Review TRENDS in Biotechnology Vol.23 No.6 June 2005

Making the most of affinity tags

David S. Waugh

Protein Engineering Section, Macromolecular Crystallography Laboratory, Center for Cancer Research, National Cancer Institute at Frederick, P.O. Box B, Frederick, MD 21702-1201, USA

Proteins do not naturally lend themselves to high-throughput analysis because of their diverse physiochemical properties. Consequently, affinity tags have become indispensable tools for structural and functional proteomics initiatives. Although originally developed to facilitate the detection and purification of recombinant proteins, in recent years it has become clear that affinity tags can have a positive impact on the yield, solubility and even the folding of their fusion partners. However, no single affinity tag is optimal with respect to all of these parameters; each has its strengths and weaknesses. Therefore, combinatorial tagging might be the only way to harness the full potential of affinity tags in a high-throughput setting.

Introduction

As we enter the post-genomic era, the focus is shifting from high-throughput analysis of genome sequences to functional and structural studies of the proteins they encode. However, proteins are much more chemically and structurally diverse than nucleic acids, making them intrinsically unsuitable for generic methodology. As a result, the use of genetically engineered affinity tags is the only means of achieving high-throughput protein purification. The variable yield and poor solubility of many recombinant proteins are also major impediments to high-throughput production. However, the yield of recombinant proteins can often be improved by the judicious use of affinity tags and some tags can enhance the solubility, and even promote the proper folding, of their fusion partners. No single tag is ideally suited for all of these purposes. Instead, two or more tags are needed to make the most of affinity tagging in a high-throughput setting. To choose an effective combination, the advantages and disadvantages of various tags must be considered (Table 1) with respect to their ability to increase the yield, enhance the solubility, and facilitate the purification of their fusion partners. Additionally, because affinity tags have the potential to interfere with structural or functional studies, provisions must also be made for removing them.

Increasing the yield of recombinant proteins

A noteworthy advantage of N-terminal tags is that they often improve the yield of recombinant proteins by providing a reliable context for efficient translation initiation. Many different N-terminal tags can confer the advantage of increased yield (Table 1), the only requirement being that ribosomes efficiently initiate translation at the N-terminal methionine residue of the tag. When problems arise, it is often because a secondary structure in the mRNA interferes with the binding of ribosomes [1]. Deleterious secondary structures are more likely to occur in conjunction with small N-terminal tags because short-range RNA–RNA interactions tend to be more stable than long-range interactions.

Sequence determinants at both the N- and C-termini of proteins can influence their rate of degradation [2–4]. Therefore, in some cases affinity tags might improve the yield of recombinant proteins by rendering them more resistant to intracellular proteolysis [5]. At present, however, the evidence for this is largely anecdotal.

Enhancing the solubility of recombinant proteins

According to recent data from several large structural genomics centers, more than half of all recombinant proteins accumulate in the form of insoluble aggregates when they are overproduced in *Escherichia coli*, irrespective of their origin [6]. This poses a major obstacle to high-throughput protein production. There is a general perception that solubility problems can often be solved by using a eukaryotic host, such as insect cells (the baculovirus expression system) or yeast, but systematic studies are lacking. Another promising alternative is cell-free protein synthesis, which has improved dramatically in recent years. Although these systems might represent the wave of the future, *E. coli* is still the work-horse of heterologous expression systems, hence, some means of overcoming the solubility problem in this host is needed.

The most common method that is used to improve the solubility of recombinant proteins in *E. coli* is to reduce the temperature at which the target protein is being produced, although this is not always effective [7]. However, for some time now it has been recognized that certain affinity tags have the ability to promote the solubility of their fusion partners [8]. Solubility-enhancing affinity tags tend to be proteins rather than peptides. Originally, it was believed that virtually any highly soluble protein could act as a solubility enhancer, but this has not proved to be the case. In parallel comparisons, several studies have shown that certain soluble proteins are consistently more effective than others [9–11]. The mechanism by which these carrier proteins promote the solubility of their fusion partners remains poorly understood and might not be universal. Perhaps the best studied and most thoroughly validated solubility-enhancing proteins are the *E. coli* maltose-binding protein (MBP) [9] and N-utilization substance A (NusA) [12]. The maltose-binding protein is also a natural affinity tag, making it a

Corresponding author: Waugh, D.S. (waughd@ncifcrf.gov).
Available online 12 April 2005

 0167-7799/$ - see front matter doi:10.1016/j.tibtech.2005.03.012

Table 1. Advantages and disadvantages of some commonly used fusion partners

Tag[a]	Advantages	Disadvantages
GST	Efficient translation initiation	High metabolic burden
	Inexpensive affinity resin	Homodimeric protein
	Mild elution conditions	Does not enhance solubility
MBP	Efficient translation initiation	High metabolic burden
	Inexpensive affinity resin	
	Enhances solubility	
	Mild elution conditions	
NusA	Efficient translation initiation	High metabolic burden
	Enhances solubility	
	Not an affinity tag	
Thioredoxin	Efficient translation initiation	Not an affinity tag[b]
	Enhances solubility	
Ubiquitin	Efficient translation initiation	Not an affinity tag
	Might enhance solubility	
FLAG	Low metabolic burden	Expensive affinity resin
	High specificity	Harsh elution conditions
BAP	Low metabolic burden	Expensive affinity resin
	Mild elution conditions	Variable efficiency of enzymatic biotinylation
	Provides convenient means of immobilizing proteins in a directed orientation	Co-purification of *E. coli* biotin carboxyl carrier protein on affinity resin
		Does not enhance solubility
His_6	Low metabolic burden	Specificity of IMAC is not as high as other affinity methods
	Inexpensive affinity resin	
	Mild elution conditions	
	Tag works under both native and denaturing conditions	Does not enhance solubility
STREP	Low metabolic burden	Expensive affinity resin
	High specificity	Does not enhance solubility
	Mild elution conditions	
SET	Enhances solubility	Not an affinity tag
CBP	Low metabolic burden	Expensive affinity resin
	High specificity	Does not enhance solubility
	Mild elution conditions	
S-tag	Low metabolic burden	Expensive affinity resin
	High specificity	Harsh elution conditions (or on-column cleavage)
		Does not enhance solubility

[a]GST, glutathione S-transferase; MBP, maltose-binding protein; NusA, N-utilization substance A; FLAG, FLAG-tag peptide; BAP, biotin acceptor peptide; His_6, hexahistidine tag; STREP, streptavidin-binding peptide; SET, solubility-enhancing tag; CBP, calmodulin-binding peptide.
[b]Derivatives of thioredoxin have been engineered to have affinity for immobilized metal ions (His-patch thioredoxin) or avidin/streptavidin [38].

particularly attractive choice as a solubility enhancer. Recently, a new type of solubility enhancing polypeptide (SET tag, Invitrogen) has been described [13]. Unlike MBP and NusA, the SET tag is not thought to adopt a stable globular fold. It has been proposed that this highly acidic tag inhibits aggregation by increasing electrostatic repulsion between nascent polypeptides [13], a conjecture that clearly merits further study.

Not every protein can be rendered soluble by fusing it to a solubility-enhancing tag. Moreover, some proteins still form insoluble aggregates after they are cleaved from a solubility-enhancing partner. Nevertheless, it is clear that this approach can lead to the recovery of more soluble, properly folded proteins than could be obtained without a solubility enhancer. Even if only one-quarter of insoluble proteins could be recovered in soluble form using this approach, which is probably an underestimate [14], this would represent a substantial increase in output (i.e. efficiency).

An N-terminal solubility tag can help to alleviate two bottlenecks at once by also providing an optimum context for translation initiation, as discussed previously. It is worth noting that the solubility-enhancing activity of MBP is manifest only if it is the N-terminal fusion partner [15]. It is not known whether the same is true of NusA. By contrast, the SET tag evidently can act as a solubility enhancer when fused to the C-terminus of a recombinant protein [13].

Although it has been firmly established that certain highly soluble proteins, such as MBP and NusA, can function as general solubility enhancers in the context of a fusion protein, virtually nothing is known about how these proteins compare in terms of their ability to promote the proper folding of their fusion partners. It is possible that solubility enhancers will differ markedly in their ability to promote the folding of their fusion partners, with some consistently outperforming others. This would imply that a solubility enhancer has an active role in the folding of its passenger proteins. Alternatively, the folding efficiency could depend primarily on the passenger protein rather than the solubility enhancer, which would be indicative of a more passive role for the enhancer. A third possibility is that neither of these trends will hold, and that multiple solubility enhancers will have to be tested to find the optimal partner for each passenger protein. This is clearly a crucial question that needs to be addressed in the future.

Facilitating the purification of recombinant proteins

High-throughput protein production must, almost by definition, entail a generic strategy for purification. This is obviously a key attribute of affinity tags. In principal, one can design affinity-based purification protocols so that the target protein never has to interact directly with a chromatographic matrix. This is an important advantage because the interaction energies of proteins with conventional matrices (e.g. ion exchange) are of the same order as their net free energy of stability [16]. When the tightest binding conformation of a protein is not the same as its native conformation, contact denaturation can occur. Hence, affinity methods might make it possible to purify proteins that would be difficult or even impossible to obtain by traditional techniques.

Generally, affinity tags fall into one of two categories: proteins that recognize small ligands or peptides that bind to immobilized proteins. Comprehensive reviews of affinity tags have been published recently [17,18]. Although many different tags have been described in the literature, few of them have been tested in a high-throughput context. Peptide epitopes like the FLAG-tag [19], the calmodulin-binding peptide [20], the Strep-tag or Strep-tag II [21,22] and the biotin acceptor peptide [23] all

exhibit a high degree of specificity for their cognate binding partners. However, the resins (immobilized proteins) that they interact with tend to be expensive, are easily fouled and have relatively low binding capacities, making them less than ideal for high-throughput applications. By contrast, large protein tags usually recognize small ligands that make for less expensive and more robust chromatography matrices. The most popular examples are glutathione S-transferase (GST) [24] and MBP [25]. One disadvantage of GST is that it is a homodimer [26], which can complicate purification of fusion proteins and renders this affinity tag unsuitable for the isolation of oligomeric proteins. Moreover, whereas MBP contains no cysteine residues, the four solvent exposed cysteines in each subunit of the GST dimer can lead to a significant degree of oxidative aggregation [26]. A disadvantage of large protein affinity tags in general is that they devour more metabolic energy during overproduction than small tags.

The hexahistidine tag (His-tag), which binds to immobilized transition metals, is by far the most commonly used affinity tag for high-throughput protein purification. Virtually all large structural genomics centers use immobilized metal affinity chromatography (IMAC) as their principal affinity strategy. Ni(II)-nitrilotriacetic acid (Ni-NTA), which exhibits a high affinity for adjacent histidine residues, is the most commonly used matrix for IMAC [27,28]. The His-tag combines the advantages of small size with the added benefit of interacting with a chromatographic matrix (e.g. Ni-NTA resin) that is relatively inexpensive, able to withstand multiple regeneration cycles under stringent sanitizing conditions, and exhibits a high binding capacity. Moreover, elution conditions are mild and flexible (100–250 mM imidazole, pH $<$ 5.0, or 10 mM EDTA). The His-tag also works well under denaturing conditions, adding yet another dimension to its versatility. If a His-tagged recombinant protein is insoluble in *E. coli*, then it can still be purified by IMAC under denaturing conditions and refolded.

Removal of affinity tags

All tags, whether large or small, have the potential to interfere with the biological activity of a protein, impede its crystallization, or otherwise influence its behavior. Consequently, it is usually desirable to remove the tag. This has always been the Achilles' heel of the fusion approach. Although highly specific endoproteases, such as those encoded by the tobacco etch virus (AcTEV, Invitrogen) [29] and the human rhinovirus (PreScission, Amersham Biotech.) [30], have largely mitigated the problem of nonspecific cleavage, processing efficiency varies with each fusion protein in an unpredictable manner. Processing efficiency can often be improved by using more protease over a longer period of time or by incorporating extra residues (e.g, polyalanine) adjacent to the protease recognition site. Nevertheless, one must be prepared to accept that some fraction of fusion proteins will be resistant to digestion and that these cases will have to be dealt with on a low-throughput basis.

Whereas the nominal sequence specificity of factor Xa (IEGR/) and enterokinase (DDDDK/) should enable these enzymes to generate target proteins with native N-termini, they are rather promiscuous and often cleave fusion proteins at locations other than the desired site [31,32]. By contrast, proteases with more stringent sequence specificity, like TEV (ENLYFQ/G) and PreScission (LEVLFQ/GP), have specificity determinants that are located on the C-terminal side of the scissile bond (/) and which will remain on the target protein after the removal of an N-terminal tag. However, at least in the case of TEV protease, it is possible to produce many proteins with no extra residues on their N-termini because this enzyme can tolerate a variety of residues in the position normally occupied by glycine with only a modest reduction in processing efficiency [33].

The removal of C-terminal tags poses a greater problem because all endoproteases that are used to remove affinity tags recognize specificity determinants that are situated mainly on the N-terminal side of the scissile bond. As a result, the endoproteolytic removal of a C-terminal tag will always leave at least 4–6 extra non-native residues on the C-terminus of the target protein. Carboxypeptidases have been used to a limited extent to remove short C-terminal tags (e.g. [27,34]). In the future, more effort needs to be expended on the development of carboxypeptidases with sequence specificity that can be controlled to limit the extent of digestion and improve the homogeneity of the product.

Finally, it should be noted that self-processing fusion partners derived from inteins (protein introns that autocatalytically splice from their host proteins) represent an alternative to conventional proteolytic cleavage in *trans* [35]. However, intein-mediated cleavage has not yet been tested in a high-throughput context. The principal drawbacks of the intein approach are: (i) the large size of the catalytic machinery that must be incorporated into the fusion protein, which increases the metabolic burden on the cells; (ii) the dependence of processing efficiency on the sequence context at the fusion junction; (iii) the slow rate of autoprocessing; and (iv) the fact that inteins neither enhance the solubility nor facilitate the purification of their fusion partners.

Combinatorial tagging

No affinity tag is ideal from every standpoint. Therefore, combinatorial tagging appears to be the only means of deriving the maximum possible benefit from affinity tags. Several groups have explored this possibility [36–40]. One combinatorial tagging strategy that is currently employed at the Macromolecular Crystallography Laboratory (NCI Frederick) [41], the Berkeley Structural Genomics Center [42] and the Center for Eukaryotic Structural Genomics (http://www.uwstructuralgenomics.org), uses a dual His_6–MBP affinity tag (Figure 1). The MBP moiety improves the yield and enhances the solubility of the passenger protein while the His_6-tag facilitates its purification. The fusion protein (His_6–MBP–passenger) is purified by IMAC on Ni-NTA resin and then cleaved *in vitro* with His_6-tagged TEV protease [29] to separate the His_6–MBP from the passenger protein. In the final step, the unwanted byproducts of the digest (His_6–MBP, His_6–TEV protease, and any undigested fusion protein)

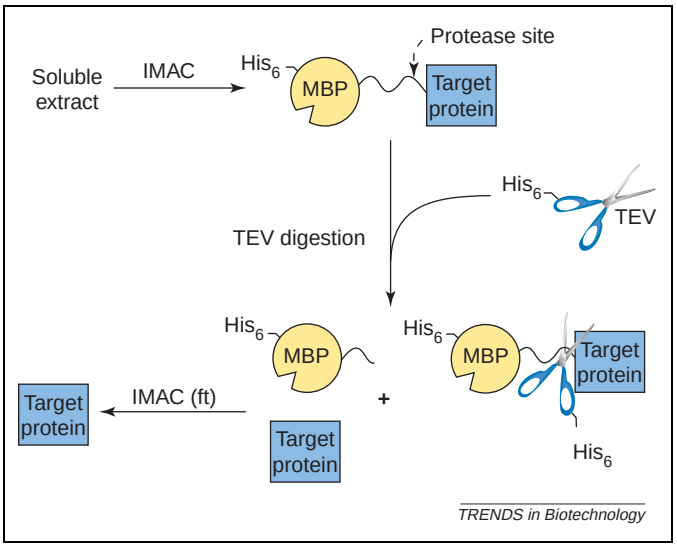

Figure 1. Schematic illustration of a generic method for high-throughput protein expression and purification in *Escherichia coli* that uses a dual His_6–MBP affinity tag. Abbreviations: ft, flow-through (unbound) fraction; IMAC, immobilized metal affinity chromatography; MBP, maltose-binding protein; TEV, tobacco etch virus protease.

are absorbed by a second round of IMAC, leaving nothing but the pure passenger protein in the flow-through fraction. A key advantage of this approach is that any endogenous proteins that bind nonspecifically to the Ni-NTA resin during the first IMAC step also do so during the second round of IMAC. Hence, the application of two successive IMAC steps, rather than just one, is the key to obtaining protein of a high purity with just a single affinity technique. The affinity of MBP for amylose resin is not exploited in this protocol but it could be if desired. Moreover, because only the His-tag is used for purification, other yield- and solubility-enhancing proteins that are not natural affinity tags (e.g. NusA) could potentially be employed in a similar configuration. Finally, it should be noted that combinatorial tagging might also prove to be a convenient means of purifying heterodimeric or even higher-order protein complexes in which each polypeptide is fused to a different tag.

Future directions

The production of recombinant proteins has traditionally been viewed as a means to an end rather than a legitimate field of scientific inquiry in its own right. Unfortunately, therefore, systematic comparisons of different affinity tagging strategies are still relatively rare. Large-scale structural genomics centers around the world have made great strides in the development of automated tools for high-throughput protein expression and purification, although most of them have experimented with only a few different strategies for the production of proteins thus far. By using a fraction of their technological resources to address basic research questions, as outlined here, these centers could make important contributions to the field that are beyond the capability of the independent researcher.

References

1 Hartz, D. *et al*. (1991) Influence of mRNA determinants on translation initiation in *Escherichia coli*. *J. Mol. Biol.* 218, 99–105
2 Varshavsky, A. (1997) The N-end rule pathway of protein degradation. *Genes Cells* 2, 13–28
3 Flynn, J.M. *et al*. (2003) Proteomic discovery of cellular substrates of the ClpXP protease reveals five classes of ClpX-recognition signals. *Mol. Cell* 11, 671–683
4 Hayes, C.S. *et al*. (2002) Proline residues at the C terminus of nascent chains induce SsrA tagging during translation termination. *J. Biol. Chem.* 277, 33825–33832
5 De Marco, V. *et al*. (2004) The solubility and stability of recombinant proteins are increased by their fusion to NusA. *Biochem. Biophys. Res. Commun.* 322, 766–771
6 Chayen, N.E. (2004) Turning protein crystallization from an art into a science. *Curr. Opin. Struct. Biol.* 14, 577–583
7 Schein, C.H. (1989) Production of soluble recombinant proteins in bacteria. *Biotechnology* 7, 1141–1148
8 Baneyx, F. (1999) Recombinant protein expression in *Escherichia coli*. *Curr. Opin. Biotechnol.* 10, 411–421
9 Kapust, R.B. and Waugh, D.S. (1999) *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* 8, 1668–1674
10 Shih, Y-P. *et al*. (2002) High-throughput screening of soluble recombinant proteins. *Protein Sci.* 11, 1714–1719
11 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
12 Davis, G.D. *et al*. (1999) New fusion protein systems designed to give soluble expression in *Escherichia coli*. *Biotechnol. Bioeng.* 65, 382–388
13 Zhang, Y-B. *et al*. (2004) Protein aggregation during overexpression limited by peptide extensions with large net negative charge. *Protein Expr. Purif.* 36, 207–216
14 Goulding, C.W. and Perry, L.J. (2003) Protein production in *Escherichia coli* for structural studies by X-ray crystallography. *J. Struct. Biol.* 142, 133–143
15 Sachdev, D. and Chirgwin, J.M. (1998) Order of fusions between bacterial and mammalian proteins can determine solubility in *Escherichia coli*. *Biochem. Biophys. Res. Commun.* 244, 933–937
16 Gerstner, J.A. *et al*. (1994) Gibbs free energy of adsorption for biomolecules in ion-exchange systems. *Biophys. Chem.* 52, 97–106
17 Hearn, M.T.W. and Acosta, D. (2001) Applications of novel affinity cassette methods: use of peptide fusion handles for the purification of recombinant proteins. *J. Mol. Recognit.* 14, 323–369
18 Terpe, K. (2003) Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. *Appl. Microbiol. Biotechnol.* 60, 523–533
19 Einhauer, A. and Jungbauer, A. (2001) The FLAG peptide, a versatile fusion tag for the purification of recombinant proteins. *J. Biochem. Biophys. Methods* 49, 455–465
20 Vaillancourt, P. *et al*. (2000) Affinity purification of recombinant proteins fused to calmodulin or calmodulin-binding peptides. *Methods Enzymol.* 326, 340–362
21 Schmidt, T.G. and Skerra, A. (1993) The random peptide library-assisted engineering of a C-terminal affinity peptide, useful for the detection and purification of a functional Ig Fv fragment. *Protein Eng.* 6, 109–122
22 Voss, S. and Skerra, A. (1997) Mutagenesis of a flexible loop in streptavidin leads to higher affinity for the Strep-tag II peptide and improved performance in recombinant protein purification. *Protein Eng.* 10, 975–982
23 Schatz, P.J. (1993) Use of peptide libraries to map the substrate specificity of a peptide-modifying enzyme: a 13 residue consensus peptide specifies biotinylation in *Escherichia coli*. *Biotechnology* 11, 1138–1143
24 Smith, D.B. and Johnson, K. (1988) Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. *Gene* 67, 31–40
25 Di Guan, C. *et al*. (1988) Vectors that facilitate the expression and purification of foreign peptides in *Escherichia coli* by fusion to maltose-binding protein. *Gene* 67, 21–30
26 Kaplan, W. *et al*. (1997) Conformational stability of pGEX-expressed *Schistosoma japonicum* glutathione S-transferase: a detoxification enzyme and fusion-protein affinity tag. *Protein Sci.* 6, 399–406
27 Houchili, E. *et al*. (1988) Genetic approach to facilitate purification of recombinant proteins with a novel metal chelate adsorbent. *Biotechnology* 6, 1321–1325

28 Gaberc-Porekar, V. and Menart, V. (2001) Perspectives of immobilized-metal affinity chromatography. *J. Biochem. Biophys. Methods* 49, 335–360
29 Kapust, R.B. *et al*. (2001) Tobacco etch virus protease: mechanism of autolysis and rational design of stable mutants with wild-type catalytic proficiency. *Protein Eng.* 14, 993–1000
30 Walker, P.A. *et al*. (1994) Efficient and rapid purification of proteins using recombinant fusion proteases. *Biotechnology* 12, 601–605
31 Choi, S.I. *et al*. (2001) Recombinant enterokinase light chain with affinity tag; expression from *Saccharomyces cerevisiae* and its utilities in fusion protein technology. *Biotechnol. Bioeng.* 75, 718–724
32 Jenny, R.J. *et al*. (2003) A critical review of the methods for cleavage of fusion proteins with thrombin and factor Xa. *Protein Expr. Purif.* 31, 1–11
33 Kapust, R.B. *et al*. (2002) The P1′ specificity of tobacco etch virus protease. *Biochem. Biophys. Res. Commun.* 294, 949–955
34 Hirabayashi, J. and Kasai, K. (1992) Arginine-tail method, an affinity tag procedure utilizing anhydrotrypsin agarose. *J. Chromatogr.* 597, 181–187
35 Paulus, H. (2000) Protein splicing and related forms of autoprocessing. *Annu. Rev. Biochem.* 69, 447–496
36 Nilsson, J. *et al*. (1996) Multiple affinity domains for the detection, purification and immobilization of recombinant proteins. *J. Mol. Recognit.* 9, 585–594
37 Pryor, K.D. and Leiting, B. (1997) High-level expression of soluble protein in *Escherichia coli* using a His_6-tag and maltose-binding protein double-affinity fusion system. *Protein Expr. Purif.* 10, 309–319
38 Smith, P.A. *et al*. (1998) A plasmid expression system for quantitative *in vivo* biotinylation of thioredoxin fusion proteins in *Escherichia coli*. *Nucleic Acids Res.* 26, 1414–1420
39 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
40 Routzahn, K.M. and Waugh, D.S. (2002) Differential effects of supplementary affinity tags on the solubility of MBP fusion proteins. *J. Struct. Funct. Genomics* 2, 83–92
41 Tropea, J.E. *et al.* A generic method for the production of recombinant proteins in *Escherichia coli* using a dual His_6-MBP affinity tag. *Methods Mol. Biol.* (in press)
42 Nguyen, H. *et al*. (2004) An automated small-scale protein expression and purification screening provides beneficial information for protein production. *J. Struct. Funct. Genomics* 5, 23–27