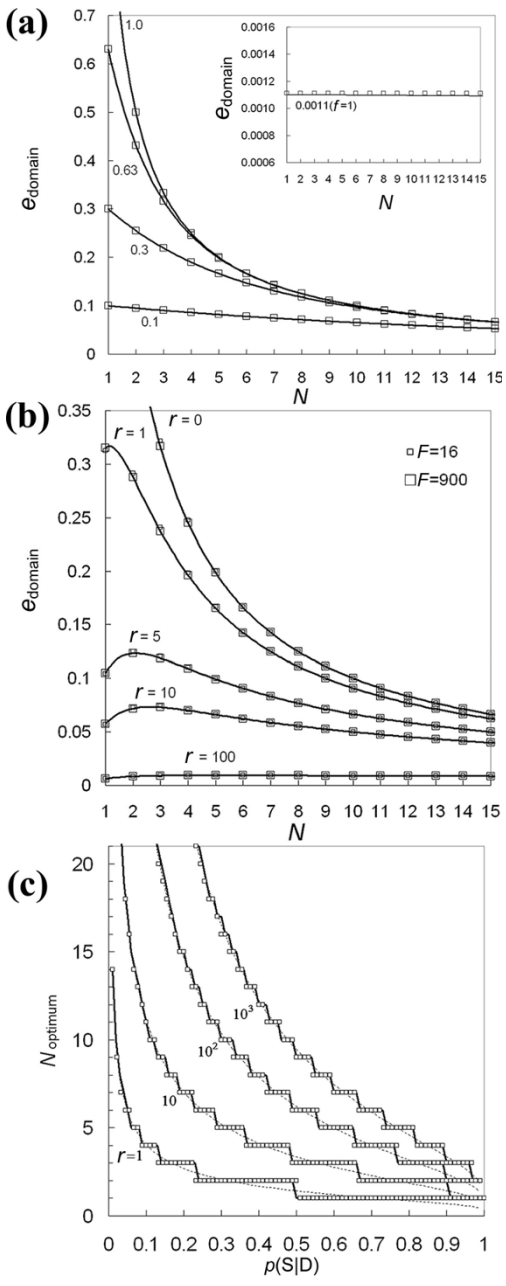

Figure 5 Dependence of e_{domain} on N. e_{domain} and $N_{optimum}$ were calculated for the multiple and single copy cases with, respectively, Eq. 6 (solid lines) and Eq. 7 (symbols). (a) e_{domain} is represented as a function of N for $p(S|D) = 1$, 0.63, 0.3, and 0.1. The curve for the special case of $f = 1$ [$p(S|D) = 0.0011$] is shown in the inserted panel, with vertical scale magnified 300 times. (b) Curves for $r > 0$ and $F = 16$ and 900. e_{domain} are shown for $r = 0$, 1, 5, 10, and 100 with $p(S|D) = 0.63$. $N_{optimum} > 1$ is clearly apparent. (c) $N_{optimum}$ is shown as a function of $p(S|D)$ for $r = 1$, 10, 100, and 1,000. $N_{optimum}$ can be any positive number. $N_{optimum}$ for the single copy case are discrete in N (symbols), while those for the multiple copy case are continuous (dotted lines). Discretization of the single copy model (solid lines) caused a small discrepancy at the very edge of switching a step of solutions.

Conclusions

We have presented a novel mathematical formulation for optimizing large-scale protein domain production experiments. Our model demonstrated that the testing of one to seven fragments per domain will fit most high throughput protein domain production experiments. The probabilistic approach presented here is not limited to protein domain production, and it can be readily modified and applied for designing various types of large-scale mutational analyses or screening libraries.

Methods

Computational protein dissection

Domains were computationally predicted from 2,032 Kazusa HUGE [21] protein sequences (KIAA0001-KIAA2033), using ProteoMix [22], an integrated protein sequence analysis system. The ProteoMix analysis included PASS [23] for domain predictions, and DLiP [24] and DLP [13,25] for domain linker predictions. PASS analyses were performed with default parameters (E-value = 1e-7, Cut-off homologues = 10); DLP with default (Threshold = 0.5, Cut-off = 0.5, Ignored terminal length = 0, Window = 19, Minimum difference = 0.05); and DLiP with no BLAST option. The other tools were HMMTOP [26] and MEMSAT [27] for trans-membrane region prediction, and BLAST [28] for removing domains homologous to structurally known protein sequences derived from the Protein Data Bank [29]. We obtained 269 putative domains according to the following rules: A putative domain was identified when it was a PASS predicted domain region whose boundary overlapped either with one of the full-length protein termini, or with a domain boundary predicted with either DLiP and DLP results (predicted domain linker regions) within ± 25 residues. We also required that the putative domain did not include transmembrane regions, as predicted by HMMTOP or MEMSAT with default parameters, to remove the inherently insoluble domains; and putative domains with sequence identities higher than 30% to PDB protein sequences were also removed. Finally, we choose 27 domains by visually inspecting the consistency among the prediction's tools, domain's, and domain linker's sizes. In addition, we completed this set with 9 domains predicted in one of the putative multi-domain protein sequences that contained one of the above 27 domains. This yielded 36 domains that were assessed experimentally. For each of the 36 dissected domains, a maximum of 20 fragments per domain, resulting in a total of 436 fragments, were designed by combining the results of PASS, DLiP, and DLP, and our termini selection rule (see Figure 2b). This resulted in a 30 residue N- or C- terminal end window (corresponding to 900 fragments per domain) on average, from which the termini residues were selected.

Experimental assessment

The cDNA clones for the selected 36 protein domains (436 fragments) were kindly provided by Kazusa DNA Research Institute (Kisaradsu, Japan). The corresponding protein fragments were expressed using an E-coli based cell-free system and were purified as described [30]. The fragments were classified as soluble, insoluble, and not applicable, according to an SDS-PAGE analysis of the supernatant and precipitate fractions. The soluble fragments were defined as the fragments that remained in the supernatant after centrifugation. The fragments that were present in the SDS-PAGE of the precipitate, but absent from the supernatant after centrifugation, were defined as insoluble fragments. All other fragments were classified as not applicable, which included all kinds of experimental obstacles, such as unsuccessful PCR, wrong cDNA sequences, no expression or ambiguous SDS-PAGE results.

Additional file 1: Additional methods. An extension for handling the distribution of $p(S|D)$ with the multiple copy model; Basic properties of P_N; Intuitive Derivation of P_N for the single copy per fragment case; Direct Derivation of P_N for the single copy per fragment case using the hypergeometric distribution.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-113-S1.PDF]

Additional file 2: Calculator of the expected number of soluble domains. A JavaScript program, implementing Eqs. 5-7.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-113-S2.HTML]

Additional file 3: Figure S1: List of expressed domains. List of expressed domains in our pilot example. Computationally dissected protein domains and their experimentally assessed solubilities.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-113-S3.PDF]

Acknowledgements
We thank Dr. Takanori Kigawa and the members of the GSC Protein Research Group for the cell-free production and screening system, and Kazusa Research Institute for the HUGE cDNA resources. This work was supported by the RIKEN Structural Genomics/Proteomics Initiative (RSGI), the National Project on Protein Structural and Functional Analyses, from MEXT.

Author details
[1]Genomic Sciences Center RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan. [2]Department of Biophysics and Biochemistry, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. [3]Department of Biotechnology and Life Science, Faculty of Technology, Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei, Tokyo 184-0012, Japan.

Authors' contributions
EC, AK, TY, SY, and YK designed the research. YK and EC derived the mathematical forms. TT, SM, AK, and EC developed the analysis programs. TY performed the experiments. EC and AK analyzed the data. EC and YK wrote the paper.

Received: 25 August 2009 Accepted: 1 March 2010
Published: 1 March 2010

References

1. Chandonia JM, Brenner SE: **The impact of structural genomics: Expectations and outcomes.** *Science* 2006, **311(5759)**:347-351.
2. Kouranov A, Xie L, de la Cruz J, Chen L, Westbrook J, Bourne PE, Berman HM: **The RCSB PDB information portal for structural genomics.** *Nucleic Acids Res* 2006, , **34 Database:** D302-305.
3. Terwilliger TC, Stuart D, Yokoyama S: **Lessons from structural genomics.** *Annu Rev Biophys* 2009, **38**:371-383.
4. Zhang C, Kim SH: **Overview of structural genomics: from structure to function.** *Current Opinion in Chemical Biology* 2003, **7(1)**:28-32.
5. Dessailly BH, Nair R, Jaroszewski L, Fajardo JE, Kouranov A, Lee D, Fiser A, Godzik A, Rost B, Orengo C: **PSI-2: structural genomics to cover protein domain family space.** *Structure* 2009, **17(6)**:869-881.
6. Card PB, Gardner KH: **Identification and optimization of protein domains for NMR studies.** *Methods Enzymol* 2005, **394**:3-16.
7. Fontana A, de Laureto PP, Spolaore B, Frare E, Picotti P, Zambonin M: **Probing protein structure by limited proteolysis.** *Acta Biochim Pol* 2004, **51(2)**:299-321.
8. Christ D, Winter G: **Identification of protein domains by shotgun proteolysis.** *J Mol Biol* 2006, **358(2)**:364-371.
9. Dokudovskaya S, Williams R, Devos D, Sali A, Chait BT, Rout MP: **Protease accessibility laddering: a proteomic tool for probing protein structure.** *Structure* 2006, **14(4)**:653-660.
10. Gao X, Bain K, Bonanno JB, Buchanan M, Henderson D, Lorimer D, Marsh C, Reynes JA, Sauder JM, Schwinn K, *et al*: **High-throughput limited proteolysis/mass spectrometry for protein domain elucidation.** *J Struct Funct Genomics* 2005, **6(2-3)**:129-134.
11. Marsden RL, McGuffin LJ, Jones DT: **Rapid protein domain assignment from amino acid sequence using predicted secondary structure.** *Protein Sci* 2002, **11(12)**:2814-2824.
12. Wheelan SJ, Marchler-Bauer A, Bryant SH: **Domain size distributions can predict domain boundaries.** *Bioinformatics* 2000, **16(7)**:613-618.
13. Miyazaki S, Kuroda Y, Yokoyama S: **Characterization and prediction of linker sequences of multi-domain proteins by a neural network.** *J Struct Funct Genomics* 2002, **2(1)**:37-51.
14. Suyama M, Ohara O: **DomCut: prediction of inter-domain linker regions in amino acid sequences.** *Bioinformatics* 2003, **19(5)**:673-674.
15. Niwa T, Ying BW, Saito K, Jin W, Takada S, Ueda T, Taguchi H: **Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins.** *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106(11)**:4201-4206.
16. Cabantous S, Waldo GS: **In vivo and in vitro protein solubility assays using split GFP.** *Nature Methods* 2006, **3(10)**:845-854.
17. Folkers GE, van Buuren BN, Kaptein R: **Expression screening, protein purification and NMR analysis of human protein domains for structural genomics.** *J Struct Funct Genomics* 2004, **5(1-2)**:119-131.
18. Hondoh T, Kato A, Yokoyama S, Kuroda Y: **Computer-aided NMR assay for detecting natively folded structural domains.** *Protein Science* 2006, **15(4)**:871-883.
19. Dumontier M, Yao R, Feldman HJ, Hogue CW: **Armadillo: domain boundary prediction by amino acid composition.** *J Mol Biol* 2005, **350(5)**:1061-1073.
20. Corpet F, Servant F, Gouzy J, Kahn D: **ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons.** *Nucleic Acids Res* 2000, **28(1)**:267-269.
21. Kikuno R, Nagase T, Nakayama M, Koga H, Okazaki N, Nakajima D, Ohara O: **HUGE: a database for human KIAA proteins, a 2004 update integrating HUGEppi and ROUGE.** *Nucleic Acids Res* 2004, , **32 Database:** D502-504.
22. Chikayama E, Kurotani A, Kuroda Y, Yokoyama S: **ProteoMix: an integrated and flexible system for interactively analyzing large numbers of protein sequences.** *Bioinformatics* 2004, **20(16)**:2836-2838.
23. Kuroda Y, Tani K, Matsuo Y, Yokoyama S: **Automated search of natively folded protein fragments for high-throughput structure determination in structural genomics.** *Protein Sci* 2000, **9(12)**:2313-2321.
24. Tanaka T, Yokoyama S, Kuroda Y: **Improvement of domain linker prediction by incorporating loop-length-dependent characteristics.** *Biopolymers* 2006, **84(2)**:161-168.
25. Miyazaki S, Kuroda Y, Yokoyama S: **Identification of putative domain linkers by a neural network - application to a large sequence database.** *BMC Bioinformatics* 2006, **7(1)**:323.

26. Tusnady GE, Simon I: **The HMMTOP transmembrane topology prediction server.** *Bioinformatics* 2001, **17(9)**:849-850.
27. Jones DT, Taylor WR, Thornton JM: **A model recognition approach to the prediction of all-helical membrane protein structure and topology.** *Biochemistry* 1994, **33(10)**:3038-3049.
28. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17)**:3389-3402.
29. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28(1)**:235-242.
30. Kigawa T, Yabuki T, Yoshida Y, Tsutsui M, Ito Y, Shibata T, Yokoyama S: **Cell-free production and stable-isotope labeling of milligram quantities of proteins.** *FEBS Lett* 1999, **442(1)**:15-19.

doi:10.1186/1471-2105-11-113
Cite this article as: Chikayama *et al.*: Mathematical model for empirically optimizing large scale production of soluble protein domains. *BMC Bioinformatics* 2010 **11**:113.