Combinatorial Domain Hunting: solving problems in protein expression

Eddy Littler

Domainex, 324 Cambridge Science Park, Milton Road, Cambridge, CB4 0WG, UK

Modern drug discovery demands large amounts of high-quality protein and, therefore, begins by expressing target genes in heterologous systems such as bacteria or insect cells. However, some of the most attractive drug targets have proven challenging to clone and express. A technology called Combinatorial Domain Hunting has been developed to express regions or domains of proteins in a non-aggregated form. This technology has been applied to more than 50 targets and, in all cases, high-expressing protein domains have been produced. In many cases, the protein is also non-aggregated, making it ideal for developing assays for screening or determining its X-ray crystal structure.

In the past 20 years, there has been a major shift in the drug discovery paradigm. Previously, drug discovery relied heavily upon testing a small number of compounds in pharmacology models. Although this approach discovered many major drugs, it was a slow and ill-defined process and could not be scaled up to handle larger numbers of targets. The Human Genome Project has identified approximately 25 000 human genes. The subsequent study of the genetics or expression patterns of these genes has identified a large number of proteins that have a causative role in many diseases. Thus, there has been a large increase in the number of targets of interest for drug discovery [1–5]. However, despite the fact that the human genome sequence (HGS) provided the drug discovery field with many targets, each with a potentially major role in tackling disease, approaches to drug discovery in use at that time would not have enabled these targets to be exploited.

Fortunately, at the same time as the DNA sequence of the human genome was being determined, technologies such as structural base drug design (SBDD), high-throughput screening (HTS) and combinatorial chemistry were being developed. The HGS and the emergent technologies coalesced into an approach now used by all pharmaceutical and biotechnology companies [6,7]. This approach involves obtaining large amounts of purified protein, usually from recombinant systems, which is used for HTS and SBDD in parallel. These complementary approaches can handle many targets and quickly identify hit compounds with potential for further development.

Unfortunately, this new approach to drug discovery has several potential bottlenecks. One early, and major, bottleneck is the production of sufficient quantities of high-quality protein to be used to develop screening assays or to determine protein structures. The approach used to obtain sufficient protein is cloning the gene of interest into a heterologous expression system, such as a bacterium or insect cell [7]. Unfortunately, in many cases, it has proven difficult or impossible to clone and express target genes. When it is not possible to clone and express such genes, the modern paradigm for drug discovery breaks down. For example, although many projects can be successful in the absence of an X-ray crystal structure, others reach an impasse when the structure–activity relationship of the compounds is not understandable. This problem can often be resolved by determining how a compound binds into the active site of a target, which might reveal that the binding is not as predicted by the activity of the rest of the chemical series [6,7]. In the absence of a protein X-ray crystal structure, this discrepancy is difficult to resolve and has proved to be a major bottleneck in drug discovery. As a result, the progression of some highly attractive targets, such as PI3-kinase delta, has been severely restricted.

Traditional techniques for high-volume protein expression

Generating large amounts of high-quality proteins for some targets has proved difficult or impossible. There are several reasons

E-mail address: eddy.littler@domainex.co.uk.

1359-6446/06/$ - see front matter © 2010 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2010.03.010

why proteins might only express poorly in recombinant systems. In many cases, it is because the proteins need to be expressed in the presence of a complex, such as a membrane or one or more proteins, to fold. Hence, in the absence of these other components, individual proteins fail to express efficiently. In many cases, however, there is no obvious reason why a protein is expressing poorly [8].

There are several traditional approaches to tackling problems of protein expression. Examples include changing the amino acid sequence of the protein by mutagenesis and stabilizing proteins by engineering them so they express with terminal extensions. Often, expression of a gene in an alternative heterologous expression host can alone solve the problem or, if this does not work, the expression conditions (temperature and time), in-cell lysis protocols, protein solubilization and purification routes can all be varied [9–27]. These approaches are time-consuming, however, because they rely largely upon trial and error and, consequently, are inefficient and unpredictable. Various authors have discussed the possibility of developing alternative approaches that are more standardized and exhaustive than trial and error [27–34]; however, these have not been demonstrated experimentally until recently.

In an attempt to solve expression problems, commercial and academic research has focused on identifying regions of the protein that are easier to manipulate and express but that contain appropriate sequences that enable the protein domain to be folded correctly and in some cases to be biologically active. A common approach to identifying these useful regions of a protein or 'domains' is by using bioinformatic analyses of the gene and the protein it encodes. Unfortunately, to date, the predicted structure of a protein does not give general rules that can predict the exact positions of the start and finish of a suitable domain. One of the most advanced approaches to predicting the precise position of a domain has been described by the structural Genomics Consortium using the Domain Boundary Analyser, which enables any relevant bioinformatic information to be visualized simultaneously to make estimates as to the best position to make constructs [35]. In this case, predicted constructs are subsequently cloned and expressed with an overall success rate of 50% but with an increase in the usefulness of the constructs for X-ray crystallography. However, differences of just a few amino acids in defining the boundaries of a domain can change the stability and level of expression of the final protein product considerably. For example, studies on the p85a subunit of class 1A phosphoinositide 3-kinase show that the C-terminal SH2 domain, when cloned as a sub-fragment starting in the bioinformatically predicted SH2 domain, gives a protein that is very prone to precipitation. However, clones generated by Combinatorial Domain Hunting (CDH), which are transcribed several base pairs upstream, express protein that is ideal for further study [36].

An accepted alternative to using bioinformatic analysis is an approach called CDH. Here, a library is created consisting of DNA fragments of a parental gene. These libraries are filtered to select clones that code for stably folded, globular domains that can be expressed at high levels and taken forward to the next stage of drug discovery (Figure 1).

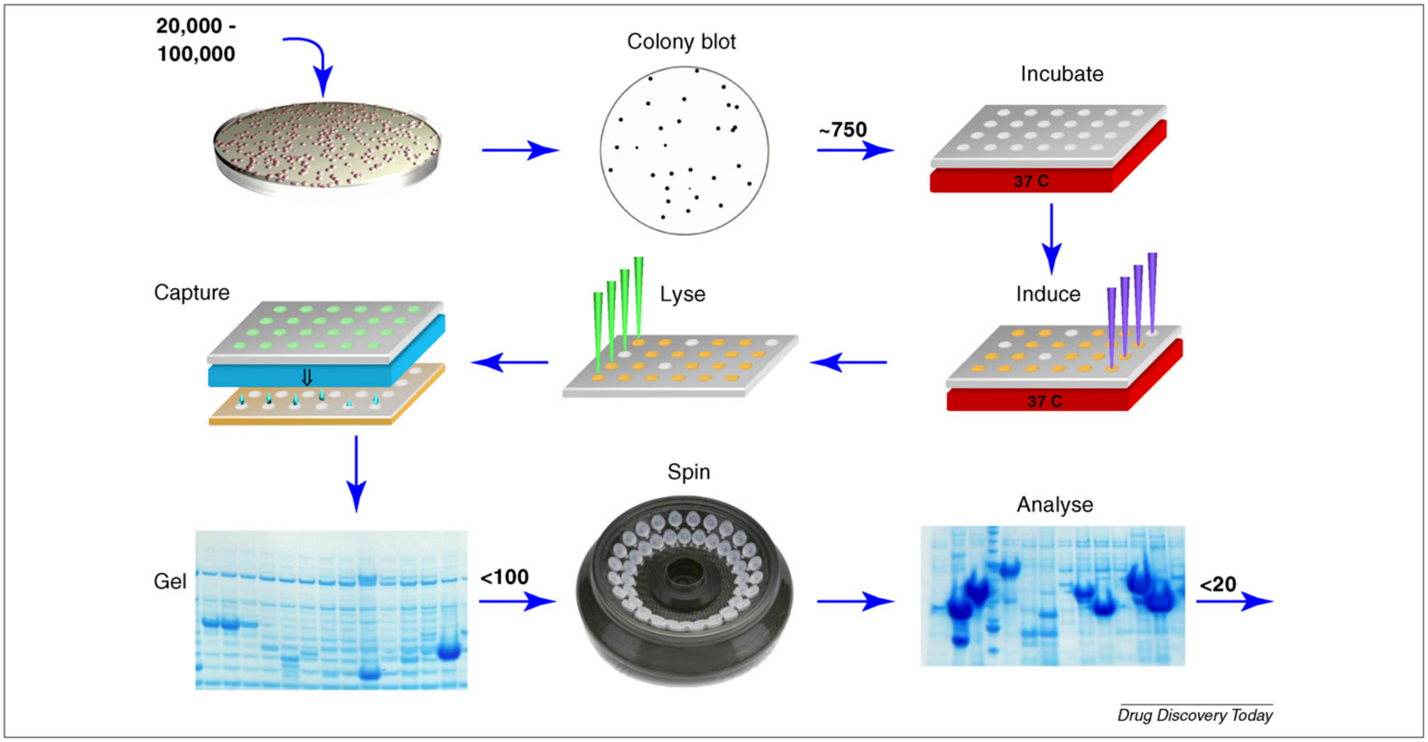

FIGURE 1

CDH in action. The figure shows how CDH starts with the generation of a library of approximately 20 000 independent clones, each containing a fragment coding for a domain or domains of a target. The libraries are subjected to colony hybridization using a monoclonal antibody that reacts with the polyHis tag on each clone. Approximately 500–700 clones are selected and grown in multiwall plates and tested for the amount and size of expressed protein. Proteins from clones that produce good levels of protein of appropriate size are then subjected to centrifugation and tested to see whether proteins are aggregated. Clones expressing non-aggregated protein domains are then sequenced and if they code for regions of interest are considered to be CDH 'hits'.

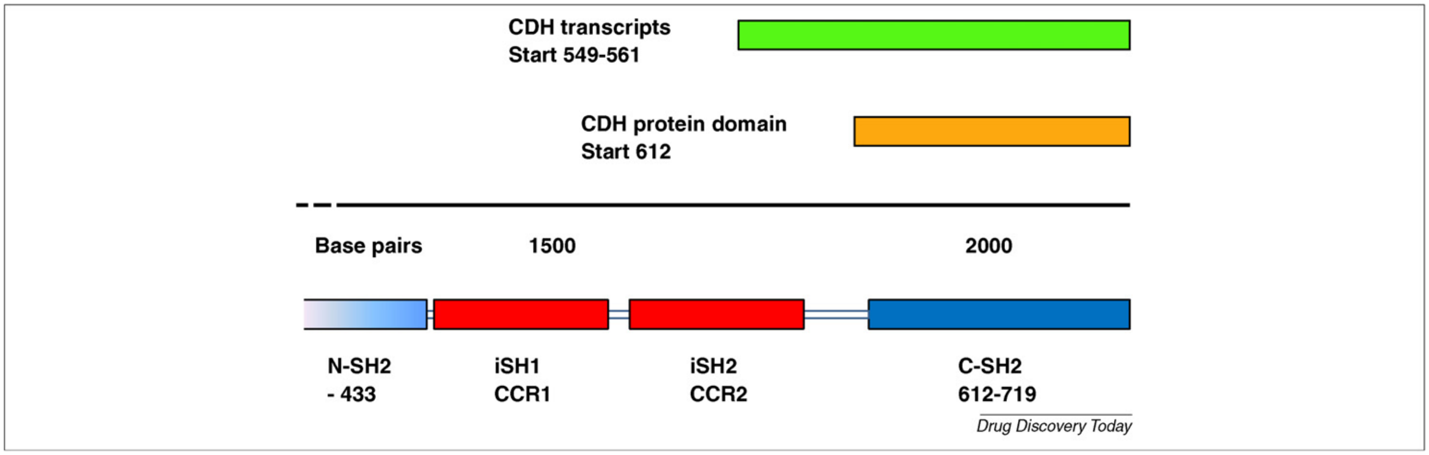

FIGURE 2

Partial results on CDH analysis of p85 alpha regulatory subunit. The diagram shows a partial genetic map of p85 starting at 1000 base pairs from the 5′ end of the gene. Regions identified by bioinformatics as having properties and sequences in common with regions such as SH2 domains are indicated, as are regions predicted to be coiled (cc). Three independent clones obtained containing the C-terminal SH2 domain were shown to contain DNA sequences starting from approximately 1600 bp from the 5′ end of the gene and extending to around 12 bp before the end of the coding region. These regions would express protein fragments starting from within a coiled-coil region of the p85 alpha containing an SH2 domain. In each case, however, the proteins produced are partially proteolyzed to proteins starting within the inter-domain region between the coiled-coil SH2 domain and the C-terminal SH2 domain. The resultant protein domains are highly stable and suitable for structural studies.

Techniques for DNA fragmentation in domain hunting

Several techniques are used to generate sub-fragments of DNA. Ordered DNA fragmentation creates uniformly sized fragments of DNA incorporating restriction sites designed for easy incorporation into the chosen expression vector. This technique uses PCR oligonucleotide primers running in parallel reactions to generate fragments with sequence-defined 5′ and 3′ ends. Although the process generates protein products with domain boundaries that can be specified to within a single amino acid residue, it is limited by the cost of oligonucleotide primers and the number of clones that have to be handled and analyzed at one time. Many organizations that have gone down this route have found serious bottlenecks in their ability to handle such large numbers of clones. The technique also presupposes that the boundaries of a domain can be predicted to some extent. Although in some cases approximate boundaries can be defined (or guessed), in many cases the location of the boundaries is not predictable. For example, when Domainex founder scientists applied CDH to the C-terminal SH2 domain of p85 alpha described above, the identified clones that expressed large amounts of good-quality protein contained DNA fragments initiated, and hence transcribed, from an internal SH2 domain located within a region of protein predicted to be coiled-coil [36]. Furthermore, the protein was subsequently found to be partially proteolytically degraded to a domain containing the C-terminal SH2 domain, whereas the proteins have their N-terminus located in the inter-domain region (Figure 2). It is simply not possible, at this time, to use bioinformatics to predict this type of construct.

Alternative approaches to generating DNA sub-fragments attempt to randomly fragment the DNA using physical, enzymatic or PCR-based methods. If these approaches are coupled with a method for screening large numbers of clones at an early stage, then libraries of fragments can be analyzed at lower cost and higher speed than with ordered fragmentation.

There are several enzymatic methods that, typically, either cut the double stranded DNA by selectively removing bases at the 3′ or 5′ ends of DNA strands or nick dsDNA strands so they can be broken by another reaction. The latter approach has an advantage over directly removing bases because both ends of the DNA molecule are varied at the same time. This gives rise to a much larger variation in the start and finish points among the gene domains produced.

A further technique generates DNA fragments from a DNA template using PCR amplification with tagged random sequence DNA primers. The primers are designed with two sections: a randomly encoded 3′ sequence and a 5′ sequence that is non-complementary to any sequence on the DNA template. This technique is called ‘tagged’ PCR, or T-PCR [37], and requires two or more PCR cycles. During the first PCR cycle, the random section of the tagged primer can attach to any complementary section on the template. In the second cycle, a second tagged primer can attach to the reaction products from the first PCR cycle. This produces DNA fragments terminated with two tag sequences. After the unreacted tag primers and primer dimers have been removed, the DNA fragments are amplified by secondary primers complementary to the non-random 5′ region on the tag primers.

The T-PCR technique requires only small amounts of DNA template. The distribution of DNA fragments is biased towards shorter fragments, however, because more than one PCR cycle is run. An additional bias occurs because GC-rich random primer sequences bond less efficiently to the template than AT-rich random primer sequences; therefore, some regions of the DNA will not be adequately represented in the final pool of clones.

A new, easy procedure that avoids the problems of physical, enzymatic and T-PCR techniques has now been demonstrated successfully on more than 50 occasions. It randomly fragments a DNA template without being biased by the sequence of bases on the template. Furthermore, it yields PCR products that can be directly incorporated into an expression vector ready for screening [31]. This technique is CDH.

CDH

CDH serves to generate a high degree of diversity in a library of protein products. The CDH process begins with a comprehensive recoding of the target gene. This enables a given gene sequence to be optimized for expression. For example, genes or portions of genes are optimized for expression in *Escherichia coli* first by changing the codons encoded by the DNA sequence to optimize them for expression. In addition, GC islands and regions of secondary structure are removed. Finally, to optimize the gene sequence for CDH, the distribution of AT base pairs in the DNA sequence is evenly distributed across the target gene.

The next stage in CDH is to run a PCR on the target gene with reaction conditions that are radically different to those in a normal PCR procedure. Specifically, while in a normal PCR, the reaction will contain the four deoxyribonucleotide triphosphates (abbreviated to dATP, dCTP, dGTP and TTP); in CDH, the TTP is replaced by a carefully controlled mixture of TTP and dUTP. The DNA polymerases used for the PCR reaction, such as Taq polymerase, are of low fidelity and will incorporate either TTP or dUTP without any preference. Therefore, the product of the CDH PCR will be a mixture of DNA fragments of the same size but containing dUTP incorporated randomly across the DNA sequence. The number of substitutions of dUTP for TTP can be controlled by carefully controlling the ratio of TTP and dUTP and the time of the PCR reaction.

The next step in CDH is to remove the randomly distributed uracil bases using uracil-DNA glycosylase, leaving abasic sites that can be cleaved by endonuclease IV to generate single-strand nicks in the DNA. The single-stranded nicks in the DNA are converted to double-stranded breaks by applying S1 nuclease. Finally, prolonged use of S1 nuclease will blunt end the fragments, which are then suitable to clone into appropriate vectors. The resultant DNA products are a collection of cloned DNA fragments of a desired size, which start and finish at a highly variable location both upstream and downstream of the DNA region of interest. The size of the domain can be controlled by three factors: the approximate choice of the domain that is recoded at the beginning of the CDH process, the PCR reaction described above and the size of the products selected with agarose gels [38,39].

CDH gains its immense power because these three factors vary simultaneously, enabling a huge number of clone permutations to be explored in parallel. For example, a domain of approximately 2000 base pairs flanked with 100 nucleotides upstream of the region of interest and a further 100 base pairs downstream of the region, when generated by CDH, would create approximately 10 000 clones that differ in their start or finish position. By contrast, an alternative technology that varied one and then another end would explore only a hundredth of the potential starts and finishes around a clone. It would, therefore, miss many (99%) potential permutations that could solve DNA expression problems.

The second phase of CDH (shown in Figure 1) is to screen the library of variant clones to identify those clones that express large amounts of well-folded proteins. Large libraries of between 20 000 and 30 000 independent clones are screened using a monoclonal antibody that reacts with a C-terminal polyHis tag. This polyHis tag is introduced when the DNA fragments generated by CDH are introduced into the plasmid vector. The incorporation of the polyHis tag must be in-frame with the expressed domain, thereby reducing the efficiency of the cloning; however, the large number of clones analyzed compensates for any such reduction. Clones are screened by colony hybridization using conditions that only partially lyze the bacterial colonies. Thus, to generate a strong signal, a clone will need to express large amounts of soluble proteins. In addition, the polyHis tag will need to be available for the monoclone, and this tag availability is a function of the folding of the recombinant protein [38,39]. Aggregated protein will be inefficiently transferred from the partially lyzed bacteria to the membrane, and any protein that is transferred will only have low reactivity with the polyHis antibody because the polyHis amino acid residues will be in a disordered state and not exposed to antibody.

Using the CDH process takes three months and can solve protein expression problems that have traditionally taken several years.

TABLE 1

Proteins considered to be difficult to express in suitable form or amounts for structural studies

Protein kinase C iota (PKC-iota)	**Respiratory syncytial virus RNA polymerase (RSV-L)**
Multiple mixed lineage leukemia (AF10-MLL)	Dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 3 (DYRK3)
Insulin like growth factor receptor (IGFR1)	Mammalian target of rapamycin (m-TOR)
Protein kinase D 1 (PKD1)	Amine oxidase flavin containing domain 2 (AOF2)
Ubiquitin ligase 3 (e3 ligase)	I kappa B kinase epsilon (IKKe)
NF-kappa-beta-inducing kinase (MAP3K14, NIK)	Protein kinase N1 (PKN1)
Ataxia telangiectasia mutated gene (ATM)	Enhancer of zeset homolog 2 (EZH2)
Mitogen-activated protein kinase activated protein kinase 2 (MAPKAPK2)	Phosphatidylinositol 3-kinase delta (PI3K delta)
Protein kinase N3 (PKN3)	Ataxia telangiectasia and Rad3 (ATR)
Fms-like tyrosine kinase 4 (Flt-4)	Hepatocyte growth factor receptor (c-MET)
Bone marrow tyrosine kinase gene in chromosome X protein (BMX)	17 beta-hydroxysteroid dehydrogenase type 1, 2 and 3 (17bHSD1-3)
Mixed lineage leukemia 2 (MLL-2)	Proto-oncogene serine/threonine-protein kinase pim-2
SET and MYND domain containing 3 (SMYD3)	Human DOT1 like, histone H3 methyl-transferase (hDOT1L)
Hepatitis C virus protein 4b (HCV NS4b)	MAP kinase-interacting serine/threonine-protein kinase 1 (MNK1)
Phosphatidylinositol 4-phosphate 5-kinase (PIP5K)	

Classical CDH targets

A survey of the literature reveals that there are many targets known by pharmaceutical companies for which either there is no structure or the structure is of limited use. For example, high-resolution structures of targets, such as telangiectasia and Rad3 gene and ataxia telangiectasia mutated gene, are not known at this time. Similarly, although structures of mesenchymal–epithelial transition factor are known, the published constructs are of limited use because they only work effectively by co-crystallization with a limited number of compounds. Table 1 shows a collection of targets for which high-resolution structures are not known or for which the structures have practical limitations. These are all prime candidates for CDH, and many of these targets are now being explored with this technology.

Application and potential drawbacks of CDH

CDH can be used to tackle the challenging targets described in the previous section because it is faster and generates more variation than alternative techniques for DNA fragmentation such as T-PCR. CDH has been successfully used commercially and in academic research on more than 50 discrete kinases, methyl-transferases, cytoskeletal proteins, insecticidal toxins, transcription activators and polymerases. High-resolution crystal structures have been obtained for several of these, and this has formed the basis of drug discovery programs (data presented at the Sixteenth Protein Structure Determination in Industry Conference 2008, Stansted, UK). CDH is a highly effective technology to tackle proteins commonly found in the cytoplasm of cells. The technology was not designed to solve expression problems of proteins or parts of proteins found as integral parts of cell membranes. In addition, in some cases there is evidence that a predicted domain might have additional amino acid residues located at distal regions of the protein that need to be present to obtain fully folded or active protein, providing a technical challenge to CDH. Interestingly, CDH^2 (described below) could provide a solution to such complex proteins.

Recent and future developments

There have been several recent major developments in CDH technology. One of these is CDH^2 – a technology developed to clone and express protein domains involved in protein–protein interactions. Protein–protein interactions mediate the normal function of cells and, in some cases, if the interaction leads to inappropriate activation of a biochemical pathway, can lead to many complex diseases such as cancer, diabetes and neurodegeneration [37]. This makes developing drugs that target these interactions a key priority for pharmaceutical research.

As with single proteins, developing targets based upon protein–protein interactions requires the generation of soluble, multi-milligram quantities of the drug target. This is extremely difficult because the complex, and every protein in the complex, must be expressed in large quantities. In particular, it has proved

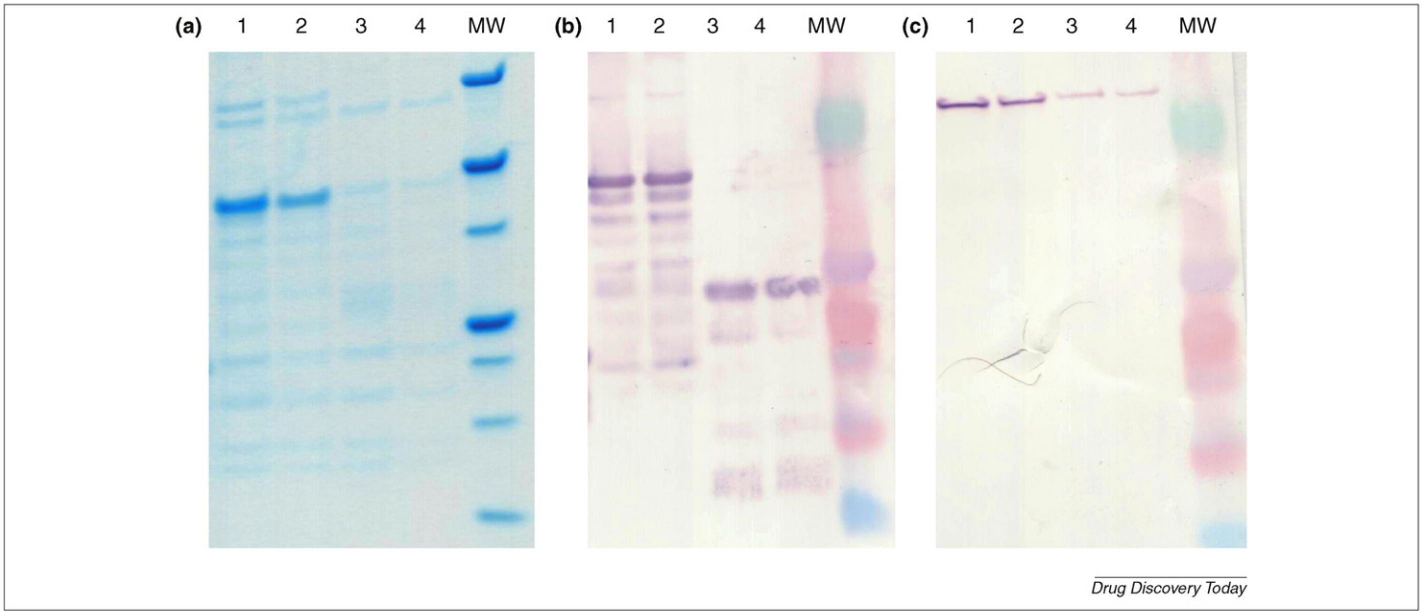

FIGURE 3

CDH^2 analysis of NF-κB and a protein known to interact. A CDH library of a gene coding for a protein known to interact with NF-κB p65 was generated. A series of clones covering the entire coding region of the gene were identified. Each CDH construct from the gene was tagged with a short polyHis region that could be detected with an appropriate monoclonal antibody. Several clones expressing CDH domains covering the entire coded protein were co-expressed with *E. coli* with full-length NF-κB p65 tagged with a Strep tag, which can also be detected with an appropriate monoclonal antibody. After expression, the proteins were extracted from the *E. coli* and mixed with beads containing nickel. The beads were used to purify any protein or protein complexes containing polyHis tags. Panel **(a)** shows a Coomasie-stained polyacrylamide gel of all the proteins that bound to nickel beads from several clones (lanes 1 and 2 are duplicate NF-κB p65 binding proteins, as are lanes 3 and 4). The lane labeled 'MW' contains a series of molecular weight markers. Panel **(b)** is a Western Blot of panel **(a)** probed with antibody that reacts with polyHis tags. As expected, protein reacts in lanes 1–4. Panel **(c)** is a Western Blot of panel **(a)** probed with an antibody that reacts with a Strep tag. Bands are clearly visible in lanes 1–4. The reactivity in panel **(c)** can only be detected if the Strep-labeled NF-κB p65 clone forms a complex with the His-tagged CDH clones. This, therefore, demonstrates that the protein that was thought to form a complex with NF-κB does interact with NF-κB p65. It also shows the region of the protein that interacts (data not shown) and provides reagents suitable to construct assays or for structural studies. Kate Maclagan PhD thesis. The Institute of Structural and Molecular Biology, Research Department of Structural and Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK.

challenging to identify and produce in large quantities the domains of both proteins that are responsible for the interaction. CDH^2 solves these problems.

CDH^2 has been demonstrated in a proof-of-principle study to identify DNA fragments that express interacting regions of human Hsp90β and human Cdc37, which can be expressed as soluble, stable single proteins or as a soluble, stable complex [40]. This study shows how CDH^2 is used where one protein or protein domain in the complex is stable in isolation. If several stable DNA fragments that code for one protein domain have been identified, these can be used in parallel screens of the CDH library.

A stable fragment of Hsp90 incorporating the known interaction site with Cdc37 was identified in a conventional CDH screen [31]. This fragment was encoded into a 'bait' N-terminal domain plasmid of Hsp90 (N-Hsp90) [40] and used to identify stable domains of Cdc37 within 15 000 colonies created from a CDH library. From this, four unique Cdc37 domains were identified. A similar Hsp90β fragment library was searched using a C-terminal domain plasmid of Cdc37 (C-Cdc37). A single Hsp90 fragment was identified, which was found to cover the known interaction site. When the four Cdc37 fragments and Hsp90 fragments were mixed and pull-down experiments performed, complexes between the protein domains produced by CDH^2 between HSP90 and Cdc37 were clearly detected. This system is now suitable for development into a screen that could be used for structural studies or to identify inhibitors of complex formation.

Further examples of protein–protein interactions solved by CDH^2 include NF-κB p65 and a partner protein whose interaction was not previously known (Figure 3). In this case, CDH^2 identified the region of NF-κB p65 that interacts with its partner and the regions of the partner protein that interact with NF-κB p65. This further example provides some excellent starting points for any subsequent study on NF-κB p65 and any interacting partner, in addition to validating the core CDH^2 technology. It is crucial for CDH^2 that either soluble, folded, full-length protein or a domain generated from CDH is used as a bait.

Finally, recent work has begun to develop eCDH – a technology aimed at reproducing the achievements gained by CDH in bacteria, but within a eukaryotic system. This system will have the advantage of enabling some relevant post-translational modification of CDH-derived domains and also expressing the domains in the presence of eukaryotic chaperones or other proteins required for optimal expression and folding of several proteins.

Concluding remarks

CDH, a revolutionary technology with a track record of resurrecting 'dead' gene targets, has recently seen several major advances that have grown its potential still further. CDH can now handle protein–protein interactions, opening up a vast number of highly validated but previously intractable targets. Soon, a eukaryotic version of CDH will also be available, making CDH core to any project involving challenging targets.

Acknowledgements

The author would like to acknowledge the contributions of Laurence Pearl, Paul Driscoll and Chris Prodromou to the CDH technology. In particular, I would like to emphasise the contributions of Stefanie Reich, who has developed CDH into an established technology, and Renos Savva, not only for his contribution to CDH but also for his pioneering work on CDH^2. The author is also indebted to Kate Maclagan for sharing results from her studies on CDH^2.

References

1 Little, P.F. (2005) Structure and function of the human genome. *Genome Res.* 15, 1759–1766

2 Bentley, D.R. (2000) Decoding the human genome sequence. *Hum. Mol. Genet.* 9, 2353–2358

3 Bentley, D.R. (2004) Genomes for medicine. *Nature* 429, 440–445

4 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

5 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351

6 Blundell, T.L. *et al.* (2002) High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov.* 1, 45–54

7 Congreve, M. *et al.* (2005) Structural biology and drug discovery. *Drug Discov. Today* 10, 895–907

8 Baneyx, F. and Mujacic, M. (2004) Recombinant protein folding and misfolding in *Escherichia coli*. *Nat. Biotechnol.* 22, 1399–1408

9 Knaust, R.K. and Nordlund, P. (2001) Screening for soluble expression of recombinant proteins in a 96-well format. *Anal. Biochem.* 297, 79–85

10 Stevens, R.C. (2000) Design of high-throughput methods of protein production for structural biology. *Structure* 8, R177–R185

11 Berrow, N.S. *et al.* (2006) Recombinant protein expression and solubility screening in *Escherichia coli*: a comparative study. *Acta Crystallogr. D Biol. Crystallogr.* 62, 1218–1226

12 Scheich, C. *et al.* (2004) Fast identification of folded human protein domains expressed in *E. coli* suitable for structural analysis. *BMC Struct. Biol.* 4, 4

13 Alzari, P.M. *et al.* (2006) Implementation of semi-automated cloning and prokaryotic expression screening: the impact of SPINE. *Acta Crystallogr. D Biol. Crystallogr.* 62, 1103–1113

14 Busso, D. *et al.* (2003) Expression of soluble recombinant proteins in a cell-free system using a 96-well format. *J. Biochem. Biophys. Methods* 55, 233–240

15 Busso, D. *et al.* (2005) Structural genomics of eukaryotic targets at a laboratory scale. *J. Struct. Funct. Genomics* 6, 81–88

16 Vincentelli, R. *et al.* (2003) Medium-sized structural genomics: strategies for protein expression and crystallization. *Acc. Chem. Res.* 36, 165–172

17 Vincentelli, R. *et al.* (2004) High-throughput automated refolding screening of inclusion bodies. *Protein Sci.* 13, 2782–2792

18 Vincentelli, R. *et al.* (2005) Automated expression and solubility screening of His-tagged proteins in 96-well format. *Anal. Biochem.* 346, 77–84

19 Esposito, D. and Chatterjee, D.K. (2006) Enhancement of soluble protein expression through the use of fusion tags. *Curr. Opin. Biotechnol.* 17, 353–358

20 Hammarstrom, M. *et al.* (2002) Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*. *Protein Sci.* 11, 313–321

21 Hammarstrom, M. *et al.* (2006) Effect of N-terminal solubility enhancing fusion proteins on yield of purified target protein. *J. Struct. Funct. Genomics* 7, 1–14

22 Sorensen, H.P. and Mortensen, K.K. (2005) Soluble expression of recombinant proteins in the cytoplasm of *Escherichia coli*. *Microb. Cell Fact.* 4, 1

23 Sorensen, H.P. and Mortensen, K.K. (2005) Advanced genetic strategies for recombinant protein expression in *Escherichia coli*. *J. Biotechnol.* 115, 113–128

24 Doyle, S.A. *et al.* (2002) High-throughput proteomics: a flexible and efficient pipeline for protein production. *J. Proteome Res.* 1, 531–536

25 Doyle, S.A. *et al.* (2005) Screening for the expression of soluble recombinant protein in *Escherichia coli*. *Methods Mol. Biol.* 310, 115–121

26 Shih, Y.P. *et al.* (2002) High-throughput screening of soluble recombinant proteins. *Protein Sci.* 11, 1714–1719

27 Dyson, M.R. *et al.* (2004) Production of soluble mammalian proteins in *Escherichia coli*: identification of protein features that correlate with successful expression. *BMC Biotechnol.* 4, 32

28 Jacobs, S.A. *et al.* (2005) Soluble domains of telomerase reverse transcriptase identified by high-throughput screening. *Protein Sci.* 14, 2051–2058

29 Christ, D. and Winter, G. (2006) Identification of protein domains by shotgun proteolysis. *J. Mol. Biol.* 358, 364–371

30 Hart, D.J. and Tarendeu, F. (2006) Combinatorial library approaches for improving soluble protein expression in *Escherichia coli*. *Acta Crystallogr. D Biol. Crystallogr.* 62, 19–26

31 Hart, D.J. *et al.* (2006) Combinatorial library approaches for improving soluble protein expression in *Escherichia coli*. *Acta Crystallogr. D Biol. Crystallogr.* 62, 19–26

32 Kawasaki, M. and Inagaki, F. (2001) Random PCR-based screening for soluble domains using green fluorescent protein. *Biochem. Biophys. Res. Commun.* 280, 842–844

33 King, D.A. *et al.* (2006) Domain structure and protein interactions of the silent information regulator Sir3 revealed by screening a nested deletion library of protein fragments. *J. Biol. Chem.* 281, 20107–20119

34 Tarendeau, F. *et al.* (2007) Structure and nuclear import function of the C-terminal domain of influenza virus polymerase PB2 subunit. *Nat. Struct. Mol. Biol.* 14, 229–233

35 Gräslund, S. *et al.* (2008) The use of systematic N- and C-terminal deletions to promote production and structural studies of recombinant proteins. *Protein Expr. Purif.* 58, 210–221

36 Reich, S. *et al.* (2006) Combinatorial Domain Hunting: an effective approach for the identification of soluble protein domains adaptable to high-throughput applications. *Protein Sci.* 15, 2356–2365

37 Pawson, T. and Nash, P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science* 300, 445–452

38 Prodromou, C. *et al.* (2007) DNA fragmentation-based combinatorial approaches to soluble protein expression. Part I. Generating DNA fragment libraries. *Drug Discov. Today* 12, 931–938

39 Savva, R. *et al.* (2007) DNA fragmentation based combinatorial approaches to soluble protein expression. Part II. Library expression, screening and scale-up. *Drug Discov. Today* 12, 939–947

40 Maclagan, K. *et al.* A combinatorial approach to the discovery and delineation of domain mediated protein–protein interactions (in press)