

METHODS FOR HIGH-THROUGHPUT PROTEIN EXPRESSION, PURIFICATION AND STRUCTURE DETERMINATION ADAPTED FOR STRUCTURAL GENOMICS

Pawel Listwan, Jennifer Martin, Bostjan Kobe and Nathan Cowieson

Department of Biochemistry and Molecular Biology, and Institute for Molecular Bioscience, University of Queensland, QLD 4072

Introduction

The Human Genome Project and other sequencing projects continue to identify thousands of genes encoding novel proteins with unknown functions. The functional annotation of these proteins, a field usually referred to as functional genomics, has therefore taken a centre stage in biological research. One approach to functional characterisation takes advantage of the dependence of the function of proteins on their three-dimensional structure, and has been termed structural genomics (or structural proteomics) (1). While structural information cannot always provide complete and accurate functional assignment, recent analyses show that the structure gives some insight into protein function in 75% of cases (2). Integration of bioinformatic and experimental approaches promises further improvements in the area of functional assignment (3).

The principal aim of the worldwide structural genomics initiative is to provide a complete repertoire of protein folds by determining a representative structure (using X-ray crystallography or nuclear magnetic resonance) for each individual protein family (4). While individual consortia approach this goal in different ways and focus on different protein realms, the nature of the structural genomics initiative requires all to share the need for implementing high-throughput methodologies. The tactics include both the adoption of existing technologies in high-throughput mode, and the development of new techniques, at all the stages of the protein structure determination pipeline from gene cloning, protein expression, purification and structure determination to data management. When integrated, these methods could allow thousands of proteins to be fed through the pipeline in a highly automated fashion.

In this paper, we briefly review recent methodology developments implemented by structural genomics initiatives, focusing in particular on the areas of recombinant protein expression in bacteria, purification and crystallization. Many of these methodologies are not restricted to high-throughput approaches, large-scale initiatives or structure determination pipelines, and are as applicable to smaller-scale projects and alternative approaches to the characterisation of biochemical and cellular functions of proteins. Where appropriate, we comment on our own experience in implementing some of the described methods to the structural and functional characterisation of proteins from macrophages (5,6).

Cloning

Following its isolation, each cDNA is traditionally cloned into a chosen expression vector. This requires a careful analysis of the restriction map of a specific cDNA; while such methods are well established and successful, each cDNA requires a different set of sequence-specific restriction enzymes, and therefore they are not easily amenable to high-throughput techniques where one would like to pursue the cloning of tens or even hundreds of cDNAs in parallel. New techniques that eliminate the use of restriction enzymes altogether are therefore much more useful. Two recently developed recombinational cloning approaches meet these criteria. The *cre-lox* recombination introduces a cDNA into a recipient vector by producing a co-integrant plasmid vector of the *cre-lox* system, rather than precise transfer of the open reading frame (ORF) into the expression vector (7). The Gateway technology (8) (Invitrogen, <http://www.invitrogen.com>) instead uses lambda Int/Xis/IHF recombination at *att* sites for transfer of the ORF into the vector (9).

The Gateway technique (Fig. 1) allows considerable flexibility, as an entry vector construct can be transferred into multiple expression (destination) vectors in a single step, for example to incorporate different affinity tags or use different expression hosts. Most of the expression vectors available on the market can be converted into Gateway-compatible vectors.

There are different protocols for cloning into the Gateway entry vectors. Directional topoisomerase-mediated cloning (TOPO) requires four extra nucleotides to be added to the PCR primers (CACC). However, the TOPO reaction requires precise vector-to-insert ratios that need be determined, and is therefore not practical for high-throughput applications. The alternative method to TOPO cloning is the 'BP reaction' that does not require precise insert-to-vector ratios. In our experience, close to 100% of colonies resulting from a BP reaction contain insert-carrying vectors after transformation into *Escherichia coli*, compared to only about 10% for the TOPO reaction. Another benefit of the BP reaction is that the gene of interest can be cloned into the entry vector and recombined into an expression vector in a single reaction, saving time and reagents. One negative aspect of the BP reaction is the need for long (~25 nucleotide) extensions to both the sense and antisense PCR primers. To circumvent this, a nested PCR

approach can be applied, where a single pair of generic primers containing these additional nucleotides is used to extend gene-specific primers with shorter 12-nucleotide overhangs. This nested approach also allows incorporation of a protease cleavage site between the recombination sites and the open reading frame, facilitating the removal of affinity tags after purification. Following cloning into the entry vector, the 'LR reaction' is used to recombine the gene of interest into any one of a range of expression vectors. This reaction is rapid, highly efficient and requires few liquid handling steps.

Bacterial Expression and Purification

Expression vectors used for protein production usually contain a specific tag for affinity purification, such as the hexa-histidine (6xHis), glutathione S-transferase (GST) or maltose-binding protein (MBP); the tags may also improve the solubility of expressed proteins (10). The 6xHis tag is small enough that the expressed proteins can often be crystallised without removing the tag (11). The removal of larger affinity tags is usually required for structural biology applications; although some progress has been made in crystallising proteins containing larger tags, these methods have not yet been used in high-

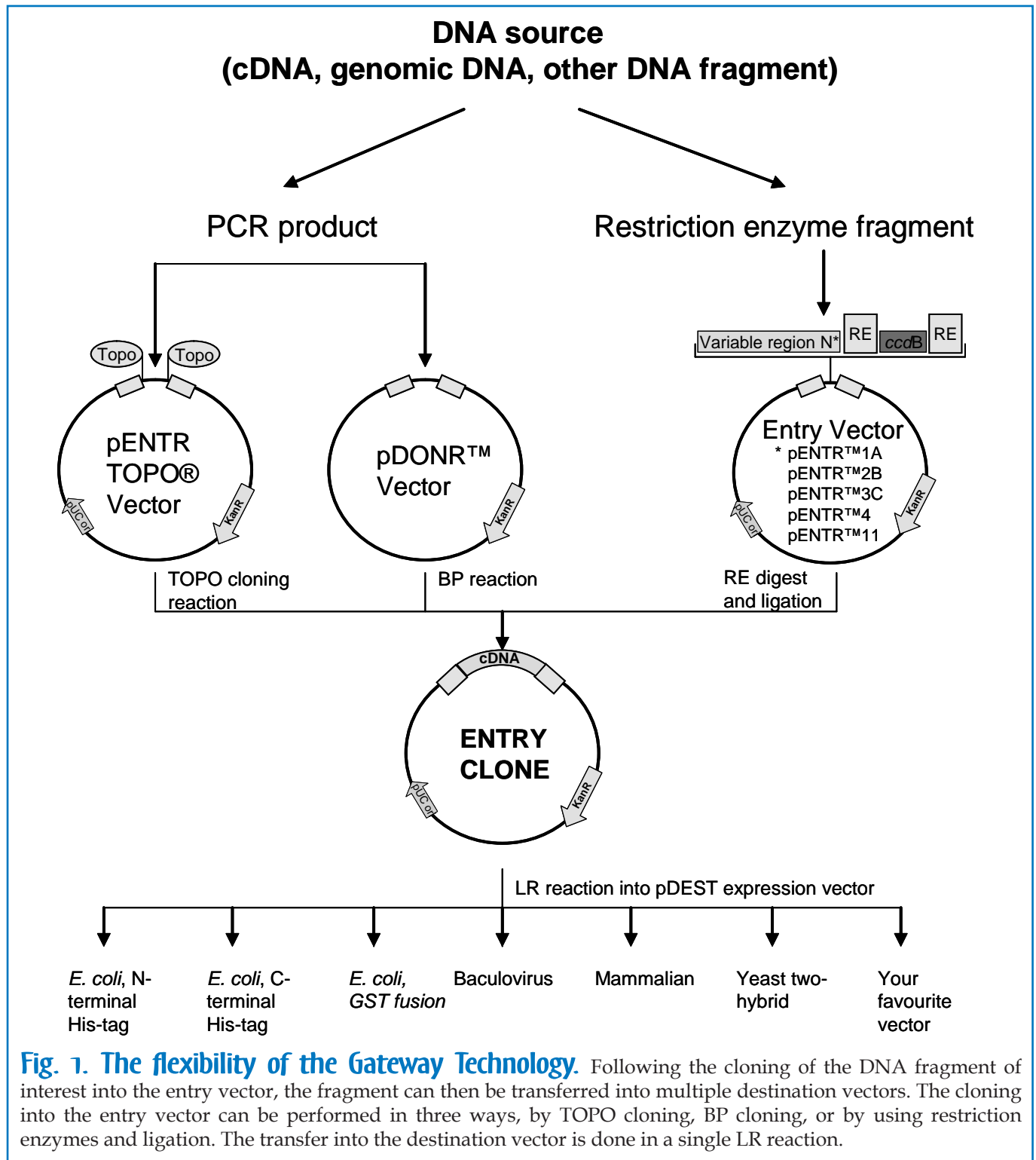


Fig. 1. The flexibility of the Gateway Technology. Following the cloning of the DNA fragment of interest into the entry vector, the fragment can then be transferred into multiple destination vectors. The cloning into the entry vector can be performed in three ways, by TOPO cloning, BP cloning, or by using restriction enzymes and ligation. The transfer into the destination vector is done in a single LR reaction.

throughput applications (12). 6xHis tags additionally allow the proteins to be purified under denaturing conditions, which is useful if refolding of insoluble proteins is to be attempted. In the case of the Gateway expression vector pDEST-17, the N-terminal 6xHis tag is followed by an additional 21 amino acids that arise from transcription through the Gateway recombination and topoisomerase sites. Although there is no mechanism in the pDEST-17 vector for removing these residues, the nested PCR approach can introduce a protease site directly adjacent to the protein.

Protein expression is usually first analysed in small-scale (up to 4 ml) cultures to test for solubility (13). Several bacterial strains are available that may be beneficial for different types of proteins; multiple rounds of trial expression are often used to decide which strain produces the highest levels of protein (14). An increasing number of laboratories use Overnight Express Autoinduction media (Novagen) for induction of protein expression in T7-based vectors (15); this media allow for fully automated growth of cultures to high optical densities without the need for either monitoring growth or induction by IPTG. Even higher cell densities can be obtained using custom 96-tube fermentors (11).

The bacterial cell pellets are lysed by chemical or mechanical means. Typically, one chooses sonication, lysozyme in combination with freeze-thawing cycles or lysis buffers such as B-PER (Pierce), based on available resources and quantity of the proteins to be processed.

Affinity chromatography using recombinant affinity tags has become almost exclusively the first step in protein purification. Recent advances in automation allow large numbers of samples to be processed with little human intervention, yielding proteins of at least 80% purity, sufficient for most applications. Additional purification steps are usually required for protein crystallography, typically ion-exchange and size exclusion chromatography. Protein purity and homogeneity is usually checked by dynamic light scattering (16) and MALDI-TOF mass spectrometry (17) before crystallisation.

Protein Crystallisation

Automation is also being applied to protein crystallography, where the crystallisation step represents the major bottleneck. Searching for diffraction-quality crystals involves the screening of a number of variables including concentrations and inclusion of different ingredients, different pH values and temperatures (18). The combination of these parameters results in a large multi-dimensional space to sample, but in most cases the amount of protein available is small and the number of crystallisation experiments needs to be minimised. One recent advance in this area is the introduction of methodology that allows the set-up of crystallization trials using minute, nanolitre amounts of protein (13,19,20). The two most common methods for initial crystallisation screening are the hanging-drop and sitting-drop vapor diffusion techniques that are both amenable to high throughput application in 96-well

plates. Some other methods including batch and liquid-liquid diffusion are even more suited to the use of robotics and monitoring systems, but are more difficult to set up manually and usually offer lower success rates. The two major uses for robots are in the preparation of crystallisation formulations and in very small-volume crystallisation set-ups; while most liquid handling robots can handle the first task, dedicated robotic systems are usually required for the small volume crystal plate set-ups (e.g. Genomic Solutions Cartesian, Molecular Dimensions Mosquito).

A number of crystallisation screens are available commercially (Hampton Research, Molecular Dimensions, Emerald Biostructures, Jena Bioscience), and new and more efficient screens are being developed based on the success rates of the different formulations in the available screens (21-23). Conventionally, crystallisation is monitored using a light microscope. The use of 96-well plates and large numbers of crystallisation set-ups resulting from high-throughput crystallisation necessitate automated inspection and crystal detection using CCD microscopes and optical recognition systems (e.g. Emerald Biostructures Crystal Monitor).

Automation and high-throughput techniques are also finding their way into the subsequent steps of the structure determination pipeline. Multi-wavelength anomalous dispersion (MAD) (and its single-wavelength equivalent, SAD) using synchrotron radiation is now a routine method for solving crystal structures, and human intervention is continuously being reduced in subsequent computational steps (24).

The Application of High-Throughput Crystallography to the Characterisation of Macrophage Proteins

At the University of Queensland, we have applied the high-throughput approach to the structural characterisation of proteins from mouse macrophages. Macrophages are the first line of defence against pathogens and are implicated in inflammatory and auto-immune disease, and therefore represent a rich source of proteins with therapeutic potential (25). We have combined gene expression analysis by cDNA microarrays and three-dimensional structure determination by X-ray crystallography to define the biochemical and cellular functions of macrophage proteins. Our experimental pipeline is illustrated in Fig. 2. We have applied the Biomek 2000 robotic system (Beckman Coulter) to automate the PCR set-up, cloning, DNA purification and small-scale expression steps. Use of the Caliper AMS 90 SE electrophoresis system has been implemented to replace DNA analysis by agarose gels; the same system will be used in place of SDS-PAGE for protein sample analysis in the near future. Crystallization is performed using a 96-well plate hanging drop setup, and Fluidigm Topaz Crystalliser is used subsequently where appropriate. Soluble expression of proteins in bacteria has been identified as one of the major bottlenecks for mouse proteins, and a substantial effort is directed towards developing inclusion body purification and refolding screens.

Conclusions

The change of the scope of structural biology and the birth of structural genomics has led to the development of new and improved methodologies and automation in every step of the structure determination pipeline. Many of these new approaches are not limited to structural genomics programs, but will find use in smaller-scale projects and fields outside structural biology. It is early days and new experience with existing methodologies is certain to yield further improvements and new imaginative uses of the techniques. For example, the success rates of the expression of eukaryotic proteins in bacteria remain low, but the alternative methods of protein production using baculovirus-infected insect cells and mammalian expression systems currently do not support high throughput applications. Cell-free protein expression systems are much more amenable to automation, however the costs remain high (26). The principal aim is now to progress from a high-throughput to a high-output mode, and to proceed from analysing the "low-hanging fruit" to technically more difficult proteins such as membrane proteins and macromolecular assemblies.

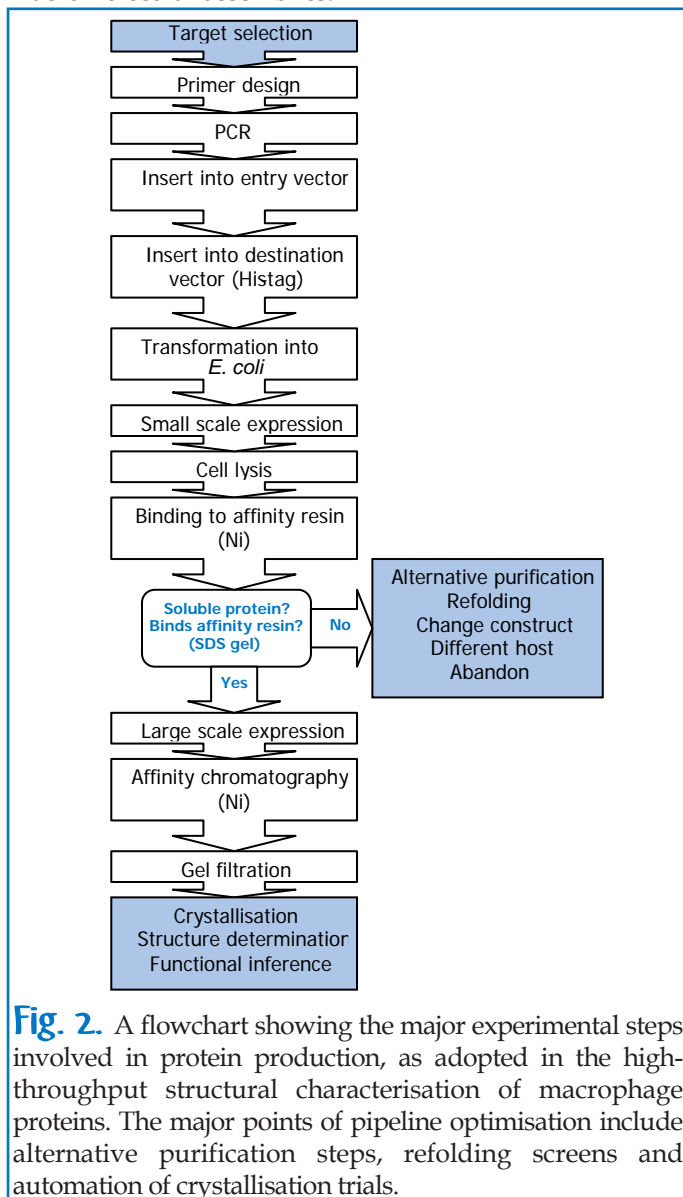


Fig. 2. A flowchart showing the major experimental steps involved in protein production, as adopted in the high-throughput structural characterisation of macrophage proteins. The major points of pipeline optimisation include alternative purification steps, refolding screens and automation of crystallisation trials.

References

- Shapiro, L., and Lima, C.D. (1998) *Structure* **6**, 265-267
- Teichmann, S.A., Murzin, A.G., and Chothia, C. (2001) *Curr. Opin. Struct. Biol.* **11**, 354-363
- Yakunin, A.F., Yee, A.A., Savchenko, A., Edwards, A.M., and Arrowsmith, C.H. (2004) *Curr. Opin. Chem. Biol.* **8**, 42-48
- Terwilliger, T.C., Waldo, G., Peat, T.S., Newman, J.M., Chu, K., and Berendzen, J. (1998) *Protein Sci.* **7**, 1851-1856
- Walsh, C., Hume, D.A., Kobe, B., and Martin, J.L. (2001) *Australian Biochemist* **32**, 13-16
- Kobe, B. (2003) *Acta Chim. Slov.* **50**, 547-562
- Liu, Q., Li, M.Z., Leibham, D., Cortez, D., and Elledge, S.J. (1998) *Curr. Biol.* **8**, 1300-1309
- Walhout, A.J., Temple, G.F., Brasch, M. A., Hartley, J.L., Lorson, M.A., van den Heuvel, S., and Vidal, M. (2000) *Methods Enzymol.* **238**, 575-592
- Hartley, J.L., Temple, G.F., and Brasch, M.A. (2000) *Genome Res.* **10**, 1788-1795
- Hammarstrom, M., Hellgren, N., van den Berg, S., Berglund, H., and Hard, T. (2002) *Protein Sci.* **11**, 313-321
- Lesley, S.A., Kuhn, P., Godzik, A., Deacon, A.M., Mathews, I., Kreusch, A., Spraggon, G., Klock, H.E., McMullan, D., Shin, T., Vincent, J., Robb, A., Brinen, L.S., Miller, M.D., McPhillips, T.M., Miller, M.A., Scheibe, D., Canaves, J.M., Guda, C., Jaroszewski, L., Selby, T.L., Elsliger, M.A., Wooley, J., Taylor, S.S., Hodgson, K.O., Wilson, I. A., Schultz, P.G., and Stevens, R.C. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 11664-11669
- Smyth, D.R., Mrozkievicz, M.K., McGrath, W.J., Listwan, P., and Kobe, B. (2003) *Protein Sci.* **12**, 1313-1322
- Vincentelli, R., Bignon, C., Gruez, A., Canaan, S., Sulzenbacher, G., Tegoni, M., Campanacci, V., and Cambillau, C. (2002) *Acc. Chem. Res.* **36**, 165-172
- Miroux, B., and Walker, J.E. (1996) *J. Mol. Biol.* **260**, 289-298
- Studier, F.W., Rosenberg, A.H., Dunn, J.J., and Dubendorff, J.W. (1990) *Meth. Enzymol.* **185**, 60-89
- Wilson, W.W. (2003) *J. Struct. Biol.* **142**, 56-65
- Leushner, J. (2001) *Exp. Rev. Mol. Diagn.* **1**, 11-18
- Kobe, B., Gleichmann, T., Teh, T., Heierhorst, J., and Kemp, B.E. (1999) in *Protein phosphorylation: A practical approach*, Vol. 2 (Hardie, D.G., ed), Oxford University Press, Oxford
- Stevens, R.C. (2000) *Curr. Opin. Struct. Biol.* **10**, 558-563
- Hansen, C.L., Skordalakes, E., Berger, J.M., and Quake, S.R. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 16531-16536
- Wooh, J.W., Kidd, R.D., Martin, J.L., and Kobe, B. (2003) *Acta Crystallogr. D Biol. Crystallogr.* **59**, 769-772
- Page, R., Grzechnik, S.K., Canaves, J.M., Spraggon, G., Kreusch, A., Kuhn, P., Stevens, R.C., and Lesley, S.A. (2003) *Acta Crystallogr. D Biol. Crystallogr.* **59**, 1028-1037
- Kimber, M.S., Vallee, F., Houston, S., Necakov, A., Skarina, T., Evdokimova, E., Beasley, S., Christendat, D., Savchenko, A., Arrowsmith, C.H., Vedadi, M., Gerstein, M., and Edwards, A.M. (2003) *Proteins* **51**, 562-568
- Lamzin, V.S., and Perrakis, A. (2000) *Nature Struct. Biol.* **7 Suppl.**, 978-981
- Duffield, J.S. (2003) *Clin. Sci. (Lond)*. **104**, 27-38
- Kigawa, T., Yabuki, T., Yoshida, Y., Tsutsui, M., Ito, Y., Shibata, T., and Yokoyama, S. (1999) *FEBS Lett.* **442**, 15-19