Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Improving the soluble expression of recombinant proteins by randomly shuffling 5′ and 3′ coding-sequence ends

Christophe Bignon, Changqing Li,‡ Julie Lichière,‡ Bruno Canard and Bruno Coutard*

Aix–Marseille Université, CNRS, AFMB UMR 7257, 13288 Marseille, France

‡ These authors contributed equally to this work.

Correspondence e-mail: bruno.coutard@afmb.univ-mrs.fr

Received 13 May 2013
Accepted 5 July 2013

Many structural genomics (SG) programmes rely on the design of soluble protein domains. The production and screening of large libraries to experimentally select these soluble protein-encoding constructs are limited by the technologies and efforts that can be devoted to a single target. Using basic technologies available in any laboratory, a method named 'boundary shuffling' was devised to generate orientated libraries for soluble domain selection without impeding the target flow.

Producing soluble proteins is a major bottleneck in functional and structural genomics. Protein solubility can be improved by expressing functionally or structurally autonomous domains rather than full-length proteins (Coutard & Canard, 2010). Two kinds of approach are used for designing these domains. The *in silico* approach relies on sequence analysis. However, to provide optimal results, completion of this approach often requires the experimental testing of different starts and stops of the *in silico* defined domain (Gräslund *et al.*, 2008). On the other hand, in the experimental approach libraries of randomly truncated DNA are produced without previous sequence analysis (Cornvik *et al.*, 2005; Pedelacq *et al.*, 2011). Unfortunately, up to 95% of the coding sequences generated by endonucleases or exonucleases may be out of frame, eventually resulting in high background libraries. As a consequence, the number of bacterial transformants to screen must be dramatically increased (up to 10^7 per library) and an additional screening step for selecting in-frame coding sequences must be performed before screening for solubility (Pedelacq *et al.*, 2011). The screening of these large libraries can then be conveniently performed by using a solubility reporter tag such as green fluorescent protein (GFP), but at the expense of an additional subcloning step to produce GFP-free protein for the downstream applications (Cabantous & Waldo, 2006). This subcloning step can be avoided by screening for soluble proteins by filtration of bacterial lysates. However, the screening throughput of this method is lower and also requires an additional step to quantitatively and qualitatively assess protein production (Cornvik *et al.*, 2005). Finally, fully automated protocols have been described to process large libraries (Yumerefendi *et al.*, 2010), but their cost restricts their use to few laboratories.

In order to simplify the truncation-library production and downstream screening steps, we have developed a strategy to drastically reduce the number of clones to test. The method, called 'boundary shuffling' (BS), relies on a PCR-based instead of a nuclease-based technology to generate the boundary diversity (Fig. 1). To make it possible, a predefined experimental space is selected by a computational approach on both sides of a core domain defined as the minimal sequence length necessary to code for a structurally or functionally standalone protein domain. This core can be designed on the basis of sequence or structure conservation, catalytic signature or hydrophobic plot and secondary-structure prediction if no other information is available (Gräslund *et al.*, 2008). Secondary-structure prediction is then used in combination with disorder prediction

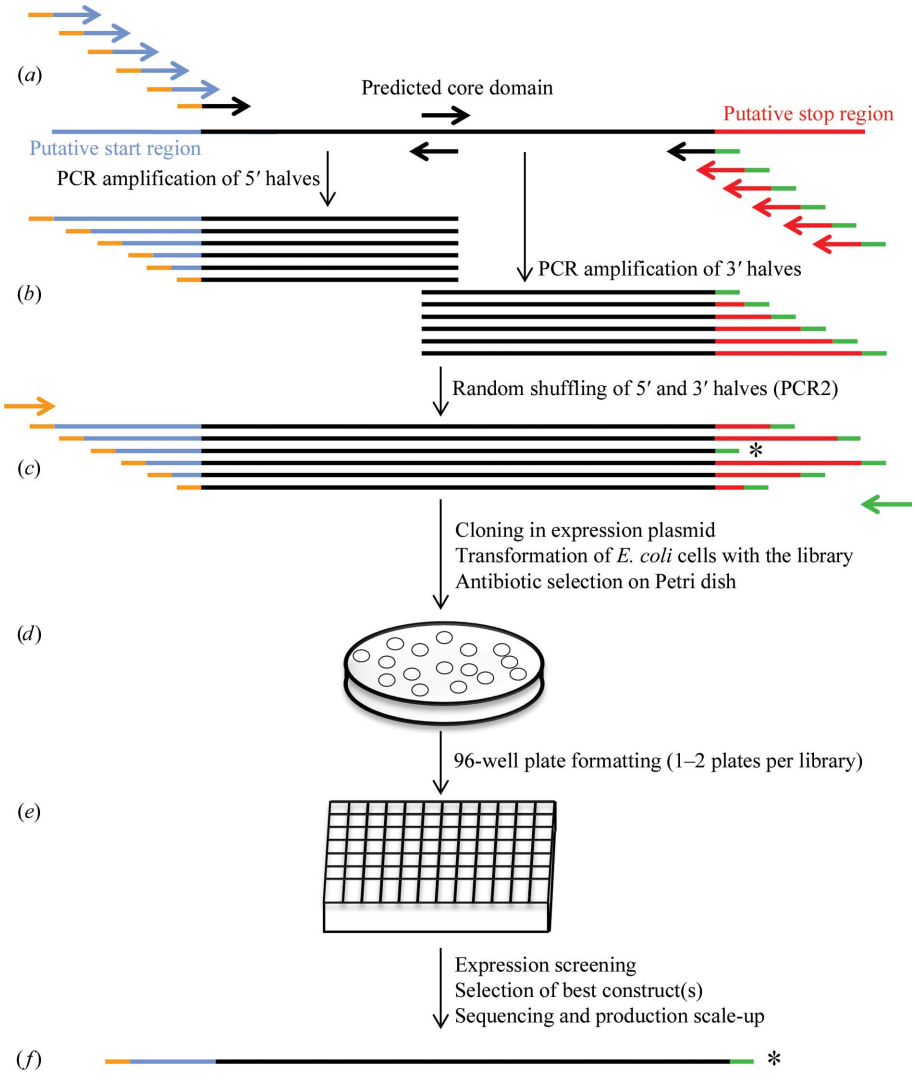

Figure 1
BS schematics from domain design to best construct selection. Putative start (light blue) and stop (red) regions, core domain (black) and corresponding primers are indicated. The sequences appended to the 5′ end of each forward and reverse primer (yellow and green, respectively) overlap the sequences of the generic primers used for the second PCR. The construct labelled with an asterisk corresponds to the selected protein domain.

(Lieutaud *et al.*, 2008; Mooij *et al.*, 2009) to define two regions of various lengths upstream and downstream of the core. In the upstream region, 5–12 putative starts can be selected. In the downstream region, 5–12 putative stops can similarly be predicted. Forward and reverse primers are designed accordingly to the starts and stops spanning the downstream and upstream regions, respectively. By doing so, all of the inserts are expected to be in frame with the plasmid sequence, which significantly reduces the number of clones to screen for soluble expression and avoids the pre-screening of in-frame sequences. Two different sequences are appended at the 5′ end of each forward and reverse primer, respectively. They are intended for hybridization with the two primers used for the second amplification, those that incorporate the vector-specific cloning sequences. Finally, a pair of complementary primers located in the middle of the core is also designed (Fig. 1*a*). In a first PCR, each forward primer is used in combination with the internal reverse primer to generate individual 5′ ORF halves. The same is performed with each reverse primer and the forward internal primer (Fig. 1*b*). The PCR efficiency is roughly assessed by individually running an aliquot of each 5′ and 3′ PCR product on an agarose gel. All 5′ and 3′ PCR products were pooled following an equimolar ratio. This pool was then used as the template for a second PCR performed in a single tube with the cloning sequence primers (Fig. 1*c*). During this second PCR, 5′ ORF halves randomly hybridize with 3′ ORF halves by means of the internal overlap corresponding to the two complementary internal primers, a method previously described for several applications (Horton *et al.*, 1989; Klock *et al.*, 2008; Figs. 1*a* and 1*b*). The self-elongation of these two PCR halves in combination with cloning-sequence primers recreates full-length cores with all possible combinations of starts and stops suitable for cloning. The resulting BS library is inserted into the favourite expression plasmid and expressed in *Escherichia coli*. To ensure that each possible start/stop combination is present at least once among the tested clones with >95% probability, a number of colonies corresponding to three times the number of combinations is tested for expression (Bosley & Ostermeier, 2005). For example, if eight starts are combined with eight stops (*i.e.* 64 combinations), 192 individual clones are analyzed for soluble expression. With such a reduced number of clones to test, classical strategies for expression screenings can easily be applied (Figs. 1*d*, 1*e* and 1*f*; Berrow *et al.*, 2006). Downstream biophysical or biochemical methods such as dynamic light scattering (DLS), thermal shift assay (TSA) or analytical size-exclusion chromatography (SEC) can be applied to the soluble proteins to select the best candidates for crystallization (Klock *et al.*, 2008; Geerlof *et al.*, 2006).

As a proof of concept, BS was tested on the macro domain of *Hepatitis E virus* (HepE MD). Based on existing structural homologues (Malet *et al.*, 2009), previous rational design of six constructs failed to produce a stable protein domain. Upstream and downstream of the core domain, two regions of about 120 base pairs each were defined on the border of the predicted disordered and structured elements that may or may not be part of the macro domain (Fig. 2*a*). Eight primers were designed in each of these two regions using the *ProteinCCD* server (Mooij *et al.*, 2009). The list of primers is reported in Supplementary Table S1[1]. The eight 5′ and eight 3′ ORFs halves were amplified individually (Fig. 2*b*). The first PCR products were pooled in an equimolar ratio. The pool was then purified on an agarose gel and used in a second PCR, as described above (Lantez *et al.*, 2011). The BS PCR library was cloned by recombination using the 'one-tube' protocol (Gateway, Life Technologies; Walhout *et al.*, 2000) into pETG20A, a vector expressing recombinant protein fused to an N-terminal thioredoxin–His_6 tag. To assess possible bias

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: YT5056). Services for accessing this material are described at the back of the journal.

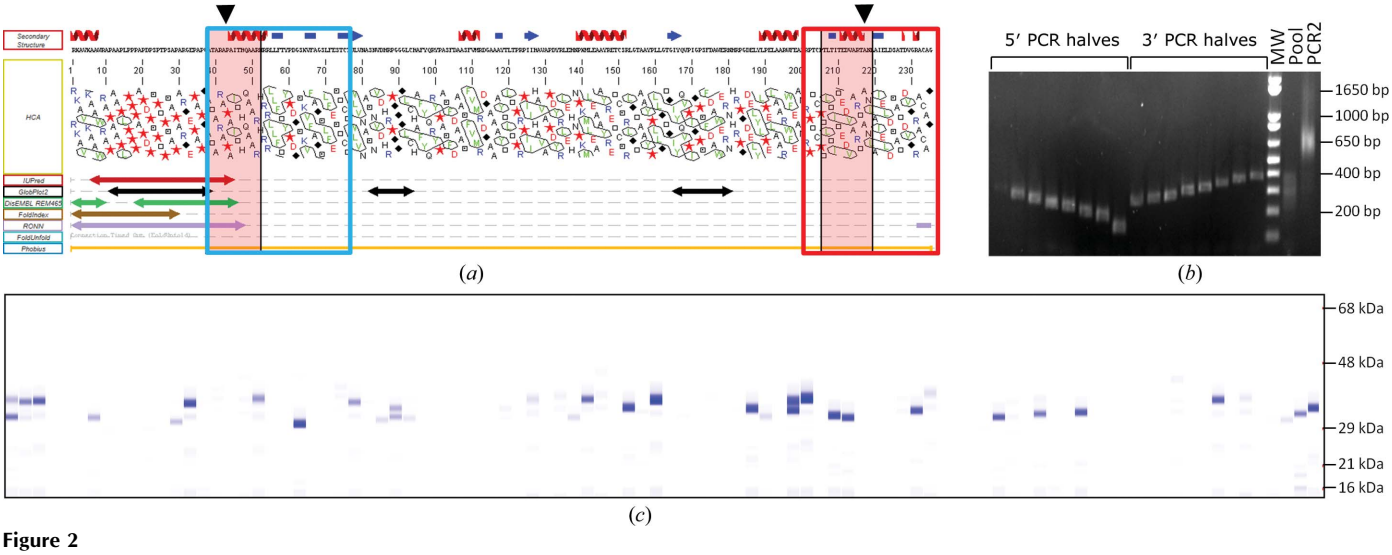

Figure 2
Results of the *Hepatitis E virus* macro domain project. (*a*) Sequence analysis of the macro domain with several predictors using the *MeDor* metaserver (secondary-structure prediction, HCA plot, disorder and hydropathy predictions; Lieutaud *et al.*, 2008). Light blue and red frames locate putative start and stop regions, respectively. The 5′ and 3′ boundaries borne by constructs encoding soluble proteins are highlighted in salmon pink. The two arrowheads indicate the limits of the construct encoding the more stable domain. (*b*) Agarose gel electrophoresis pattern of individual 5′ and 3′ PCR halves, the PCR halves pool before PCR2, molecular-weight markers and PCR2. (*c*) Coomassie Blue-like representation of capillary electrophoresis (LabChip GXII, PerkinElmer) pattern of IMAC micropurified expression results of 96 BS clones.

introduced during library production, 12 colonies were randomly picked among the ∼200 colonies obtained by heat-shock transformation. The inserts borne by pETG20A were then sequenced. All 12 constructs contained a HepE MD domain in the expected reading frame. Moreover, no start and stop combination was over-represented, suggesting that BS did not bias the library composition (data not shown). The expression of HepE MD was analyzed in the lysate of 192 clones by capillary electrophoresis following IMAC micro-purification in 96-well format (Fig. 2*c*). About one sixth of the clones expressed a soluble protein with no detectable proteolysis and with yields compatible with the demands of crystallization trials. The sequencing of the insert borne by pETG20A of 11 of these clones showed that their starts and stops are all located in short upstream and downstream regions, respectively (Fig. 2*a*). Six constructs led to stable recombinant proteins, all of which were homogenous in solution, as revealed by SEC. Four protein domains showed good denaturation profiles by TSA and were suitable for crystallization assays. That showing the highest melting temperature by TSA led to crystallization hits (Fig. 2*a*). Following this example, BS has been successfully applied with deletion regions as large as 600 nt for proteins with poor functional and structural information (data not shown).

Truncation experiments are a major strategy to improve the production of soluble and crystallizable proteins. Using boundary shuffling to perform 30–200-amino-acid truncations on both sides of a targeted domain, the process from library cloning to expression screening can be completed within a week and at a lower cost than processing ten individual constructs in parallel, mainly because BS uses a single cloning reaction and only selected clones are sequenced. Moreover, all of the clones expressing soluble proteins contain the domain of interest, preventing the unwanted core-domain truncations that can occur with nuclease-based approaches. Finally, boundary shuffling does not require the use of specific cloning methods, tags or expression plasmids and can therefore be adapted in any structural biology laboratory with little modification of production processes.

This work was supported by the EU through the European Virus Archive (EVA) project (European FP7 Capacities Project No. 228292; http://www.european-virus-archive.com/) and EUVIRNA (Marie Curie Initial Training Network Project No. 264286; http://euvirna.phrmy.cf.ac.uk/).

References

Berrow, N. S. *et al.* (2006). *Acta Cryst.* D**62**, 1218–1226.
Bosley, A. D. & Ostermeier, M. (2005). *Biomol. Eng.* **22**, 57–61.
Cabantous, S. & Waldo, G. S. (2006). *Nature Methods*, **3**, 845–854.
Cornvik, T., Dahlroth, S. L., Magnusdottir, A., Herman, M. D., Knaust, R., Ekberg, M. & Nordlund, P. (2005). *Nature Methods*, **2**, 507–509.
Coutard, B. & Canard, B. (2010). *Antiviral Res.* **87**, 85–94.
Geerlof, A. *et al.* (2006). *Acta Cryst.* D**62**, 1125–1136.
Gräslund, S. *et al.* (2008). *Nature Methods*, **5**, 135–146.
Horton, R. M., Hunt, H. D., Ho, S. N., Pullen, J. K. & Pease, L. R. (1989). *Gene*, **77**, 61–68.
Klock, H. E., Koesema, E. J., Knuth, M. W. & Lesley, S. A. (2008). *Proteins*, **71**, 982–994.
Lantez, V., Dalle, K., Charrel, R., Baronti, C., Canard, B. & Coutard, B. (2011). *PLoS Negl. Trop. Dis.* **5**, e936.
Lieutaud, P., Canard, B. & Longhi, S. (2008). *BMC Genomics*, **9**, Suppl. 2, S25.
Malet, H., Coutard, B., Jamal, S., Dutartre, H., Papageorgiou, N., Neuvonen, M., Ahola, T., Forrester, N., Gould, E. A., Lafitte, D., Ferron, F., Lescar, J., Gorbalenya, A. E., de Lamballerie, X. & Canard, B. (2009). *J. Virol.* **83**, 6534–6545.
Mooij, W. T., Mitsiki, E. & Perrakis, A. (2009). *Nucleic Acids Res.* **37**, W402–W405.
Pedelacq, J. D., Nguyen, H. B., Cabantous, S., Mark, B. L., Listwan, P., Bell, C., Friedland, N., Lockard, M., Faille, A., Mourey, L., Terwilliger, T. C. & Waldo, G. S. (2011). *Nucleic Acids Res.* **39**, e125.
Walhout, A. J., Temple, G. F., Brasch, M. A., Hartley, J. L., Lorson, M. A., van den Heuvel, S. & Vidal, M. (2000). *Methods Enzymol.* **328**, 575–592.
Yumerefendi, H., Tarendeau, F., Mas, P. J. & Hart, D. J. (2010). *J. Struct. Biol.* **172**, 66–74.