

Co-expression of protein complexes in prokaryotic and eukaryotic hosts: experimental procedures, database tracking and case studies

Christophe Romier,^{a,‡} Marouane Ben Jelloul,^{b,‡} Shira Albeck,^c Gretel Buchwald,^b Didier Busso,^a Patrick H. N. Celie,^b Evangelos Christodoulou,^b Valeria De Marco,^b Suzan van Gerwen,^b Puck Knipscheer,^b Joyce H. Lebbink,^b Valerie Notenboom,^b Arnaud Poterszman,^a Natacha Rochel,^a Serge X. Cohen,^b Tamar Unger,^c Joel L. Sussman,^c Dino Moras,^a Titia K. Sixma^b and Anastassis Perrakis^{b*}

^aInstitut de Génétique et Biologie Moléculaire et Cellulaire, 1 Rue Laurent Fries, BP 163, 67404 Illkirch CEDEX, France, ^bDivision of Molecular Carcinogenesis, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands, and ^cThe Israel Structural Proteomics Center, The Department of Structural Biology, The Weizmann Institute of Science, Rehovot 76100, Israel

‡ These authors contributed equally to this work.

Correspondence e-mail: a.perrakis@nki.nl

Structure determination and functional characterization of macromolecular complexes requires the purification of the different subunits in large quantities and their assembly into a functional entity. Although isolation and structure determination of endogenous complexes has been reported, much progress has to be made to make this technology easily accessible. Co-expression of subunits within hosts such as *Escherichia coli* and insect cells has become more and more amenable, even at the level of high-throughput projects. As part of SPINE (Structural Proteomics In Europe), several laboratories have investigated the use co-expression techniques for their projects, trying to extend from the common binary expression to the more complicated multi-expression systems. A new system for multi-expression in *E. coli* and a database system dedicated to handle co-expression data are described. Results are also reported from various case studies investigating different methods for performing co-expression in *E. coli* and insect cells.

1. Introduction

Most functional units within the eukaryotic cell are assemblies of proteins, of nucleic acids or of proteins and nucleic acids, rather than single macromolecules (Gavin & Superti-Furga, 2003). Proper characterization of a cellular function has to be carried out on such complexes or at least on a subset of these multi-component entities. One major challenge of the post-genomic era is to produce these complexes in sufficient amounts to be studied by biochemical and structural means. Several studies have shown the feasibility of purifying endogenous complexes for structure determination, including RNA polymerase II (Cramer *et al.*, 2000) and the ribosome (Yusupov *et al.*, 2001). More recently, technical advances such as Tap-tagging have allowed easier purification of large multi-protein complexes (Dziembowski & Séraphin, 2004). This latter approach is currently restricted by the low abundance of many complexes within the cell. Techniques such as *in vitro* reconstitution from separately purified components can be used to study small or mid-size assemblies. The major drawback of this technique is that it is relatively slow and often requires refolding steps: in many cases, proteins that form complexes in cells are unfolded without their cellular partners in a heterologous expression system.

Co-expression of multiple proteins in the same cell has emerged as a good compromise between endogenous purification and *in vitro* reconstitution from individually expressed components. Co-expression offers the possibility of cofolding protein partners that would otherwise be insoluble if

Received 15 November 2005

Accepted 7 August 2006

expressed alone and can enable *in vivo* reconstitution with higher yields of the desired complex. An additional important advantage is the possibility of deciphering protein–protein interactions within a complex *in vivo*, either in *Escherichia coli* (Li *et al.*, 1997; Copeland, 1997; Johnston *et al.*, 2000; Fribourg *et al.*, 2001) or also in eukaryotic cells such as insect cells (Jawhari *et al.*, 2002; Berger *et al.*, 2004 and references therein). Limitations or difficulties are also observed with co-expression, such as decreasing yields upon increasing the number of expressed genes (Johnston *et al.*, 2000). One way to solve this problem, especially in *E. coli*, is the use of micro-fermentors with media reaching high cell densities (see Geerlof *et al.*, 2006) or the use of the ‘auto-induction’ media (Studier, 2005) that allows the *E. coli* to reach considerably higher cell densities (typically five to ten times more than LB medium) in flasks. Furthermore, complex formation is sometimes not observed when too long or too short constructs are used or when one of the proteins bears a purification tag (Fribourg *et al.*, 2001). Here, again, a high-throughput robotized strategy should perfectly be suited to overcome these problems by allowing the testing of multiple constructs, with different tags, in a systematic and efficient manner.

Having within SPINE (Structural Proteomics In Europe) a major interest in the structure determination of protein complexes, we have investigated the use of co-expression both in *E. coli* and in recombinant baculovirus-infected insect cells in order to solve the difficulties encountered in our different projects. This report reviews the different techniques used to perform co-expression experiments, both in *E. coli* and in insect cells. A novel multi-expression system is also presented which is well suited for high-throughput strategies. Several case studies are discussed both for the *E. coli* and insect-cell systems.

The organization and archiving of data has been a central concern of the SPINE project (see Bahar *et al.*, 2006). The development of a Laboratory Information Management System (LIMS) for Structural Biology and Genomics has made great progress (Prilusky *et al.*, 2005), but it has various shortcomings especially when complicated data, such as information on expression of protein complexes, has to be handled. We present a database system that efficiently addresses issues about handling data from protein complexes and is considered as a prototype/case study towards an integrated LIMS system for Structural Biology, such as the Protein Information Management System (PIMS) currently being developed in Europe (<http://www.pims-lims.org>).

2. Materials and methods

Many different techniques are used to co-express proteins in cells. The short description that follows is limited to the *E. coli* and insect-cell systems, which are the two most common hosts within our laboratories; we are however aware that other kind of cells have been used for co-expression, *e.g.* yeast and mammalian cells (Burgers, 1999; Aricescu, Assenberg *et al.*, 2006).

2.1. Co-expression of protein complexes in *E. coli*

2.1.1. Multiple vectors. In *E. coli*, the easiest approach for co-expressing proteins is to use vectors with different resistance markers, each vector bearing a single gene. Vectors with the same origin of replication can be used as long as their resistance markers are different. However, plasmids with same origin of replication can ‘compete’ inside a cell and one of the plasmids might end up being less amplified than the others within the cell (especially in high-density cultures). This can lead to complexes with poor stoichiometry or even missing subunits. In this respect, a theoretically sound and practically proven approach is to use vectors with different origins of replication (Copeland, 1997; Johnston *et al.*, 2000; Fribourg *et al.*, 2001). More than two vectors can be used, but the question of the correct maintenance of all the plasmids within the cell remains since the amplification of some of them could be down-regulated.

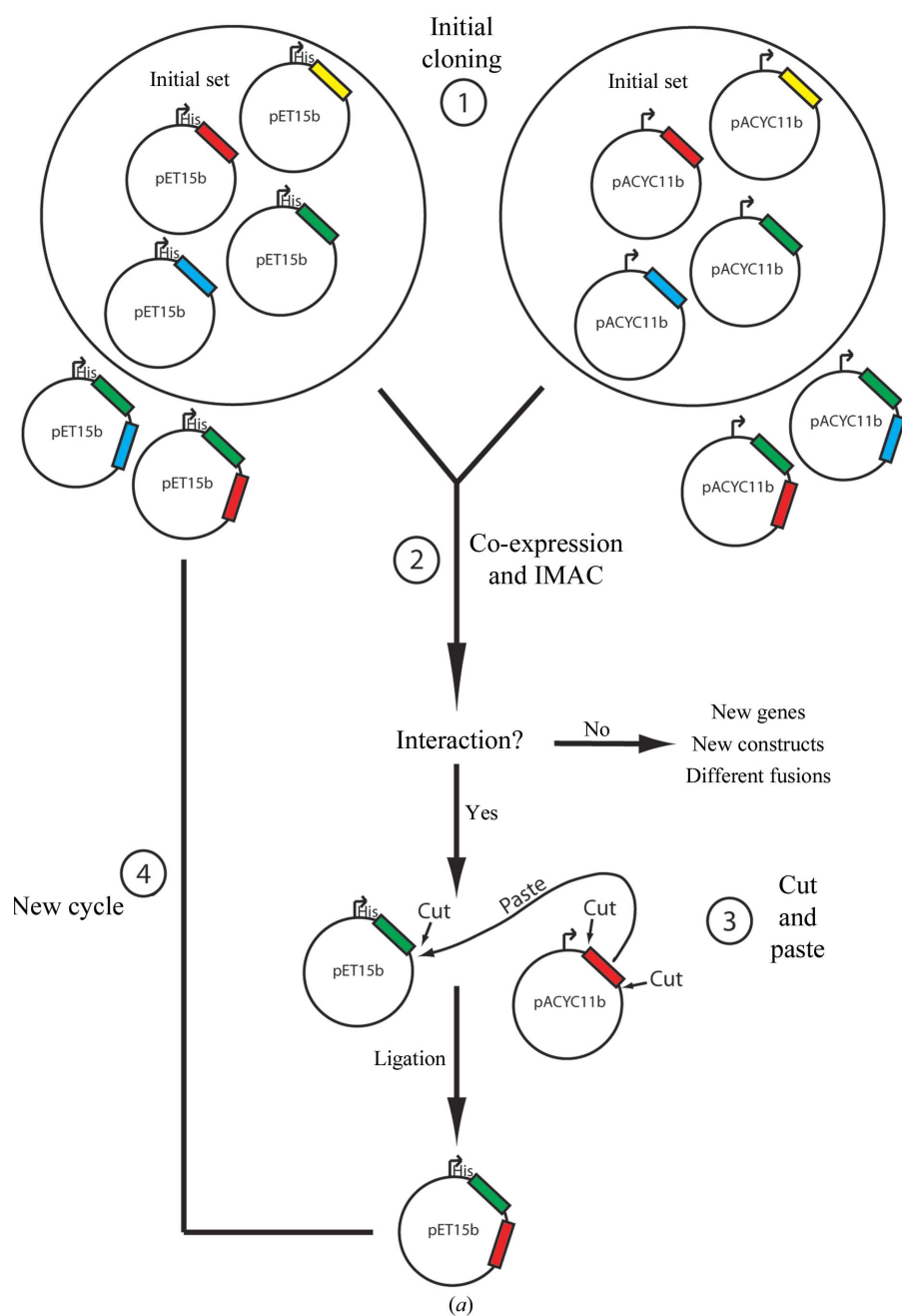
2.1.2. Single vector, single RNA transcript. Another co-expression strategy is to have the various genes cloned on a single vector. This can be achieved by cloning all genes under the control of a single promoter, each gene having its own ribosome-binding site (RBS; single-transcript; Li *et al.*, 1997; Stebbins *et al.*, 1999; Buhler *et al.*, 2001; Neumann *et al.*, 2003; Tan *et al.*, 2005). This strategy results in a long polycistronic mRNA: its length is restricted by the capabilities of the polymerase used and the intrinsic stability of mRNA. Moreover, the efficiency of ribosome binding to the RBS depends on mRNA structure. Typically, a linker DNA sequence between the end of one gene and the RBS for the next one needs to be introduced. Expression levels and efficiency are sometimes, albeit not always, crucially dependent on the order with which the complex components appear on the common mRNA.

2.1.3. Single vector, multiple RNA transcripts. The use of separate promoters for each gene on a single plasmid (multi-transcripts; Novy *et al.*, 2002; Loomis *et al.*, 2003; Alexandrov *et al.*, 2004; Kim *et al.*, 2004) results in multiple RNA transcripts, offering an alternative to the method described in §2.1.2. In this case, the plasmid is larger but the mRNAs are smaller and different promoters can be used. A recent study suggests that a multi-transcript approach appears to give higher yields than a single-transcript polycistronic system (Kim *et al.*, 2004).

It must be noted that all these techniques (*i.e.* several vectors, several genes on a single promoter and several promoters on a single plasmid) are compatible and therefore can be combined to increase the number of proteins co-expressed within a single *E. coli* cell.

2.2. Co-expression of protein complexes in insect cells

Expression and co-expression of proteins in baculovirus-infected insect cells has been in use for some time (for a recent review, see Kost *et al.*, 2005). Initially, most experiments were performed with multiple viruses, each one bearing a single gene to be overexpressed. Since no selection can be performed on these viruses, the major problem of this approach is that it gives partial as well as full co-infection and often very careful



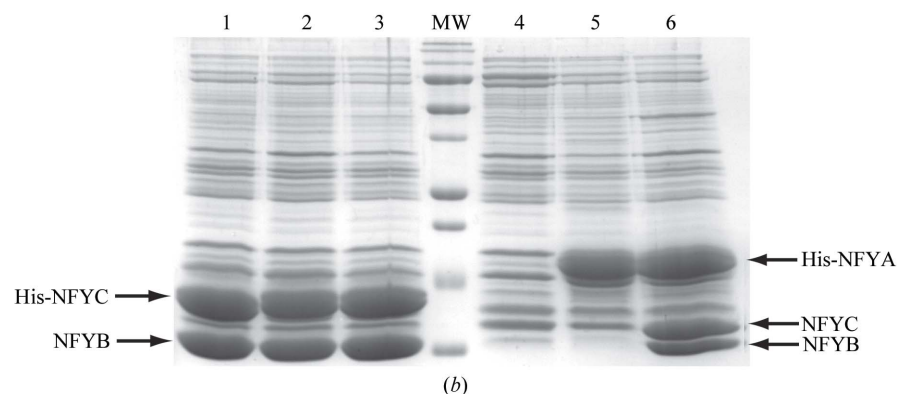
quantification of the virus titre is required to obtain reasonable results. However, large complexes have been obtained and purified using this technique, albeit in small amounts (Tirode *et al.*, 1999). The use of a single vector with multiple promoters, in a fashion similar to *E. coli* (§2.1.3), clearly increases the yields and has also been used in a number of cases (Berger *et al.*, 2004).

2.3. A novel co-expression strategy for *E. coli*

The Strasbourg laboratory has previously developed a co-expression system (Fribourg *et al.*, 2001) based on two vectors [pET15b (Novagen) and pACYC11b, having different resistance markers and compatible origins of replication] that has proved successful in numerous cases (Gangloff *et al.*, 2001; Werten *et al.*, 2002; Romier *et al.*, 2003). The major limitation of this system was that only two genes could be co-expressed at the same time. To overcome this limitation, we developed a new set of vectors (pET-MCN; **pET multi-cloning and expression**) that allows the expression of more than two genes, but is also sufficiently flexible to test the influence of the purification tag without requiring extensive cloning. Initially, all genes of interest are cloned into both vectors; cloning into the pET15b vector results in a His-tagged version of each gene, while cloning into

Figure 1

(a) The suggested workflow for using the pET-MCN (**pET multi-cloning and expression**) system. See §2.3 for additional details. (b) Co-expression of NFYB–NFYC heterodimer and NFYA–NFYB–NFYC heterotrimer from human transcription factor NFY (all subunits were truncated to their evolutionary conserved regions). All samples represent complexes retained on cobalt beads. MW, molecular-weight markers. Lanes 1–3, co-expression of His-tagged NFYC and non-tagged NFYB using the original pET15b/pACYC11b vectors (1) and the pET-MCN vectors either from two vectors (2) or a single vector with both genes on the promoter (3). Lane 4, same as lane 3 (bi-cistronic) but with no tag encoded in front of the NFYC coding gene (no retention). Lane 5, expression of His-tagged NFYA from pET-MCN vector. Lane 6, combination of the vectors used for lanes 4 and 5 (His-tagged NFYA and bi-cistronic non-tagged NFYB/NFYC complex) to produce the full complex; all three proteins are retained on the beads.



the pACYC11b vector results in an untagged version. Performing all pairwise co-expression experiments, the possible influence of the tag in complex formation can be tested easily. To be able to express more than two genes at a time (one from each vector), we modified both the pET15b and pACYC11b vectors to enable an easy ‘cut-and-paste’ strategy. In the ‘cut’ step, a piece of the T7 promoter containing the RBS and the gene of interest (that is already cloned to this vector) is excised (cut). In the ‘paste’ step, the excised piece is ligated into the T7 promoter of another vector that already contains another gene (paste). This construction leads to a single promoter with two different genes, each one preceded by its own RBS. This mechanism is based on the compatibility of the restriction sites *SpeI*, *XbaI*, *NheI* and *AvrII*, which are not affected in each cloning step, and thus this procedure can be repeated as many times as desired, leading to multi-cistronic constructs.

Further modifications of the pET15b vector in Strasbourg allowed replacement of its N-terminal His tag by either a C-terminal His tag or an N-terminal fusion protein (thioredoxin, glutathione-S-transferase, maltose-binding protein or NusA) associated or not with an N- or a C-terminal His tag. These different combinations have been made to improve the success rate for expressing poorly soluble proteins as soluble fusion proteins that could then be stabilized by interacting

with their natural partner(s). So far, two protease sites have been used: thrombin or TEV. Restriction sites have been introduced to be able to easily replace either the fusion protein or the protease site onto the plasmids and therefore abolish possible solubility problems or degradation through cryptic protease sites. Altogether, when considering all different fusions, 34 vectors have been generated that are all compatible with the cut-and-paste cloning procedure.

A typical workflow using the pET-MCN system is shown in Fig. 1(a) and is described below.

(i) The different constructs generated from the genes of interest are cloned into both the modified pET15b and pACYC11b vectors. Different fusions could be used for constructs that seem to be poorly soluble.

(ii) At this stage each possible pair can be tested for interaction by co-expression in *E. coli*.

(iii) Constructs coding for interacting protein pairs are transferred onto a single vector following the cut-and-paste procedure described above.

(iv) These vectors are used for interaction studies against the initial pool of vectors (looking for trimeric complexes) but also against themselves (looking for tetrameric complexes). If novel interactions are found in this round, new vectors are generated and a new cycle is initiated. Since one cannot exclude the case that soluble complexes will only be obtained

upon co-expression of three or four subunits, it might be interesting to generate ‘random’ bi-cistronic vectors for step (iv) regardless of results in step (ii).

Our control experiments in Strasbourg showed that in either single expression or two-vector pairwise expression no decrease in protein yields was observed (Fig. 1b, lanes 1–2). Moreover, bi-cistronic vectors showed identical expression as with two vectors (Fig. 1b, lane 3). The presence of the pRare plasmid (encoding several *E. coli* rare tRNAs; Novagen), which can be used when all genes are cloned onto a single plasmid with different antibiotic resistance, sometimes results in an increase of the yield of soluble purified protein (data not shown).

Finally, we co-expressed the three-subunit transcription factor NFY, with His-tagged NFYA cloned into modified pET15b and the non-tagged NFYB/NFYC histone-like pair cloned as a bi-cistron into modified pACYC11b. Expression tests show proper formation of the ternary NFY (Fig. 1b, lane 6). The pET-MCN system has also been tested in other laboratories in Europe. Notably, at EMBL a trimeric complex was expressed as a combination of two

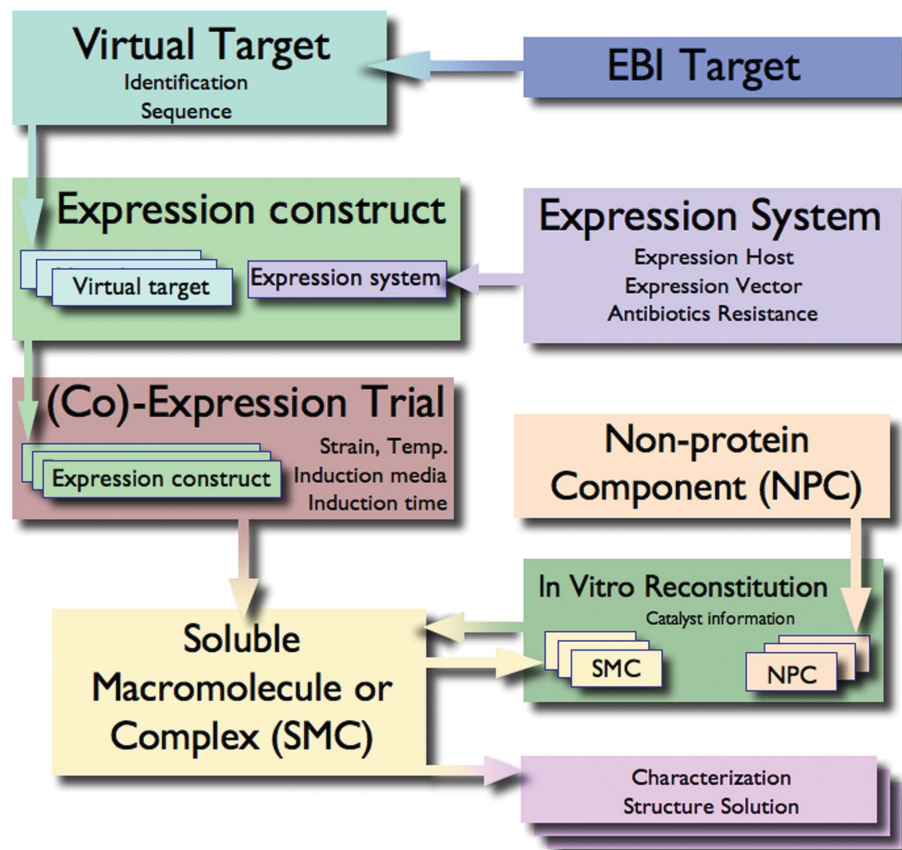


Figure 2

A graphical depiction of the main elements of the data model for the database system described in §2.1.

vectors, one containing a single gene and the second containing two genes on a single promoter. This complex was expressed, purified and its structure solved (Bono *et al.*, 2004). More recently, the genes coding for the three subunits were appended into a single promoter and the complex was expressed and purified as the one expressed from two vectors (E. Conti, personal communication). All vectors and their respective maps can be obtained from the Strasbourg laboratory upon request.

3. Results: a database system for co-expression data

The organization and archiving of data has been a central concern of the SPINE project (see Bahar *et al.*, 2006). In this approach, the central concept has been the notion of the 'target': a single gene product whose structure we attempt to solve. Although this approach is perfectly justifiable in the context of a 'classic' structural genomics approach, it was soon realised that SPINE goals extend beyond that; the study of medically relevant targets invariably involves the study of macromolecular complexes. To archive such data, the Amsterdam group have developed the Complex-3D database system presented here. A modification of this database system is in use by the EU 3D Repertoire project. It is foreseen that this experiment-tracking database will be incorporated into PIMS.

3.1. Workflow and data model

The data model is presented in Fig. 2. We opted for a specific data model rather than a more abstract one: this choice guarantees functionality and speed of development, since the graphical user interface directly refers to the data structure. However, future extendibility is compromised and this should be seen as a short-term solution that will be superseded by PIMS. Although the workflow and the model are clearly designed to enable experiments in macromolecular complexes, they can also be used for simple single-protein experiments (this functionality will not be described).

The EBI Targets are a pre-requisite for starting an experiment. As a first step, Virtual Targets can be created. Virtual Targets are related to a Target but can be a subset of it; *e.g.* a small domain of the Target or a small deletion that was designed for practical reasons. The next step in the workflow is to create an Expression Construct. An Expression Construct consists of one or more proteins (Virtual Targets) and an Expression System (*E. coli*, insect cells *etc.*) that this Expression Construct was designed for. At this point, the user can create a multi-cistronic plasmid that expresses more than one protein (by including many Virtual Targets), as discussed in §§2.1.2, 2.1.3 and 2.2. At the same time, the user can add affinity tags and protease-cleavage sites to each component of the expression construct. It should be noted that an Expression Construct is designed for a certain Expression System; a variety of Expression Systems are pre-registered in the database, while this list can easily be appended as new systems emerge. All commonly used vectors and viruses can be

described in a flexible manner. Given an Expression Construct the user can perform Expression Trials, as in the real laboratory, by varying the temperature, growth and induction conditions, the expression host *etc.* During an Expression Trial, more than one Expression Construct can be combined, thus allowing the efficient description of the experiments presented in §§2.1.1, 2.2 and 2.3. When an Expression Trial yields a Soluble Macromolecular Complex (or single macromolecule) this molecule becomes available for forming higher order complexes with other macromolecules and/or non-protein components (In Vitro Reconstitution) and for performing Experiments. The description of the latter section of the information-management system developed by the Amsterdam group is beyond the scope of this paper.

3.2. Implementation

The MySQL relational database server (<http://www.mysql.com/>) was used to provide the database support for the Complex-3D database back-end and WebObjects, a Java-based technology, was used to implement a stable and maintainable web interface to the database. The application is deployed to servers running Linux or MacOSX. A Linux server providing http bridges the application servers to the internet.

3.3. Sharing information and privacy of sensitive data

In a project with many participants, the need for privacy is often in conflict with the obligation to share information. Any group/participant can participate in one or more projects. For example, the NKI (the Amsterdam group) participates in SPINE but is also involved in 'NKI-in-house projects'. When registering an experiment in the database, a user has to also clarify to which project this experiment belongs; for example, a user from NKI would have to clarify if the registration is for SPINE or for the NKI in-house project. The user could thus register experiment *A* as a SPINE experiment and experiment *B* as an NKI in-house experiment. After both experiments are registered, all users from NKI have access to the full details of both experiments *A* and *B*. Users that are outside the NKI, however, can only see experiment *A* and that only if they are registered participants of the SPINE project; these users will have no access to experiment *B*, which contains NKI in-house data. Moreover, certain information that is highly sensitive, for example the exact sequence of the Virtual Targets or the exact conditions of Expression Trials, are only visible to the members of the group: in our example, all NKI users would be able to see all details of experiments *A* and *B*; other SPINE participating users would see that experiment *A* exists and is connected to certain Targets, but would not obtain details such as the exact sequence of each Virtual Target that is part of this experiment. Should anyone need more data on a particular experiment in order to reproduce it, the user that entered that experiment can be contacted directly from within the graphical user interface of the database.

Table 1

Co-expression in *E. coli* and insect cells.

The symbols denote: –, system not tested for this project; X, no expression of soluble proteins or the intact soluble complex; +, expression of marginally soluble but unstable proteins or complex in amounts visible in Coomassie-stained SDS gel that behaved poorly in purification; ++, expression of small quantities of soluble well behaved proteins or complex; +++, crystallization amounts of good quality soluble well behaved proteins or complex.

Complex name	Section	<i>E. coli</i>				Insect cells		
		Single proteins	Co-expression			Single proteins	Co-expression	
			Multiple vectors	Single vector, single RNA	Single vector, multiple RNAs		Multiple viruses	Single virus
TAF _{II} 6/TAF _{II} 9	4.1.1	X/X	+++†	–	–	–	–	–
hTAF _{II} 4 _{II} b/hTAF _{II} 12	4.1.2	X/X	–	–	+++	–	–	–
ER/SRC-1	4.2.1	+ / ++	–	++	–	–	–	–
VDR/RXR/Drip	4.2.2	X / +++ / +	+++‡	–	–	–	–	–
VirE1/VirE2	4.3	+ / X	–	–	+++	–	–	–
Cdt1/geminin	4.4	+ / ++	+++	++	X	+ / ++	–	–
Ring1b/Bmi1	4.5	+ / +	++	+++	–	–	–	–
HR6B/Rad18	4.6	+++ / X	+++	+++	–	++ / +	++	–
MSH2/MSH6	4.7	+ / X	X	–	–	+ / +	++	++
Skp1/Fbox/Rbx1/Cul1	4.8	++ / X / X / X	–	X	–	– / – / + / X	+++§	+++§
Cdk7/cyclin H/MAT1	4.9	X / +++ / X	X	X	–	+ / +++ / +	++	++

† Each vector had a different ORI. ‡ Two of the three vectors with the same ORI. § Two of the four genes were in one virus, while the other two were in separate viruses.

All information is of course available at all times to the database manager and data mining (querying) for reporting and scientific reasons is straightforward.

4. Results: case studies

The following examples are drawn from different projects that were or are pursued in SPINE laboratories. They provide different snapshots on how co-expression can be carried out to obtain soluble complexes, mostly for the purpose of crystallization. These examples range from ‘routine’ benchmarks in *E. coli* co-expression to more complicated multi-expression in insect cells (see Table 1 for an overview). Many of the case studies include comparisons of the different systems described in §2.1 and §2.2. Together, they constitute an informative, albeit incomplete, picture of which methods are especially useful for trying to co-express proteins to obtain soluble complexes in a single step. Here, we present a case-by-case study aiming to analyse the particularities of each case and provide clues of experimental protocols that should be tried in similar cases.

4.1. TFIID subcomplexes

The general transcription factor TFIID is a multi-protein complex involved in transcription initiation by RNA polymerase II. It is composed of 15 subunits: the TATA-box binding protein (TBP) and 14 TBP-associated factors (TAF_{II}s).

4.1.1. TAF_{II}6/TAF_{II}9. Two of the TAF_{II}s, TAF_{II}6 and TAF_{II}9, show sequence similarity with histones H4 and H3, respectively. The structure of the heterodimer from *Drosophila* has already been solved, confirming that these two proteins interact through histone motifs (Xie *et al.*, 1996). In that study, both partners were expressed independently as insoluble GST

fusions in *E. coli*, purified from inclusion bodies and the heterodimer reconstituted by refolding.

The Strasbourg group have tried to obtain the yeast and human TFIID sub-complex pairs by co-expressing their subunits in *E. coli* BL21 (DE3) using a two-plasmid strategy: Both plasmids, pET15b (Novagen) and pACYC11b (Fribourg *et al.*, 2001), have different resistance markers and different origins of replication, the pACYC11b encoding no tag. Fig. 3(a) shows that upon single expression of the His-tagged proteins, no soluble product is obtained at the expected molecular weight (lanes 1/2 and 6/7). In contrast, upon co-expression of both proteins, one tagged and the other untagged, soluble complexes are obtained, the untagged partner being retained on the affinity column through its interaction with the tagged protein (lanes 3/4 and 8/9). The position of the tag is not important for formation of the yeast complex (lanes 3/4), but in the human case His-tagging of TAF_{II}9, but not of TAF_{II}6, strongly reduces complex formation (compare lanes 8 and 9).

4.1.2. hTAF_{II}4_{II}b/hTAF_{II}12. hTAF_{II}4_{II}b and hTAF_{II}12 are two other TFIID subunits containing histone motifs (Gangloff *et al.*, 2001; Werten *et al.*, 2002). The complex between hTAF_{II}12 and hTAF_{II}4_{II}b binds specifically to DNA (Shao *et al.*, 2005). Both proteins were found in inclusion bodies when expressed separately by the Weizmann group. After cloning of both genes into a pACYCDuet expression vector (Novagen), which provides two promoters in a single vector (see §2.1.3), co-expression yielded a soluble complex (Fig. 3b) that could be purified through the His tag encoded at the N-terminus of hTAF_{II}12.

4.2. Nuclear receptor complexes

The superfamily of nuclear receptors (NRs) is a large class of metazoan transcriptional regulators that control most aspects of mammalian physiology from development to

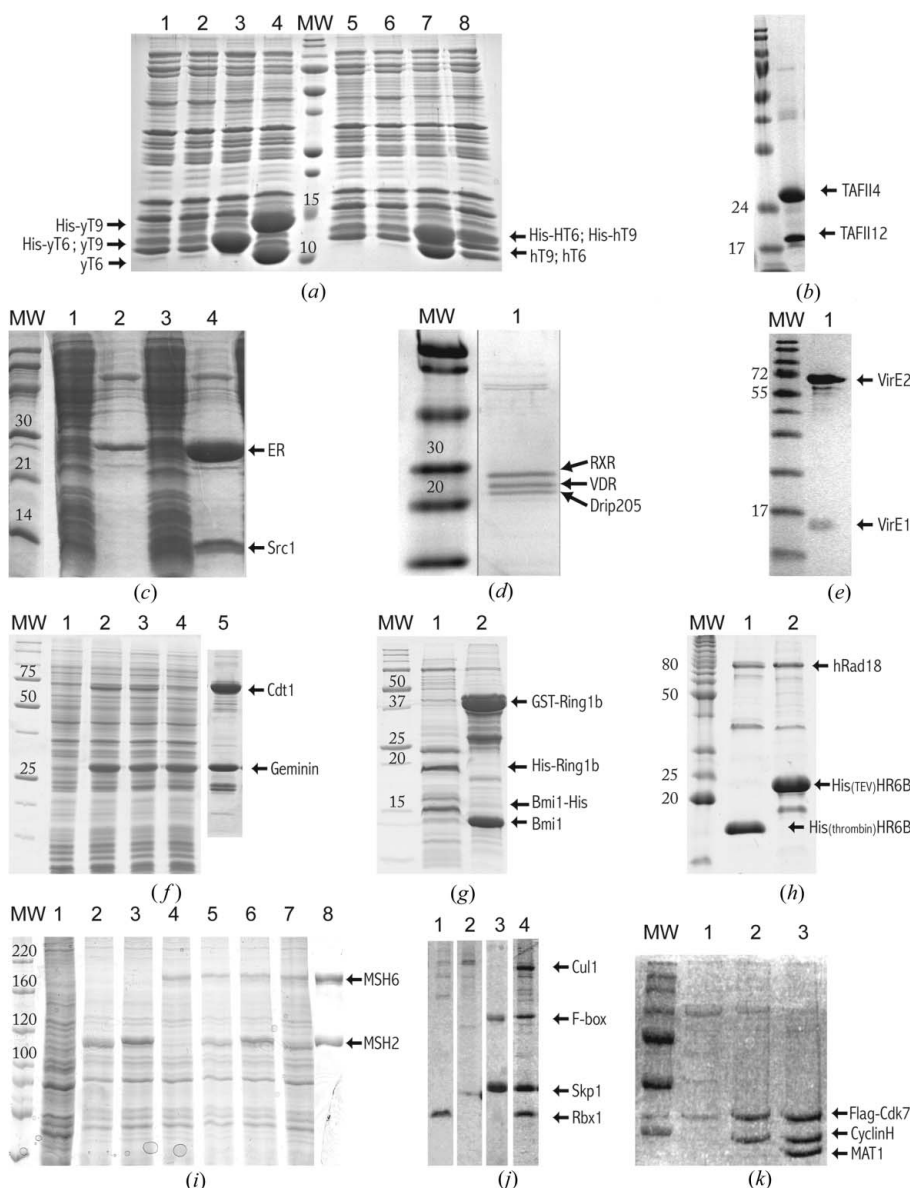


Figure 3

Coomassie blue-stained polyacrylamide gels of various stages of protein production referring to the case studies. In all gels MW is molecular-weight markers and the size of relevant bands is indicated with a number at the side or above each band. A label indicating each protein of interest is at the side of each gel, with an arrow pointing to the corresponding height of each protein band. (a) Expression and co-expression of yeast (y) and human (h) TAF_{II}6 (T6) and TAF_{II}9 (T9). All samples represent proteins retained on cobalt beads. Lanes 1, 2, 5 and 6, single expression of His-tagged yTAF_{II}6, yTAF_{II}9, hTAF_{II}6 and hTAF_{II}9, respectively. Lanes 3, 4, 7 and 8, co-expression of His-yTAF_{II}6/yTAF_{II}9, His-yTAF_{II}9/yTAF_{II}6, His-hTAF_{II}6/hTAF_{II}9 and His-hTAF_{II}9/hTAF_{II}6, respectively. (b) Co-expression of the complex formed between hTAF_{II}4 and hTAF_{II}12. Lane 1 shows the purified complex. (c) Expression of His-ER-LBD and co-expression of His-ER-LBD with SRC-1 fragment. Lane 1, soluble fraction of His-ER-LBD expression; lane 2, His-ER-LBD after affinity purification on cobalt beads; lane 3, soluble fraction of His-ER-LBD and SRC-1 co-expression; lane 4, His-ER-LBD–SRC-1 complex after affinity purification on cobalt beads through the His-tag on ER-LBD. (d) Co-expression of the complex formed between human VDR, RXR and a fragment of Drip205. Lane 1 shows the purified complex. Note that MW and lane 1 are both from the same gel, but at opposing sides and are depicted together. (e) Co-expression of the complex formed between VirE1 and VirE2. Lane 1 shows the purified complex. (f) Co-expression trials of human Cdt1 and geminin. Lane 1, cells before induction. Lane 2, co-expression from two plasmids. Lane 3, co-expression from one plasmid, producing one transcript. Lane 4, co-expression from pETDuet. All lanes present total cell extract. Lane 5 (pasted from a different gel), the complex after affinity purification through the His tag on Cdt1. (g) Co-expression trials of mouse Ring1b and Bmi1 Ring domain fragments. Lane 1, proteins retained in Ni²⁺ beads after co-expression of His-tagged Ring domains. Lane 2, proteins retained in glutathione beads after co-expression of GST-tagged Ring1b and untagged Bmi1. (h) Co-expression trials of human Rad18 and HR6B. Lanes 1 and 2, proteins retained in Ni⁺ beads after co-expression of His-tagged constructs, using one plasmid (lane 1) or two plasmids (lane 2) for the co-expression experiment. (i) Co-expression of human MSH2 and MSH6 in baculovirus-infected Sf9 insect cells. Lane 1, uninfected cells; lane 2, cells infected with virus for MSH2 72 h post-infection; lane 3, cells infected with virus for His-tagged MSH2; lane 4, cells infected with virus for MSH6; lane 5, co-expression of MSH2 and MSH6 from separate viruses; lane 6, co-expression from separate viruses coding for His6-MSH2 and MSH6; lane 7, co-expression of MSH2 and MSH6 from a single baculovirus; lane 8 (pasted from a different gel), purified His-tagged MutS α . (j) Co-expression of the SCF components in insect cells followed by Ni²⁺-bead affinity purification. Lanes 1–3, using separate viruses for Rbx1 (1), Cul1 (2) and Skp1-F-box (3); lane 4, after co-transfection and optimization using all three viruses. (k) Co-expression of CTD-activating kinase (CAK) subunits co-expressed in insect cells. Lane 1, purified Flag-Cdk7; lane 2, Flag-Cdk7/cyclin H; lane 3, Flag-Cdk7/cyclin H/MAT1.

homeostasis and constitute an important target for pharmaceutical action. They specifically bind small hydrophobic molecules that regulate their transcriptional activity (Renaud & Moras, 2000). These ligands constitute regulatory signals which modify the NR transcriptional activity in a three-step mechanism: repression, de-repression and transcription activation. Most NRs are physiologically functional in a dimeric form, either as homodimer like the oestrogen nuclear receptor (ER) or as heterodimer like the vitamin D nuclear receptor (VDR). Transcriptional activity relies on the coordinated action of the NRs and a variety of co-activator complexes.

4.2.1. ER/SRC-1. ER is responsible for mediating all of the physiological and pharmacological effects of natural and synthetic oestrogens and anti-oestrogens. The Amsterdam group studied the complex between ER and one of its co-activators SRC-1 (steroid receptor coactivator-1). The ligand-binding domain of the ER gene was initially cloned alone into a pET25b vector, in frame with an N-terminal His-tag coding sequence. A bi-cistronic expression vector of the ER domain and an 82-amino-acid SRC-1 construct was constructed by PCR concatenation of the two genes, each preceded by a linker and ribosome-binding site (RBS) (*TGATAATCTA-GAAATTTGTTTAACTTTAAGAAGGAGATGGATCC-ATG*, the linker sequence being in italics, the RBS in bold, the *Bam*HI cloning site underlined and the stop codon of the ER gene and the start codon of the SRC-1 gene in bold italics), and insertion into a pET vector. Upon expression of the ER protein alone, most of the protein was found in inclusion bodies. Upon co-expression with SRC-1, an increase of the yields of expressed and soluble ER protein was observed (Fig. 3c) and up to 10 mg of ER/SRC1 complex was obtained per litre of culture.

4.2.2. VDR/RXR/Drip. VDR mediates the action of the active form of vitamin D, 1,25(OH)₂D₃. The Strasbourg group studied the complex between a VDR/retinoid X (RXR) receptor heterodimer and a fragment of the Drip205 co-activator. The VDR construct was cloned into a pET28b vector downstream of a His-tag coding sequence. The RXR construct was cloned into the pACYC11b vector (Fribourg *et al.*, 2001) and the Drip205 construct was cloned into a pGEX-4T2 vector. In this case, the two vectors, pET28b and pGEX-4T2, have the same origin of replication, the latter having a higher copy number. Only the pACYC11b plasmid has a different origin of replication. However, all plasmids have different resistance markers allowing selection of multi-transformants. Expression of the proteins either alone or in complex was carried out in *E. coli* BL21(DE3) cells. Expressed alone, VDR was found mainly in inclusion bodies, but co-expression with untagged RXR allowed the purification of a soluble complex. GST-tagged Drip205 is soluble, but removal of the GST fusion results in insoluble protein. Co-expression of the three proteins together leads to soluble trimeric complex. Purification was performed by using sequentially the GST and His-tag fusions followed by tag removal and gel filtration. A trimeric complex could be purified to homogeneity, as seen in Fig. 3(d).

4.3. VirE1/VirE2

VirE2 and its binding partner VirE1 are two proteins involved in the transfer of DNA from the soil bacterium *Agrobacterium tumefaciens* into plant cells, causing plant transformation (Deng *et al.*, 1999; Abu-Arish *et al.*, 2004). It has been suggested that VirE1 associates with VirE2 in the bacterium to prevent binding of VirE2 to ssDNA and VirE2 self-aggregation. The Weizmann group found that VirE2 alone is expressed in inclusion bodies and attempts to refold the denatured inclusion bodies of VirE2 yielded soluble aggregates. When VirE2 was co-expressed with VirE1 in pACYC-Duet, a soluble complex of the two proteins was obtained (Fig. 3e) which was purified through an N-terminal His-tag on the VirE1 protein. The complex was functional as shown by ssDNA-binding activity (data not shown).

4.4. Cdt1/geminin

Replication of DNA within eukaryotic cells requires the formation of pre-replication complexes (Pre-RCs, replication license) that bind to chromatin and recruit MCM helicases, establishing the conditions for DNA replication. Pre-RC assembly is an ordered process in which the origin recognition complex (ORC), Cdc6 and Cdt1 sequentially bind to chromatin during the G1 phase (Bell & Dutta, 2002). As the S phase proceeds, replicated origins are prevented from becoming re-replicated by inhibiting further licensing. This is achieved by the combined activity of cyclin-dependent kinases and an inhibitory protein called geminin; geminin binds tightly to Cdt1 after initiation of DNA replication, preventing Cdt1 re-assembly onto pre-RC (Nishitani & Lygerou, 2002).

Three different co-expression strategies have been pursued by the Amsterdam group to produce the complex between human Cdt1 and geminin in *E. coli*. (i) Co-expression from two different vectors having different resistance markers but the same origin of replication: pET28b and pET22b (Novagen). (ii) Co-expression from one vector producing one transcript. In this case, part of the pET28b or the pET22b promoter (linker region and RBS) was cut and pasted downstream the first gene of the other vector (with a technique similar to that described in §2.3), leading to a single promoter having two coding sequences preceded by an RBS. All four possible combinations were produced, with either Cdt1 as upstream gene or as downstream gene, but also with the His tag coding sequence in front of either the upstream or the downstream gene. (iii) We have also used the pET-Duet (Novagen) system and thus performed co-expression from one vector producing two transcripts. Once again, both combinations with Cdt1 or geminin His-tagged were constructed.

All three expression systems were tested in parallel for co-expression in Rosetta2 (DE3) cells. All attempts to generate untagged Cdt1 failed in all expression systems used. On the contrary, when Cdt1 was tagged, expression of both proteins was observed (Fig. 3f, lanes 2–3), regardless of the order of the genes on the bi-cistronic vectors (data not shown). The highest yields were obtained with the two-vector strategy (Fig. 3f, lane 2). Cdt1 was also relatively well produced in the one vector,

one transcript strategy, but was barely detectable when expressed from the pET-Duet (Fig. 3*f*, compare lanes 3 and 4). Expressed complexes are easily purified using the His tag available only in Cdt1 (Fig. 3*f*, lane 5).

4.5. Ring1b/Bmi1

The Ring finger proteins Ring1b and Bmi1 belong to the Polycomb group proteins (PcG), which play an important role in controlling developmental patterning by maintaining transcriptional repression of specific genes. Recent evidence supports a role of Ring1b and Bmi1 in the ubiquitination pathway. These Ring finger proteins are essential for the function of the PRC1 complex, which forms an E3 ubiquitin ligase that monoubiquitinates histone H2A (Wang *et al.*, 2004).

To produce the Ring1b–Bmi1 complex, the Amsterdam group used two co-expression strategies in *E. coli*. (i) Co-expression from a modified pGEX6P vector with both genes on a single promoter, whereas Ring1b was expressed fused to GST and Bmi1 was untagged. (ii) A two-vector strategy with Ring1b into a pET28b vector and Bmi1 into a pET22b vector, leading to proteins having an N-terminal and a C-terminal His tag, respectively. In both cases, the proteins expressed well and show that both systems work equally well (Fig. 3*g*). Purification utilizing the GST-tagged Ring1b and cleavage of the GST fusion leads to soluble purified Ring1b–Bmi1 complex. No purification of the His-tagged complex was attempted, since the experiment was only performed to compare expression levels in the hope of improving protein yields.

4.6. HR6B/Rad18

The E2/E3 ubiquitin ligase complex HR6B–Rad18 targets replication processivity factor PCNA for ubiquitination (Hoegge *et al.*, 2002). Human Rad18 is a modular protein containing a RING domain, a Zn finger, a SAP domain and a predicted coiled-coil region, in addition to several predicted unfolded areas which may be sites of protein interaction or recognition, *e.g.* HR6B, Pol-eta, PCNA (Ulrich & Jentsch, 2000; Prakash *et al.*, 2005).

Rad18 alone is insoluble when expressed in bacteria, unless fused to GST; removal of the GST, however, leads to severe and irreversible aggregation. Two separate approaches were followed by the Amsterdam group to obtain soluble and active protein in complex with one of its partners, human HR6B. (i) A bi-cistronic (one promoter, one transcript) construction was made by three-point ligation with a pET25b vector, each gene having its own RBS and the downstream gene, *hr6b*, being preceded by a His-tag coding sequence. (ii) Two separate plasmids were used to express both proteins, Rad18 cloned untagged in PET22b and HR6B cloned in a PET28a expression vector in-frame with the N-terminal His-tag.

Both expression systems yielded soluble human HR6B–Rad18 complex. The single plasmid system PET25b-hRad18-His6hHR6B yields good amounts of complex (~2 mg per litre of culture), with an approximate 20-fold excess of HR6B, despite this protein being encoded as the second gene (Fig. 3*h*, lane 1). The double-plasmid system yielded slightly larger

amounts of protein complex (~5 mg per litre of culture) with a similar amount of excess HR6B (Fig. 3*h*, lane 2). In both expression systems an apparent alternative starting codon resulted in a short Rad18 fragment starting at residue Met312 as confirmed by N-terminal sequencing.

4.7. MutS α mismatch-repair complex

DNA-mismatch repair is involved in maintaining genomic stability by repairing base–base mismatches and small loops that are introduced during DNA replication (Kunkel & Erie, 2005). In eukarya, mismatches are recognized by MutS α , a heterodimer of MSH2 and MSH6, which binds the mismatch and initiates repair. Co-expression of the human MSH2 and MSH6 subunits was tried in *E. coli* using pET-derived expression vectors with different antibiotic resistance but the same origin of replication. Both genes were cloned in various pET vectors, resulting in untagged or N-terminally His-tagged and GST-tagged proteins. Several *E. coli* expression strains, growth protocols and induction procedures were tested, but in all cases soluble expression of only MSH2 was observed (Amsterdam; results not shown).

The MutS α complex was co-expressed in Sf9 insect cells infected by recombinant baculovirus. Expression of either subunit independently resulted in production of small amounts of each protein 72 h post-infection (Fig. 3*i*, lanes 2, 3 and 4). Co-expression of both subunits by simultaneous infection with separate viruses for both subunits, as well as co-expression from a single baculovirus carrying the *msh2* and *msh6* genes under the control of the p10 and the polyhedrin promoter, respectively, resulted in similar amounts of both subunits (Fig. 3*i*, lanes 5 and 7). The native (untagged) MutS α complex could be purified as a stable heterodimer from these cells at more than 95% purity. Stability of the amplified virus appears to be variable and yields are typically less than 1 mg pure protein per litre of insect-cell culture. An N-terminal His-tagged fused MSH2 increases yield for this subunit and allows purification to more than 98% purity (Fig. 3*i*, lane 8), but yields do not exceed 1 mg per litre of culture.

4.8. SCF ubiquitin ligase

Many cellular processes depend on tight regulation and high specificity of protein degradation by the ubiquitin/proteasome system. Polyubiquitination of target proteins is a three-step process involving first an activating enzyme (E1), then a conjugating enzyme (E2) and in the last step a ubiquitin ligase (E3) (Pickart, 2001). The SCF complex is one of many ubiquitin ligases and is involved in the degradation of numerous cell-cycle regulatory proteins and transcription factors (Cardozo & Pagano, 2004). This multi-subunit E3 ligase consists of four proteins, Skp1, Cul1, Rbx1 and an F-box protein, the latter component being variable and ensuring a broad substrate specificity.

Bacterial overexpression of these proteins singly as His-tag fusions has proven to be extremely difficult in the experience of the Amsterdam group. All proteins of the complex, except Skp1, gave very low or undetectable expression levels. Other

fusions (*e.g.* GST or MBP) resulted in some cases in slightly increased expression, but cleaving off these tags resulted in aggregated or precipitated proteins. Co-expression in bacteria was tried by constructing a quadri-cistronic expression vector by PCR concatenation of all four genes, each preceded by a linker and ribosome-binding site (*TGATCTAGAAATAATTTTGTTAACTTTAAGAAGGAGATCCATGG*, the linker sequence being in italics and the RBS in bold), and insertion into a pET vector. Only one protein was His-tagged. Unfortunately, expression was extremely low except for Skp1. Changing the order of the genes on the plasmid did not result in an increased level of expression.

We subsequently attempted expression of this complex in insect cells. Rbx1 and Cul1 were expressed from two different viruses, whereas Skp1 and an F-box protein were expressed from a virus coding for both proteins from two different viral promoters. Using separate viruses for protein expression, we obtained good expression for the Skp1-F-box combination ($\sim 2 \text{ mg l}^{-1}$) and reasonable expression of Rbx1 ($\sim 0.2 \text{ mg l}^{-1}$), but still very poor expression of Cul1 (Fig. 3j). Upon combination of the various viruses, we observed some Cul1 expression and after optimization of virus amounts used for infection, we were able to express all four proteins simultaneously in similar amounts (Fig. 3j).

4.9. CAK subcomplex of TFIIF

Cdk-activating kinase (CAK) is a trimeric complex consisting of cdk7, cyclin H and MAT1 and activates the cell-cycle regulating Cdk through T-loop phosphorylation. In addition, other substrates of the CAK complex have been identified when CAK is assembled with the TFIIF core proteins, thereby regulating transcription and nucleotide-excision repair (Fisher *et al.*, 1995; Tirode *et al.*, 1999).

Attempts by the Strasbourg group to produce the cdk7–cyclin H binary complex in *E. coli* were all unsuccessful, either by co-expression of bi-cistronic cdk7 and cyclin H from a single plasmid or by co-expression from two plasmids with compatible antibiotic resistance and replication origins. Various induction strategies and growth conditions were tested, but in all cases cdk7 was found in the insoluble pellet whilst His-tagged cyclin H was produced in the soluble extract.

When expressed in insect cells, cdk7 can be obtained as a soluble protein, either alone, in complex with cyclin H or in complex with both cyclin H and MAT1; this was tested on a small scale with anti-Flag antibody (see Fig. 3k for details). For production, the genes coding for a histidine-tagged cyclin H and cdk7 were cloned under the control of a P10 and polyhedron promoters. Up to 3 mg of homogenous complex could be purified from 10^9 cells.

5. Discussion

Reconstitution of protein complexes is a tedious task that may require a lot of time and effort. The various studies presented here and many others that can partially be found referenced throughout this manuscript demonstrate that co-expression of

the different subunits of a complex, either in *E. coli* or in eukaryotic cells such as insect cells, can alleviate the need for *in vitro* reconstitution and allow the complexes to form directly *in vivo*. Clearly, the major advantage of this technique is to enable the expression as soluble complex of protein partners that are improperly folded when expressed independently. As part of SPINE, we have investigated various approaches for co-expressing proteins in order to obtain sufficient amounts of multi-protein complexes that may be amenable to structural studies. In this manuscript, different cases are presented which recapitulate most of the different ways of performing co-expression experiments. Although not exhaustive, these results together with other data that were not included allow a certain number of conclusive lines of evidence to be drawn together and suggest some directions which should be explored.

The use of *E. coli* as host seems a very good initial choice for co-expression of protein complexes, as for single proteins, even if some partners are barely detectable or insoluble when expressed alone. This may come from the stabilizing effect observed upon complex formation that could prevent protein degradation or aggregation. Coupling co-expression to other techniques that are known to allow better folding of the proteins is beneficial and should be implemented before trying other expression systems: reducing growth speed by lowering temperature during expression, trying various culture media and various culture-media additives (*e.g.* sucrose), varying the induction starting time and total time, supplementing with rare codon tRNAs or chaperones and generating different constructs from the full-length protein that span different putative structural domains. Co-expression in insect cells infected by recombinant baculovirus should be exploited when expression attempts in *E. coli* fail, while co-expression in mammalian cells (see Aricescu, Lu *et al.*, 2006) should not be overlooked as an attractive alternative.

A somewhat unexpected result from our case studies comes from the fact that in *E. coli* using a multi-vector strategy with plasmids having an identical origin of replication does not seem to have a strong influence on complex formation or stoichiometry when compared to bi-cistronic vectors: if anything, some examples show better behaviour upon using multi-vectors compared with one-vector strategies. However, our study is inconclusive whether identical or different origins of replication affect yield or complex stoichiometry.

Given that multi-vector strategies and single-vector strategies (with single or multiple promoters) on average perform equally well, the attractive possibility of combining both strategies and achieving expression of multiple component complexes, where a group of components resides in one vector whereas another group resides in another plasmid, is open. In specific cases, smart combinatorial approaches can dramatically decrease cloning effort. The system we present in §3 capitalizes on this principle and provides attractive possibilities.

Request for materials described in §2.3 should be addressed to CR. Requests for materials relevant to the case studies

should be addressed to JLS (§§4.1.2, 4.3), DM (§§4.1.1, 4.2.2, 4.9), TKS (§§4.2.1, 4.5, 4.6, 4.7, 4.8) and AP (§4.4). Requests for the database should be addressed to AP. The research presented in this project in general, and GB and SVG in particular, have been supported by the European Commission as SPINE, Structural Proteomics In Europe, Contract No. QL2-CT-2002-00988 under the Integrated Programme 'Quality of Life and Management of Living Resources'. MBJ is supported by the European Union grant 3D Repertoire, contract No. LSHG-CT-2005-512028. VD is an EMBO long-term fellowship recipient. JHL is a VENI/NWO grant recipient. PC was employed under KWF grant No. NKI 199-2052.

References

- Abu-Arish, A., Frenkiel-Krispin, D., Fricke, T., Tzifira, T., Citovsky, V., Wolf, S. G. & Elbaum, M. (2004). *J. Biol. Chem.* **279**, 25359–25363.
- Alexandrov, A., Vignali, M., LaCount, D. J., Quartley, E., de Vries, C., De Rosa, D., Babulski, J., Mitchell, S. F., Schoenfeld, L. W., Fields, S., Hol, W. G., Dumont, M. E., Phizicky, E. M. & Grayhack, E. J. (2004). *Mol. Cell. Proteomics*, **3**, 934–938.
- Aricescu, A. R., Assenberg, R. *et al.* (2006). *Acta Cryst.* **D62**, 1114–1124.
- Aricescu, A. R., Lu, W. & Jones, E. Y. (2006). *Acta Cryst.* **D62**, 1243–1250.
- Bahar, M. *et al.* (2006). *Acta Cryst.* **D62**, 1170–1183.
- Bell, S. P. & Dutta, A. (2002). *Annu. Rev. Biochem.* **71**, 333–374.
- Berger, I., Fitzegerald, D. J. & Richmond, T. J. (2004). *Nature Biotechnol.* **22**, 1583–1587.
- Bono, F., Ebert, J., Unterholzner, L., Guttler, T., Izaurralde, E. & Conti, E. (2004). *EMBO Rep.* **5**, 304–310.
- Buhler, C., Lebbink, J. H., Bocs, C., Ladenstein, R. & Forterre, P. (2001). *J. Biol. Chem.* **276**, 37215–37222.
- Burgers, P. M. (1999). *Methods*, **18**, 349–355.
- Cardozo, T. & Pagano, M. (2004). *Nature Rev. Mol. Cell Biol.* **5**, 739–751.
- Copeland, W. C. (1997). *Protein Expr. Purif.* **9**, 1–9.
- Cramer, P., Bushnell, D. A., Fu, J., Gnat, A. L., Maier-Davis, B., Thompson, N. E., Burgess, R. R., Edwards, A. M., David, P. R. & Kornberg, R. D. (2000). *Science*, **288**, 640–649.
- Deng, W., Chen, L., Peng, W. T., Liang, X., Sekiguchi, S., Gordon, M. P., Comai, L. & Nester, E. W. (1999). *Mol. Microbiol.* **31**, 1795–1807.
- Dziembowski, A. & Séraphin, B. (2004). *FEBS Lett.* **556**, 1–6.
- Fisher, R. P., Jin, P., Chamberlin, H. M. & Morgan, D. O. (1995). *Cell*, **83**, 47–57.
- Fribourg, S., Romier, C., Werten, S., Gangloff, Y. G., Poterszman, A. & Moras, D. (2001). *J. Mol. Biol.* **306**, 363–373.
- Gangloff, Y. G., Romier, C., Thuault, S., Werten, S. & Davidson, I. (2001). *Trends Biochem. Sci.* **26**, 250–257.
- Gavin, A. C. & Superti-Furga, G. (2003). *Curr. Opin. Chem. Biol.* **7**, 21–27.
- Geerloff, A. *et al.* (2006). *Acta Cryst.* **D62**, 1125–1136.
- Hoegge, C., Pfander, B., Moldovan, G. L., Pyrowolakis, G. & Jentsch, S. (2002). *Nature (London)*, **419**, 135–141.
- Jawhari, A., Uhring, M., Crucifix, C., Fribourg, S., Schultz, P., Poterszman, A., Egly, J. M. & Moras, D. (2002). *Protein Expr. Purif.* **24**, 513–523.
- Johnston, K., Clements, A., Venkataramani, R. N., Trievel, R. C. & Marmorstein, R. (2000). *Protein Expr. Purif.* **20**, 435–443.
- Kim, K.-J., Kim, H.-E., Lee, K.-H., Han, W., Yi, M.-J., Jeong, J. & Oh, B.-H. (2004). *Protein Sci.* **13**, 1698–1703.
- Kost, T. A., Condreay, J. P. & Jarvis, D. L. (2005). *Nature Biotechnol.* **23**, 567–575.
- Kunkel, T. A. & Erie, D. A. (2005). *Annu. Rev. Biochem.* **74**, 681–710.
- Li, C., Schwabe, J. W. R., Banayo, E. & Evans, R. M. (1997). *Proc. Natl Acad. Sci. USA*, **94**, 2278–2283.
- Loomis, K., Sternard, H., Rupp, S., Held, D., Yaeger, K., Novy, R. & Wong, S. (2003). *Novagen InNovations Newsl.* **18**, 7–12.
- Neumann, D., Woods, A., Carling, D., Wallimann, T. & Schlattner, U. (2003). *Protein Expr. Purif.* **30**, 230–237.
- Nishitani, H. & Lygerou, Z. (2002). *Genes Cells*, **7**, 523–534.
- Novy, R., Yaeger, K., Held, D. & Mierendorf, R. (2002). *Novagen InNovations Newsl.* **15**, 2–6.
- Pickart, C. M. (2001). *Annu. Rev. Biochem.* **70**, 503–533.
- Prakash, S., Johnson, R. E. & Prakash, L. (2005). *Annu. Rev. Biochem.*, **74**, 317–353.
- Prilusky, J., Oueillet, E., Ulryck, N., Pajon, A., Bernauer, J., Krimm, I., Quevillon-Cheruel, S., Leulliot, N., Graille, M., Liger, D., Trésaugues, L., Sussman, J. L., Janin, J., van Tilbeurgh, H. & Poupon, A. (2005). *Acta Cryst.* **D61**, 671–678.
- Renaud, J. P. & Moras, D. (2000). *Cell. Mol. Life Sci.* **57**, 1748–1769.
- Romier, C., Cocchiarella, F., Mantovani, R. & Moras, D. (2003). *J. Biol. Chem.* **278**, 1336–1345.
- Shao, H., Revach, M., Moshonov, S., Tzuman, Y., Gazit, K., Albeck, S., Unger, T. & Dikstein, R. (2005). *Mol. Cell Biol.*, **25**, 206–219.
- Stebbins, C. E., Kaelin, W. G. Jr & Pavlevitch, N. P. (1999). *Science*, **284**, 455–461.
- Studier, F. W. (2005). *Protein Expr. Purif.* **41**, 207–234.
- Tan, S., Kern, R. C. & Selleck, W. (2005). *Protein Expr. Purif.* **40**, 385–395.
- Tirode, F., Busso, D., Coin, F. & Egly, J. M. (1999). *Mol. Cell*, **3**, 87–95.
- Ulrich, H. D. & Jentsch, S. (2000). *EMBO J.* **19**, 3388–3397.
- Wang, H., Wang, L., Erdjument-Bromage, H., Vidal, M., Tempst, P., Jones, R. S. & Zhang, Y. (2004). *Nature (London)*, **431**, 873–878.
- Werten, S., Mitschler, A., Romier, C., Gangloff, Y. G., Thuault, S., Davidson, I. & Moras, D. (2002). *J. Biol. Chem.* **277**, 45502–45509.
- Xie, X., Kokubo, T., Cohen, S. L., Mirza, U. A., Hoffmann, A., Chait, B. T., Roeder, R. G., Nakatani, Y. & Burley, S. K. (1996). *Nature (London)*, **380**, 316–322.
- Yusupov, M. M., Yusupova, G. Z., Baucom, A., Lieberman, K., Earnest, T. N., Cate, J. H. D. & Noller, H. F. (2001). *Science*, **292**, 883–896.