Published in final edited form as:

J Struct Funct Genomics. 2015 December ; 16(3-4): 113–128. doi:10.1007/s10969-015-9200-y.

Gene selection and cloning approaches for co-expression and production of recombinant protein–protein complexes

György Babnigg^{#1}, Robert Jedrzejczak^{#1}, Boguslaw Nocek¹, Adam Stein¹, William Eschenfeldt¹, Lucy Stols¹, Norman Marshall¹, Alicia Weger¹, Ruiying Wu¹, Mark Donnelly¹, Andrzej Joachimiak¹

¹Midwest Center for Structural Genomics, Biosciences Division, Argonne National Laboratory, 9700 S Cass Ave., Argonne, IL 60439, USA

[#] These authors contributed equally to this work.

Abstract

Multiprotein complexes play essential roles in all cells and X-ray crystallography can provide unparalleled insight into their structure and function. Many of these complexes are believed to be sufficiently stable for structural biology studies, but the production of protein–protein complexes using recombinant technologies is still labor-intensive. We have explored several strategies for the identification and cloning of heterodimers and heterotrimers that are compatible with the high-throughput (HTP) structural biology pipeline developed for single proteins. Two approaches are presented and compared which resulted in co-expression of paired genes from a single expression vector. Native operons encoding predicted interacting proteins were selected from a repertoire of genomes, and cloned directly to expression vector. In an alternative approach, *Helicobacter pylori* proteins predicted to interact strongly were cloned, each associated with translational control elements, then linked into an artificial operon. Proteins were then expressed and purified by standard HTP protocols, resulting to date in the structure determination of two *H. pylori* complexes.

Keywords

Multiprotein complexes; High-throughput structural biology; Ligation independent cloning; Multigene expression

Introduction

Many proteins in cellular systems are organized in multiprotein complexes and assemblies as they perform various functions. In these complexes proteins are held together by non-

György Babnigg, gbabnigg@anl.gov; Andrzej Joachimiak, andrzejj@anl.gov.

Author contributions GB, R.J., M.D. and A.J. designed the experiments. R.J. and W.E. performed the cloning and small scale expression of the ‘Eps-RBS-fusion-strategy’ and ‘Operon-strategy’ targets. L.S., N.M., A.W. and R.W. purified and crystallized complexes, B.N and A.S. determined the structure of the protein complexes. All authors read and revised the manuscript.

Electronic supplementary material

The online version of this article (doi:10.1007/s10969-015-9200-y) contains supplementary material, which is available to authorized users.

covalent protein–protein interactions with a broad range of affinities that are tailored to the specific processes. There are many well-documented examples of such interactions (ribosomal proteins, ATP synthase subunits, transporters and regulatory proteins [7]). Some classic examples include enzymes of the trpLEGDCEBA operon of *Escherichia coli* (the prototypical tryptophan synthase, a stable $\alpha_2\beta_2$ complex constituting TrpA and TrpB1 and anthranilate synthase composed of TrpE and TrpG [41]). Many other examples of enzymes participating in a given enzymatic pathway have been described [47]. Similarly the association of proteins involved in the stress response and protein folding such as Hsp10 and Hsp60 [37] and Hsp40 and Hsp70 [69] are well documented. The formation of complexes is the core in regulation of gene expression [36] and protein translation [27]. The important issue is to recognize that although many of these complexes are high-affinity and stable, significant fractions are transient, metastable [54, 68] or conditional [37], and the proteins are associated only during a particular step or function. Some assemblies may have a stable core and additional components may bind and dissociate during the process cycle. The ribosome [67], the RNA degradosome [28], and the protein translocation machinery [17] are classic examples of such associations. Moreover, although some complexes can form simply by association of the subunits, many require protein co-expression (and sometimes specific chaperones [18]) to assemble a functional complex [65]. It has been shown that protein interactions can be highly dynamic and components can be shared during biochemical transformation, especially in small genomes (e.g. *Mycoplasma pneumoniae* [35]) and that the transcriptional regulation of this dynamic repertoire is more complex than previously thought [22].

Past and current structural genomics efforts focus mainly on the structural studies of individual proteins or homo-oligomeric complexes utilizing well-established HTP structural biology pipelines. Obtaining structures of hetero-oligomeric protein complexes clearly is a challenge, as <13 % of PDB deposits contain two or more different polypeptides (<7 % when sequences are clustered using a 95 % sequence identity cut-off value). This is in striking contrast with estimates that the majority of proteins in the cell are involved in protein–protein interactions with various affinities and half-lives. Typically, multiprotein complexes for structural studies are obtained through direct purification from crude extract of the natural host cell, which impacts the quality and quantity of the sample and makes low abundance complexes practically inaccessible. In order to apply recombinant technologies developed for individual proteins, there are a number of issues that must be addressed. Identification of stable complexes suitable for structural studies is a challenge. HTP cloning and co-expression that addresses stoichiometry, high expression level, efficient purification and crystallization of stable and soluble protein–protein complexes (PP-CX) is also problematic. While commercial kits are available for the co-expression of proteins, the procedures are not suitable for HTP approaches. One can reconstitute a protein–protein complex using in vitro methods when the individual protein components express at high levels and are soluble by mixing together the bacterial cells, lysates, or purified proteins and isolating the complex by size exclusion chromatography (SEC) or affinity chromatography [40]. Alternatively one of the components can be purified by affinity chromatography, followed by incubation of the resin containing the bound protein with a lysate(s) of the other protein(s) in order to capture additional components of the complex [39]. When one of the

components of a protein–protein complex is not soluble in vivo approaches are used via the co-expression of the components by insertion of multiple genes in the same vector, or co-transformation with compatible plasmids with different selection markers and replication origins, each carrying a single gene. Both approaches have certain advantages and disadvantages. Several bi- or multi-cistronic vectors for co-expression of proteins have been described [66] and are commercially available (Novagen's pET-Duet vectors). Modified versions of these vectors allowing Ligation Independent Cloning (LIC) produced co-expressed proteins amenable for structural characterization have been described previously [65]. These systems however require the serial assembly of the complex mostly via conventional cloning techniques (restriction digest followed by ligation). Another disadvantage is that when more than one vector is used the desired stoichiometry of the complex components may not be obtained due to plasmid copy number variation. For example, the co-expression of 8 genes requires the use of 4 separate plasmids with replicons that may vary more than tenfold in copy numbers [e.g. p15 (10–12), CloDF13 (20–40), ColE1 (50–70), RSF1030 (>100); <http://bionumbers.hms.harvard.edu>].

The cloning of an entire operon eliminates this problem when an identical upstream region for every gene is introduced in a polycistronic plasmid. The pST44 polycistronic expression system enables the expression of up to 4 translation cassettes with a translational enhancer (e) and the Shine-Dalgarno sequence (SD) preceding each coding region of the given target gene [62]. Each gene is cloned into a separate plasmid by traditional cloning technique via unique restriction endonuclease sites prior the assembly of the polycistronic cassette, making it difficult to use for the co-expression of proteins in a HTP manner. A LIC-based construction of a tandem plasmid carrying two ORFs has been described previously exploiting the compatibility of two LIC overhangs generated by compatible restriction endonucleases [1]. While this is an attractive method to combine two genes, the resulting plasmid is enormous and does not allow further addition of complex members. An automated recombineering, Acembl, employs a battery of donor and acceptor vectors for the cloning of individual members of protein complexes via Sequence and Ligation Independent Cloning (SLIC) [5]. The system enables the rapid assembly of multigene vectors by either a labor intensive in vitro Cre-recombinase-mediated fusion, a multigene assembly by homing endonucleases, or by the rapid annealing followed by SLIC [25]. This method was successfully used for the expression of protein complexes in *E. coli* [24]. A cut-and-paste technique was described by Romier and coworkers and used in a structural genomics pipeline (Structural Proteomics in Europe; SPINE) [50]. The individual subunits of a putative complex are cloned into compatible vectors and assayed for interaction. Positive interactors are assembled via cutting one partner from the donor vector and pasting into the acceptor construct recreating the acceptor site, therefore enabling the insertion of additional genes. While this strategy is powerful, it is not compatible with HTP operation due to the traditional digestions and ligations used. The above methods use cloning techniques to create larger assemblies.

An obvious alternative is simply synthesizing the entire cassette containing the individual genes with identical upstream regions. Today's technologies enable the synthesis of not only cassettes of several genes, but entire genomes using commercial gene synthesis and assembly in yeast [20]. While this is an attractive alternative to the cloning methods it does

not allow combinatorial approaches and is still much more expensive compared to the methods presented here.

In order to address the expression of a large number of protein complexes using a standard LIC protocol, we recently developed techniques amenable to HTP operation for the expression of small PP-CX composed of 2–3 different polypeptides. Such complexes represent a majority of predicted PP-CX in an average microbial cell (Fig. 1). Target selection strategies are based on bioinformatics analyses and the mining of interactome data generated by HTP experimental methods. Currently using this approach we can co-express hundreds of proteins in a cassette format for up to three genes. Individual proteins can be suitably tagged on the N- or C-terminus for rapid affinity chromatography purification using robotic workstations and a HTP protocol. This approach is inexpensive and enables a significant increase in throughput of protein complex production in a 96-well format. Here we describe two cloning approaches: (a) a direct amplification of bacterial operons and expression in *E. coli* ('Operon-strategy'), and (b) amplification of genes from the genome with simultaneous introduction of the translation enhancer sequence followed by *E. coli* ribosome binding site ('Eps-RBS-fusion-strategy' or 'ERF-strategy'). We demonstrate that our target selection scheme combined with the experimental approaches results in the successful and rapid structure determination of PP-CX.

Experimental procedures

Cloning, expression, and purification of PP-CX

The selected genes were amplified by PCR from the genomic DNA with KOD Hot Start DNA polymerase. The PCR products were purified, T4 polymerase treated [11], cloned into the pMCSG7 vector series according to the LIC procedure [16], and transformed into the *E. coli* BL21 (DE3)-Goldstrain (Stratagene, Santa Clara, CA, USA) (which harbors an extra plasmid (pMgk) encoding one rare tRNA (corresponding to rare Arg codons, AGG and AGA)). These vectors are available from the PSI-MR material repository [8]. The 'Eps-RBS-fusion-strategy' creates an internal Eps-RBS site by utilizing the forward 5'-GTTTAACTTTAAGAAGGAGATATACAT-3' and reverse 5'-TCCTTCTTAAAGTTAAACACCATTTCTA-3' primers in the first step of the PCR reaction. The expression vectors were confirmed by DNA sequencing. The PP-CX expression and solubility is assessed in a small-scale (1 mL). A single colony is picked and grown overnight in 1 mL LB media with 150 µg/mL of ampicillin and 30 µg/mL of kanamycin at 37 °C. The following morning, 120 µL of overnight culture is transferred to 48-well culture plates containing 4 mL of "pink" media. After reaching OD₆₀₀ = 1 (~3 h of incubation at 37 °C, 270 rpm) the 48-well plates are cooled down to 18 °C and IPTG is added to a final concentration of 0.5 mM. The culture is grown for an additional 16–20 h at 18 °C, 270 rpm. The 96-well plates are then spun down, the supernatant is discarded, and the pellets are resuspended in 100 µL of buffer A [lysis buffer, 20 mM imidazole, 10 mM β-mercaptoethanol (β-ME)]. The plates are flash frozen in liquid nitrogen for about 10 min followed by batch sonication. After sonication and thawing, 20 µL of lysozyme and 20 µL of benzonase (Novagen) are added to 15 mL of buffer A, then 150 µL of this solution is added to each well. Plates are then incubated at 26 °C, 200 rpm. After 90 min of incubation,

disrupted cells are transferred from 48- to 96-well plates. At this point, expression samples are collected. The 96-well plates are spun at $3200 \times g$ for 60 min. The supernatant is mixed with Ni Sepharose and applied onto a 1 μ m 96-well filter plate. Any unbound sample is washed out by 3 spinning/washing cycles of 250 μ L at $50 \times g$. In the next step, 100 μ L of elution buffer (500 mM imidazole in lysis buffer, TEV protease 0.2 mg/mL) is added to each well. The filter plate is spun down the following day and collected samples are analyzed by SDS-PAGE for the presence of the PP-CX. Only the targets with confirmed high-level expression and good solubility are moved into the large-scale fermentation.

To produce the protein, the bacterial culture was grown at 37 °C, 190 rpm in 1L of enriched M9 medium [38] until it reached $OD_{600} = 1.0$. After air-cooling at 4 °C for 60 min, methionine synthesis inhibitory amino acids (25 mg each/L of L-valine, L-isoleucine, L-leucine, L-lysine, L-threonine, L-phenylalanine) and 90 mg/L of selenomethionine (SeMet) (Medicillin, Inc., catalog number MD045004D) were added. Then the protein expression was induced by 0.5 mM isopropyl- β -D-thiogalactoside (IPTG). The cells were incubated overnight at 18 °C, harvested and resuspended in lysis buffer [500 mM NaCl, 5 % (v/v) glycerol, 50 mM HEPES pH 8.0, 20 mM imidazole, and 10 mM β -mercaptoethanol]. Cells were disrupted by lysozyme treatment (1 mg/mL) and sonication, and the insoluble cellular material was removed by centrifugation (75 min, $30,000 \times g$, 4 °C). The SeMet labeled protein was purified from other contaminating proteins using Ni-NTA affinity chromatography (IMAC I) and the AKTExpress system (GE Health Systems) with the addition of 10 mM β -mercaptoethanol in all buffers as described previously [34]. This was followed by the cleavage of the His₆-tag using recombinant His₇-tagged TEV protease (20 °C for 3 h at 1:30 TEV to protein molar ratio) and an additional step of Ni-NTA affinity chromatography (IMAC II) was performed to remove the protease, uncut protein, and affinity tag. The pure protein was concentrated using Amicon Ultra-15 centrifugal concentrators (Millipore, Bedford, MA, USA) in 20 mM HEPES pH 8.0 buffer, 250 mM NaCl, and 2 mM dithiothreitol (DTT). Protein concentrations were determined from the absorbance at 280 nm using the calculated molar absorption coefficient for the complex and Nanodrop spectrophotometer (Thermo Scientific, Wilmington, DE, USA) [21]. The concentration of protein samples used for crystallization ranged from 10 to 80 mg/mL. Individual aliquots of purified PP-CX were stored in -80 °C.

Protein crystallization

The PP-CX were screened for crystallization conditions with the help of the Mosquito liquid dispenser (TTP Labtech, Cambridge, MA, USA) using the sitting-drop vapor-diffusion technique in 96-well Crystal Quick plates (Greiner Bio-one, Monroe, NC, USA). For each condition, 0.4 μ L of protein (10–80 mg/mL) and 0.4 μ L of crystallization formulation were mixed; the mixture was equilibrated against 140 μ L of the reservoir in the well. Several commercially available crystallization screens were used including: ANL-1, ANL-2, MCSG-1–4 (Microlytic Inc. Burlington, MA, USA) at 24 and 4 °C. Crystals of 3-oxoadipate CoA-transferase (space group C2₁) grew from a solution containing 0.2 M ammonium sulfate, 25 % (w/v) PEG 3350, pH 8.5, while crystals of molibdopterin-converting factor belong to space group F222 and grew from 0.2 M sodium citrate, 20 % (w/v) PEG 3000

pH5.5. Prior to data collection, crystals of PP-CX were cryoprotected with glycerol (15–30 %) and flash-cooled directly in liquid nitrogen.

X-ray diffraction and structure determination

Single-wavelength diffraction (SAD) data were collected from the single crystals at 100 K near the selenium K-absorption edge at the wavelength 0.9794 Å up to 2.29 Å (for 3-oxoadipate CoA-transferase) and 1.90 Å (for molibdopterin-converting factor) at the 19-ID beamline of the Structural Biology Center at the Advanced Photon Source at Argonne National Laboratory using the program SBC-collect [29] on ADSC Quantum 315R CCD detector. The reflection intensities were integrated and scaled with the HKL3000 program suite [42]. The structures of molibdopterin-converting factor and 3-oxoadipate CoA-transferase were determined using the SAD method as implemented in the HKL3000 software. Positions of heavy atoms were determined in SHELXD and initial phases were obtained from SHELXE. The heavy atom sites were refined and improved phases were calculated by iterations of MLPHARE and DM. The initial protein model was built in ARP/wARP. Model re-building for each structure was performed manually using the program COOT [13]. The final refinement of structures was performed using the program Phenix.Refine [44]. The stereochemistry of the structure was checked with PROCHECK and the Ramachandran plot. The atomic coordinates and structure

Results

Operon selection

Target selection strategies for the identification of PP-CX can be based on bioinformatics, experimental, text mining tools, or a combination of them. For the expression of PPCX in this manuscript, targets were selected via bioinformatics approaches and/or mining of interactome databases. The MicrobesOnline [10] and SEED [3] resources were used for the identification of PP-CX with focus on predicted small stable heterodimers and heterotrimers. The analysis of operon prediction for the 1334 genomes (MicrobesOnline) was carried out by clustering them based on the number of genes. More than 10 % of operons contain just two genes (Fig. 1). A clear inverse correlation can be observed between the relative abundance of operons and the number of genes per operon (Fig. 1a). The largest operons are those encoding ATP synthase and ribosomal proteins. We also calculated the relative percentage of genes arranged in operons for all 1334 genomes (Fig. 1b). On average 60 % of genes are parts of an operon, but the relative percentage varied greatly and showed no correlation with the size of the genome (not shown).

The operon selection was performed using more than 60 available reagent genomes used at the Midwest Center for Structural Genomics (MCSG). The analysis of bicistronic (~50,000) and tricistronic operons (~30,000) identified ~1900 bicistronic and ~1700 tricistronic operons that were annotated as part of a larger assembly ('subunit') and had no trans-membrane regions as predicted by Phobius [30]. ~1100 (bicistronic) and ~460 (tricistronic) operons were conserved across many genomes. We used this latter set of operons to select 468 distinct genes residing in 288 distinct genomic regions accounting for 245 regions with bicistronic operons, and 43 regions with tricistronic operons. All selected genomic regions

were between 651 and 4369 bp long, excluding larger assemblies such as polymerases, helicases, topoisomerases, and excinucleases. The HTP gene cloning, expression, and solubility analysis of PP-CX encoded by the selected operons is described below.

We also tested the utility of target selection techniques based on genomic inference methods, namely on gene fusion. The simple sequence comparison of completed genomes can identify potential protein–protein interactions from gene fusion events. Homologs of a given gene are detected with BLAST from all sequenced genomes and fusions are detected after the elimination of fusions as a result of gene duplication. The gene fusion events were analyzed in the *H. pylori* 26695 genome. Homolog search from the bacterial reference genome set encompassing more than 8 million sequences for this relatively small genome (1600 ORFs) identified more than 25,000 sequences. After a reciprocal BLAST validation and filtering 13 sets of *H. pylori* 26695 genes were found to be fused in certain RefSeq genomes (Fig. 2). Only a single instance was found where three neighboring *H. pylori* 26,695 genes were found to be fused in other genomes. The fused gene is annotated as adenine specific DNA methyltransferase in *H. pylori* HPAG1, *H. pylori* J99, *H. pylori* Cuz20, and *H. pylori* Sat464 strains only, while their homologs appear to be single genes in most cases (Supplementary Figure S1).

Integration of proteomics data

Protein–protein interaction data are available in interactome databases and are amenable for data mining. We tested the feasibility of using interactome data along with other bioinformatics analysis to select targets for HTP structural characterization. We focused initially on *H. pylori* for which more than a dozen complexes were verified experimentally using a low-throughput two-dimensional blue native/SDSPAGE [55] and more than 1200 interactions were identified using the yeast two-hybrid (Y2H) assay [59]. We have selected 16 PP-CX from *H. pylori* based on the IrefIndex protein–protein interaction database, where the proteins are encoded by adjacent genes in the genome and are predicted to be part of an operon [2]. The expression, solubility, and structural characterization of these proteins are described below.

Cloning strategies

Two cloning approaches were used and compared for the expression of PP-CX: (a) the ‘Operon-strategy’ and (b) the ‘Eps-RBS-fusion-strategy’ (Fig. 3). The ‘Operon-strategy’ entails the simple amplification of neighboring genes as part of an operon using standard HTP LIC [12, 64]. Briefly, gene-specific LIC-tagged primers are used in the amplification of 2–3 consecutive genes enabling the N-terminal tagging of the first gene or the C-terminal tagging of the last gene (Fig. 3a). This simple method requires minor modifications in the HTP cloning pipeline, such as PCR conditions and DNA polymerase used, and can be used as the default cloning strategy of PP-CX. On average the size of intergenic region in the selected operons was 19 bases.

As an alternative we used an ‘Eps-RBS-fusion-strategy’, where an identical upstream region of the first gene is generated upstream of the second gene via a PCR extension method (Fig. 3c). Briefly, the individual genes are amplified via PCR using gene specific primers fused

with regular LIC extensions and another gene-specific primer fused with a fragment coding for a translation enhancer and a ribosomal binding site (Eps-RBS). The selection of the appropriate primers is based on the order of the selected genes in genome. In this approach both genes can be tagged either on the N- or C-termini giving rise to four possible arrangements (Fig. 3d). This method requires 4 LIC primers per clone versus 2 primers per clone for the ‘Operons-strategy’. Both strategies are compatible with HTP operations. A comparison of factors differentiating these two methods from each other and from a method used previously to clone paired proteins is presented in Table 1.

Gene fusion: case of petrobactin biosynthesis

One of the very strong indicators for PP-CX formation is when homologous members are fused in another organism. We explored the utility of this fact using the petrobactin biosynthesis operon as an example. *Bacillus anthracis* employs two distinct siderophores for iron acquisition; bacillibactin and petrobactin. The human innate immunity protein, siderocalin, can compete with bacillibactin for iron, effectively stripping off the ferric iron chelated by the 2,3-dihydrobenzoic acid (DHBA) moiety. In contrast petrobactin contains a 3,4-DHBA moiety that escapes the siderocalin system enabling iron acquisition by this pathogen. The inhibition of the petrobactin synthesis pathway therefore provides an attractive way to neutralize this virulence factor. The six enzymes responsible for petrobactin synthesis are organized into the *asb* operon (*asb*ABCDEF). Our laboratory has determined the structure of *AsbF* dehydratase which participates in the synthesis of the 3,4-DHBA precursor (Fig. 4a) [51] and the structure of siderophore synthetase (*AsbB*) [45] that contribute to the synthesis of the final product, petrobactin. The condensation of the 3,4-DHBA and spermidine is carried out by three enzymes of the operon, *AsbCDE*. Bioinformatics analysis revealed that not only the order of these genes is highly conserved but also in some organisms, such as in *Rhodopseudomonas palustris*, the *asbD* and *asbE* homologs are fused (Fig. 4b). Indeed, the *B. anthracis* N-terminal His₆-tagged *AsbD* protein forms a stable complex with the coexpressed *AsbE* protein, and the His₆-tagged *AsbC* interacts with the untagged *AsbD* (Fig. 4c) indicating a strong interaction between the components of this enzyme complex. It is important to tag the lower abundance partner in order to purify a stoichiometric complex, which was not the case for the *AsbDE* complex when *AsbD* was tagged (Fig. 4c). An alternative is to tag both partners with different purification tags and perform a two-step affinity purification to obtain the complex with desired stoichiometry. Nevertheless, the identification of fused genes in orthologs is an excellent indicator of an interaction.

Expression, solubility, purification and analysis of complexes: ‘Operon-strategy’

Operons enable the co-expression of components of cellular machines not only in the host organisms but often also heterologously in *E. coli*, enabling the biochemical and structural characterization of PP-CX in a HTP pipeline. We tested the expression of 384 selected bicistronic and tricistronic operons in a small-scale. A representative set of SDS-PAGE gels reveals that in the ‘Operon-strategy’ approach a large proportion of clones express all proteins encoded by the operons (Fig. 5a). However, there are variations observed in the expression levels of the complex members, often resulting in lower expression of the second gene. Bicistronic orthologous operons show similar stoichiometries and expression levels in

spite of sequence divergence. For example 2-oxoglutarate synthase from *Bacteroides* displays an almost identical expression pattern in spite of considerable sequence divergence. Similarly, acetolactate synthase orthologs produced in *E. coli* show high levels of expression. Several tricistronic operons expressed all three proteins but at moderate levels. The orthologous tricistronic operons also expressed similarly as shown for xanthine dehydrogenase from *Clostridium* and *Bacillus* (Fig. 5a). Highly expressed clones with the expected stoichiometry were further analyzed via a small-scale IMAC I survey using 96-well plate format and analyzed on SDS-PAGE (Fig. 5b). More than 80 soluble operons were tested and shown to be sufficiently stable to endure chromatography and remained intact during the large-scale purification. These PP-CX were purified in milligram scale using the HTP pipeline using the standard operating procedures (SOP) [34]. In addition, many clones provided PP-CX yields compatible with typical crystallization trials. There were varying successes however with the removal of the N-terminal His₆ tag from the first gene in the operon after IMAC purification resulting in protein loss and the introduction of heterogeneity in the sample (Fig. 6). The glycyl-tRNA synthetase complex from *Klebsiella pneumoniae* subsp. *pneumoniae* MGH 78578 purified at the expected stoichiometry for $\alpha_2\beta_2$ complex at the IMAC I purification step. However, the TEV treatment resulted in heterologous sample with several species that can be separated by SEC. Samples from each peak were further analyzed by native PAGE. While the glycyl-tRNA synthetase complex resulted in 3 clearly identifiable peaks, another complex, acetyl-CoA carboxylase, from the same species showed 2 peaks only (Fig. 6a). Pure PP-CX were screened for crystallization condition using commercially available formulations (ANL-1, ANL-2, and MCSG1–4) and a number of crystals was obtained (Fig. S3).

Expression, solubility, purification and analysis of target complexes: ‘Eps-RBS-fusion-strategy’

As alternative to the ‘Operon-strategy’ we tested a cloning approach where an identical Eps-RBS region is introduced in front of the second gene to aid the equal translation of components from the polycistronic mRNA. We selected targets from *H. pylori* for which not only bioinformatics analyses were available, such as operon predictions, but also interactome data derived from Y2H experiments [59]. This set comprised of thirteen proteins forming six protein complexes (FtsA/FtsZ, CheW/CheA, Rpl7/Rpl12/Rpl10, YxjD/YxjE, MoaE/MoaD and UreB/UreA) (Table 2). The urease UreAB was used as a control throughout the cloning, purification, and crystallization process as this complex was previously purified directly from bacteria, crystallized and its structure was determined [23]. None of thirteen proteins generated high yields of soluble protein when expressed individually. The cells expressing the full-length cell division protein complex (FtsA/FtsZ) had growth defects preventing the expression of the complex. The purine-binding chemotaxis protein complex (CheW/CheA) and ribosomal protein complex (Rpl7/Rpl12/Rpl10) showed poor solubility and provided no crystals in crystallization trials. The 3-oxoadipate CoA-transferase (YxjD/YxjE), molybdopterin converting factor (MoaE/MoaD) and urease (UreB/UreA) formed soluble and stable complexes that were easily purified to near homogeneity using standard protocols. All three purified complexes produced diffraction quality crystals and structures of 3-oxoadipate CoA-transferase and molybdopterin converting factor were determined (Fig. 7). The structure of recombinant

urease was not refined since the X-ray structure has been reported previously by others [63]. The structures of 3-oxoadipate CoA-transferase and molybdopterin converting factor were solved using SeMet labeled proteins and the anomalous scattering. These structures were refined to 2.29 and 1.90 Å resolutions, respectively (Table 3). The 3-oxoadipic acid CoA transferase, (also annotated as acetoacetate: succinyl-CoA CoA transferase [9]) crystallized as a tetramer made up of two $\alpha\beta$ subunit pairs, identical to the quaternary structure observed in crystals of the transferase from *B. subtilis* [65]. The unit cell for the molybdopterin converting enzyme was a simple $\alpha\beta$ dimer. The urease formed a very large complex of 12 α and 12 β subunits ($\alpha_{12}\beta_{12}$). This large structure of recombinant complex, very distinct from the much smaller ABC complex of other ureases, is consistent with that determined previously for the purified *H. pylori* enzyme [63]).

Discussion

Knowing structures of PP-CX can significantly contribute to our understanding of many cellular processes. However the complexes and assemblies may be composed of a few or up to hundreds of subunits and show broad range of stability. Therefore it is important to identify complexes that are stable in composition and show long half-life to withstand purification and crystallization. The production of protein complexes is still labor-intensive and involves: (1) the expression of the individual components separately and mixing together the purified components, or (2) the co-expression of the interactive partners in a vector capable of accepting multiple genes. At present most of these techniques are either not compatible with HTP operations or are not economical at that scale. Here we present target selection strategies and show the feasibility of producing stable bacterial PP-CX in a HTP manner using well-established and tested procedures that can yield X-ray diffraction quality crystals and lead to structures. Several bioinformatics approaches have been reported to identify interacting proteins, based on genomic inference methods (phylogenetic profile, gene cluster, Rosetta Stone, and gene neighbor methods [6, 14, 26, 48, 49, 70]), classification methods [56, 58], inclusion of text mining [6], or inclusion of proteomics data [57]. Co-regulated genes can be predicted via the automatic prediction of operon structures in prokaryotic genomes. Genes of a functional pathway are typically transcribed in a single mRNA and organized into operons [52]. Early operon prediction methods utilized evolutionarily preserved gene adjacency [15] or intergenic distance [43] for example and were trained on the *E. coli* and *B. subtilis* genomes, but with the flood of the newly sequenced genome information it was evident that operon structures can be drastically different from these model organisms. An alternative method infers operon structures by creating a training set via the combination of intergenic distance information and comparative transcriptomics studies. The predictor can be used to assign operons using sequence information only and is insensitive to variations in intergenic spacing between species [53].

The proteomics tools, such as the Y2H screens or affinity purification mass-spectrometry (AP-MS, [19]) enable the construction single species or multi-species interactome databases including stable and transient interactions [61]. Several public protein–protein interaction databases are available reporting experimental findings or 3D structures of interacting proteins. These resources list protein–protein interaction information with focus mainly on

eukaryotes (~90 %), but include information for common bacteria such as *E. coli*, *Campylobacter jejuni*, *Treponema pallidum*, *Synechocystis* sp. PCC 6803, and *H. pylori* for example. The IrefIndex [60] is an integrated resource providing a non-redundant dataset from the above publicly available databases utilizing SEGUID-based identifiers which are directly compatible with our proteomics informatics platform [4]. The v.10.0 public release of the IrefIndex was used in this study listing more than 540,000 interactions and more than 1 million interactors.

Target selection strategies employing bioinformatics, interactomics, and text mining are effective in identifying high quality complexes. The operon predictions from the MicrobesOnline resource provided the basis for the selection of many targets. The selection of orthologs from more than 60 microbial species enabled the comparative analysis. One may consider extrapolating protein–protein interaction patterns from one organism to others using well curated databases, such as EcoCyc [31]. Indeed the orthologous complexes have similar physicochemical properties: the complexes expressed at comparable levels and showed similar solubilities. For example we have successfully purified and crystallized glycyl-tRNA synthetase α_2/β_2 heterodimers from *H. ducreyi*, *K. pneumoniae*, *P. fluorescens*, *S. enterica*, *S. flexneri*, and *Y. enterocolitica*. We found that the expression and solubility of heterodimers was generally higher (>35 %) than expressing single genes, suggesting that the current SOPs are compatible with the production of small recombinant complexes, and also provides a salvage pathway for high-value targets.

We compared two HTP cloning strategies. The expression of operons is very simple and efficient. The expression of an operon in the MCSG type vectors ensures the high-level expression of the first ORF. However the other members of the operon may express at different levels driven by upstream sequences. Interestingly, *E. coli* correctly translated overlapping genes from other microbes. The translation of the first gene is governed by a strong Eps-RBS site [46] provided as part of the LIC vector, while the downstream genes on the polycistronic mRNA have their own RBS native upstream sequences with differing efficacies and unequal expression of the components can result. Changes in spacing in operons are a fine-tuning mechanism for the regulation of the expression levels of the individual components. Long spacers for example enable the tight regulation of second gene preventing unnecessary high-level production, while short spacers can lead to tight translational regulation. Nevertheless, we found that a fair number of operons expressed interacting partners near equal levels. Eps-RBS-fusion-strategy circumvents the unequal translation problem and enables the high flexibility in N- and C-terminal affinity tagging of the components. It also permits the co-expression of genes of interacting proteins with large intergenic regions on an operon or those positioned at a long distance on the chromosome and it eliminates concerns with the expression of overlapping genes.

The integration of proteomics data, namely interactome data generated by HTP methods, is valuable in target selection. In fact, the heterodimers selected from *H. pylori*, for which we were able to determine the structure, were identified based on bioinformatics analysis (operon), and supporting Y2H data. We selected a number of targets from *H. pylori* that are not part of an operon but have interactome data supporting interaction. We were able to reproduce interaction in the small-scale tests for a number of these, but none of them were

stable enough to remain intact during the large-scale purification and crystallization pipeline. It is important to note that although we were able to purify and crystallize a number of heterooligomeric complexes, the crystals often do not diffract to high resolution requiring crystallization optimization. In spite of using purified stable complexes in crystallization trials, in some cases crystals contain only one component [like the β subunit of phenylalanyl-tRNA synthetase from *Bacteroides fragilis* (PDB entry 3IG2) or the α subunit of glycyl-tRNA synthetase from *Campylobacter jejuni* (PDB entry 3RF1)].

In addition to the large-scale testing of HTP methods for the cloning, expression, and purification of protein complexes, the structures determined from the *H. pylori* set provide important biological insights since both of them are potential drug targets. 3-oxoadipate CoA-transferase is believed to play a critical role in the energy metabolism of *H. pylori* by completing the strain's tricarboxylic acid cycle [9, 32]. *H. pylori* lacks three enzymes of the canonical TCA cycle, but contains three others, including acetoacetate:succinyl-CoA transferase, which completes the cycle. Thus the transferase is likely essential for successful infection by the *H. pylori*. The 3-oxoadipate CoA-transferase subunits reside on an operon sandwiched between a 3-ketoacyl-CoA thiolase and a short chain fatty acid transporter, strengthening the postulated role of this enzyme in energy metabolism. Mammalian orthologs were identified in the brain and heart mitochondria. These organs are capable of utilizing ketone bodies as fuel, mainly in the form of acetoacetate. In rat, similar enzymatic activities are found in the gastric grandular mucosa, suggesting its role in the stimulation of acid secretion after refeeding following starvation. *H. pylori* could exploit this highly acidic environment rich in ketone bodies or small chain fatty acids. Finally, the molybdenum cofactor biosynthetic enzyme may be important for *H. pylori*'s survival as well. A known molybdopterin containing *H. pylori* enzyme, bisC (Gene ID: 8208053), is thought to be an S-oxide reductase based on homology [33], and thus may be involved in repairing oxidized amino acids or important cofactors of the strain.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to thank members of the Structural Biology Center at Argonne National Laboratory for their help with data collection at the 19-ID beamline. This work was supported by the National Institutes of Health Grants GM074942 and GM094585, and by the US Department of Energy, Office of Biological and Environmental Research, under contract DE-AC02-06CH11357.

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a US Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The US Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

References

1. Alexandrov A, Vignali M, LaCount DJ, Quartley E, de Vries C, De Rosa D, Babulski J, Mitchell SF, Schoenfeld LW, Fields S, Hol WG, Dumont ME, Phizicky EM, Grayhack EJ (2004) A facile

method for high-throughput co-expression of protein pairs. *Mol Cell Proteomics* 3(9):934–938 [PubMed: 15240823]

2. Alm EJ, Huang KH, Price MN, Koche RP, Keller K, Dubchak IL, Arkin AP (2005) The MicrobesOnline Web site for comparative genomics. *Genome Res* 15(7):1015–1022 [PubMed: 15998914]
3. Aziz RK, Devoid S, Disz T, Edwards RA, Henry CS, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Stevens RL, Von-stein V, Xia F (2012) SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. *PLoS One* 7(10):e48053 [PubMed: 23110173]
4. Babnigg G, Giometti CS (2006) A database of unique protein sequence identifiers for proteome studies. *Proteomics* 6(16):4514–4522 [PubMed: 16858731]
5. Bieniossek C, Nie Y, Frey D, Olieric N, Schaffitzel C, Collinson I, Romier C, Berger P, Richmond TJ, Steinmetz MO, Berger I (2009) Automated unrestricted multigene recombineering for multiprotein complex production. *Nat Methods* 6(6):447–450 [PubMed: 19412171]
6. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* 5(5):R35 [PubMed: 15128449]
7. Chaban Y, Boekema EJ, Dudkina NV (2014) Structures of mitochondrial oxidative phosphorylation supercomplexes and mechanisms for their stabilisation. *Biochim Biophys Acta* 1837(4):418–426 [PubMed: 24183696]
8. Cormier CY, Mohr SE, Zuo D, Hu Y, Rolfs A, Kramer J, Taycher E, Kelley F, Fiocco M, Turnbull G, LaBaer J (2010) Protein structure initiative material repository: an open shared public resource of structural genomics plasmids for the biological community. *Nucleic Acids Res* 38(Database issue):D743–D749 [PubMed: 19906724]
9. Corthesy-Theulaz IE, Bergonzelli GE, Henry H, Bachmann D, Schorderet DF, Blum AL, Ornston LN (1997) Cloning and characterization of *Helicobacter pylori* succinyl CoA:acetoacetate CoA-transferase, a novel prokaryotic member of the CoA-transferase family. *J Biol Chem* 272(41):25659–25667 [PubMed: 9325289]
10. Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D, Friedland GD, Huang KH, Keller K, Novichkov PS, Dubchak IL, Alm EJ, Arkin AP (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res* 38(Database issue):D396–D400 [PubMed: 19906701]
11. Dieckman L, Gu M, Stols L, Donnelly MI, Collart FR (2002) High throughput methods for gene cloning and expression. *Protein Expr Purif* 25(1):1–7 [PubMed: 12071692]
12. Donnelly MI, Zhou M, Millard CS, Clancy S, Stols L, Eschenfeldt WH, Collart FR, Joachimiak A (2006) An expression vector tailored for large-scale, high-throughput purification of recombinant proteins. *Protein Expr Purif* 47(2):446–454 [PubMed: 16497515]
13. Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60(Pt 12 Pt 1):2126–2132 [PubMed: 15572765]
14. Enright AJ, Iliopoulos I, Kyripides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402(6757):86–90 [PubMed: 10573422]
15. Ermolaeva MD, White O, Salzberg SL (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res* 29(5):1216–1221 [PubMed: 11222772]
16. Eschenfeldt WH, Stols L, Millard CS, Joachimiak A, Donnelly MI (2009) A family of LIC vectors for high-throughput cloning and purification of proteins. *Methods Mol Biol* 498:105–115 [PubMed: 18988021]
17. Falguieres T, Luyet PP, Gruenberg J (2009) Molecular assemblies and membrane domains in multivesicular endosome dynamics. *Exp Cell Res* 315(9):1567–1573 [PubMed: 19133258]
18. Frydman J, Hartl FU (1996) Principles of chaperone-assisted protein folding: differences between in vitro and in vivo mechanisms. *Science* 272(5267):1497–1502 [PubMed: 8633246]
19. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G (2006) Proteome

- survey reveals modularity of the yeast cell machinery. *Nature* 440(7084):631–636 [PubMed: 16429126]
20. Gibson DG, Benders GA, Andrews-Pfannkoch C, Denisova EA, Baden-Tillson H, Zaveri J, Stockwell TB, Brownley A, Thomas DW, Algire MA, Merryman C, Young L, Noskov VN, Glass JI, Venter JC, Hutchison CA 3rd, Smith HO (2008) Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* 319(5867):1215–1220 [PubMed: 18218864]
 21. Gill SC, von Hippel PH (1989) Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem* 182(2):319–326 [PubMed: 2610349]
 22. Guell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kuhner S, Rode M, Suyama M, Schmidt S, Gavin AC, Bork P, Serrano L (2009) Transcriptome complexity in a genome-reduced bacterium. *Science* 326(5957):1268–1271 [PubMed: 19965477]
 23. Ha NC, Oh ST, Sung JY, Cha KA, Lee MH, Oh BH (2001) Supramolecular assembly and acid resistance of *Helicobacter pylori* urease. *Nat Struct Biol* 8(6):505–509 [PubMed: 11373617]
 24. Haffke M, Marek M, Pelosse M, Diebold ML, Schlattner U, Berger I, Romier C (2015) Characterization and production of protein complexes by co-expression in *Escherichia coli*. *Methods Mol Biol* 1261:63–89 [PubMed: 25502194]
 25. Haffke M, Viola C, Nie Y, Berger I (2013) Tandem recombineering by SLIC cloning and Cre-LoxP fusion to generate multigene expression constructs for protein complex research. *Methods Mol Biol* 1073:131–140 [PubMed: 23996444]
 26. Huynen MA, Bork P (1998) Measuring genome evolution. *Proc Natl Acad Sci USA* 95(11):5849–5856 [PubMed: 9600883]
 27. Jackson RJ, Hellen CU, Pestova TV (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol* 11(2):113–127 [PubMed: 20094052]
 28. Janga SC, Babu MM (2009) Transcript stability in the protein interaction network of *Escherichia coli*. *Mol Biosyst* 5(2):154–162 [PubMed: 19156261]
 29. Joachimiak A (2009) High-throughput crystallography for structural genomics. *Curr Opin Struct Biol* 19(5):573–584 [PubMed: 19765976]
 30. Kall L, Krogh A, Sonnhammer EL (2007) Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res* 35(Web Server issue):W429–W432 [PubMed: 17483518]
 31. Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S (2002) The EcoCyc database. *Nucleic Acids Res* 30(1):56–58 [PubMed: 11752253]
 32. Kather B, Stingl K, van der Rest ME, Altendorf K, Molenaar D (2000) Another unusual type of citric acid cycle enzyme in *Helicobacter pylori*: the malate:quinone oxidoreductase. *J Bacteriol* 182(11):3204–3209 [PubMed: 10809701]
 33. Kawai M, Furuta Y, Yahara K, Tsuru T, Oshima K, Handa N, Takahashi N, Yoshida M, Azuma T, Hattori M, Uchiyama I, Kobayashi I (2011) Evolution in an oncogenic bacterial species with extreme genome plasticity: *Helicobacter pylori* East Asian genomes. *BMC Microbiol* 11:104 [PubMed: 21575176]
 34. Kim Y, Babnigg G, Jedrzejczak R, Eschenfeldt WH, Li H, Maltseva N, Hatzos-Skintges C, Gu M, Makowska-Grzyska M, Wu R, An H, Chhor G, Joachimiak A (2011) High-throughput protein purification and quality assessment for crystallization. *Methods* 55(1):12–28 [PubMed: 21907284]
 35. Kuhner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, Yamada T, Maier T, Bader S, Beltran-Alvarez P, Castano-Diez D, Chen WH, Devos D, Guell M, Norambuena T, Racke I, Rybin V, Schmidt A, Yus E, Aebersold R, Herrmann R, Bottcher B, Frangakis AS, Russell RB, Serrano L, Bork P, Gavin AC (2009) Proteome organization in a genome-reduced bacterium. *Science* 326(5957):1235–1240 [PubMed: 19965468]
 36. Latchman DS (1997) Transcription factors: an overview. *Int J Biochem Cell Biol* 29(12):1305–1312 [PubMed: 9570129]
 37. Li Z, Srivastava P (2004) Heat-shock proteins. *Curr Protoc Immunol* Appendix 1:Appendix 1T

38. Makowska-Grzyska M, Kim Y, Maltseva N, Li H, Zhou M, Joachimiak G, Babnigg G, Joachimiak A (2014) Protein production for structural genomics using *E. coli* expression. *Methods Mol Biol* 1140:89–105 [PubMed: 24590711]
39. Malecki M, Jedrzejczak R, Puchta O, Stepień PP, Golik P (2008) In vivo and in vitro approaches for studying the yeast mitochondrial RNA degradosome complex. *Methods Enzymol* 447:463–488 [PubMed: 19161856]
40. Malecki M, Jedrzejczak R, Stepień PP, Golik P (2007) In vitro reconstitution and characterization of the yeast mitochondrial degradosome complex unravels tight functional interdependence. *J Mol Biol* 372(1):23–36 [PubMed: 17658549]
41. Merino E, Jensen RA, Yanofsky C (2008) Evolution of bacterial trp operons and their regulation. *Curr Opin Microbiol* 11(2):78–86 [PubMed: 18374625]
42. Minor W, Cymborowski M, Otwinowski Z, Chruszcz M (2006) HKL-3000: the integration of data reduction and structure solution—from diffraction images to an initial model in minutes. *Acta Crystallogr D Biol Crystallogr* 62(Pt 8):859–866 [PubMed: 16855301]
43. Moreno-Hagelsieb G, Collado-Vides J (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* 18(Suppl 1):S329–S336 [PubMed: 12169563]
44. Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* 53(Pt 3):240–255 [PubMed: 15299926]
45. Nusca TD, Kim Y, Maltseva N, Lee JY, Eschenfeldt W, Stols L, Schofield MM, Scaglione JB, Dixon SD, Oves-Costales D, Challis GL, Hanna PC, Pflieger BF, Joachimiak A, Sherman DH (2012) Functional and structural analysis of the siderophore synthetase AsbB through reconstitution of the petrobactin biosynthetic pathway from *Bacillus anthracis*. *J Biol Chem* 287(19):16058–16072 [PubMed: 22408253]
46. Olins PO, Rangwala SH (1989) A novel sequence element derived from bacteriophage T7 mRNA acts as an enhancer of translation of the lacZ gene in *Escherichia coli*. *J Biol Chem* 264(29):16973–16976 [PubMed: 2676996]
47. Osbourn AE, Field B (2009) Operons. *Cell Mol Life Sci* 66(23):3755–3775 [PubMed: 19662496]
48. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA* 96(6):2896–2901 [PubMed: 10077608]
49. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 96(8):4285–4288 [PubMed: 10200254]
50. Perrakis A, Romier C (2008) Assembly of protein complexes by coexpression in prokaryotic and eukaryotic hosts: an overview. *Methods Mol Biol* (Clifton, NJ) 426:247–256
51. Pflieger BF, Kim Y, Nusca TD, Maltseva N, Lee JY, Rath CM, Scaglione JB, Janes BK, Anderson EC, Bergman NH, Hanna PC, Joachimiak A, Sherman DH (2008) Structural and functional analysis of AsbF: origin of the stealth 3,4-dihydroxybenzoic acid subunit for petrobactin biosynthesis. *Proc Natl Acad Sci USA* 105(44):17133–17138 [PubMed: 18955706]
52. Price MN, Arkin AP, Alm EJ (2006) The life-cycle of operons. *PLoS Genet* 2(6):e96 [PubMed: 16789824]
53. Price MN, Huang KH, Alm EJ, Arkin AP (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res* 33(3):880–892 [PubMed: 15701760]
54. Prudencio M, Ubbink M (2004) Transient complexes of redox proteins: structural and dynamic details from NMR studies. *J Mol Recognit* 17(6):524–539 [PubMed: 15386621]
55. Pyndiah S, Lasserre JP, Menard A, Claverol S, Prouzet-Mauleon V, Megraud F, Zerbib F, Bonneu M (2007) Two-dimensional blue native/SDS gel electrophoresis of multiprotein complexes from *Helicobacter pylori*. *Mol Cell Proteomics* 6(2):193–206 [PubMed: 17092930]
56. Qi Y, Balem F, Faloutsos C, Klein-Seetharaman J, Bar-Joseph Z (2008) Protein complex identification by supervised graph local clustering. *Bioinformatics* 24(13):i250–i258 [PubMed: 18586722]
57. Qi Y, Bar-Joseph Z, Klein-Seetharaman J (2006) Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins* 63(3):490–500 [PubMed: 16450363]

58. Qi Y, Klein-Seetharaman J, Bar-Joseph Z (2007) A mixture of feature experts approach for protein–protein interaction prediction. *BMC Bioinformatics* 8(Suppl 10):S6
59. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, Chemama Y, Labigne A, Legrain P (2001) The protein–protein interaction map of *Helicobacter pylori*. *Nature* 409(6817):211–215 [PubMed: 11196647]
60. Razick S, Magklaras G, Donaldson IM (2008) iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* 9:405 [PubMed: 18823568]
61. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 437(7062):1173–1178 [PubMed: 16189514]
62. Selleck W, Tan S (2008) Recombinant protein complex expression in *E. coli*. *Curr Protoc Protein Sci Chapter 5:Unit 5 21*
63. Shi R, Munger C, Asinas A, Benoit SL, Miller E, Matte A, Maier RJ, Cygler M (2010) Crystal structures of apo and metal-bound forms of the UreE protein from *Helicobacter pylori*: role of multiple metal binding sites. *Biochemistry* 49(33):7080–7088 [PubMed: 20681615]
64. Stols L, Gu M, Dieckman L, Raffin R, Collart FR, Donnelly MI (2002) A new vector for high-throughput, ligation-independent cloning encoding a tobacco etch virus protease cleavage site. *Protein Expr Purif* 25(1):8–15 [PubMed: 12071693]
65. Stols L, Zhou M, Eschenfeldt WH, Millard CS, Abdullah J, Collart FR, Kim Y, Donnelly MI (2007) New vectors for co-expression of proteins: structure of *Bacillus subtilis* ScoAB obtained by high-throughput protocols. *Protein Expr Purif* 53(2):396–403 [PubMed: 17363272]
66. Tan S (2001) A modular polycistronic expression system for overexpressing protein complexes in *Escherichia coli*. *Protein Expr Purif* 21(1):224–234 [PubMed: 11162410]
67. Timsit Y, Acosta Z, Allemand F, Chiaruttini C, Springer M (2009) The role of disordered ribosomal protein extensions in the early steps of eubacterial 50 s ribosomal subunit assembly. *Int J Mol Sci* 10(3):817–834 [PubMed: 19399222]
68. Ubbink M (2009) The courtship of proteins: understanding the encounter complex. *FEBS Lett* 583(7):1060–1066 [PubMed: 19275897]
69. Vos MJ, Hageman J, Carra S, Kampinga HH (2008) Structural and functional diversities between members of the human HSPB, HSPH, HSPA, and DNAJ chaperone families. *Biochemistry* 47(27):7001–7011 [PubMed: 18557634]
70. Wojcik J, Schachter V (2001) Protein–protein interaction map inference using interacting domain profile pairs. *Bioinformatics* 17(Suppl 1):S296–S305 [PubMed: 11473021]

Fig. 1.

Operon statistics. **a** The operon prediction data available from MicrobesOnline of 1334 genomes was used to generate overall operon statistics. The percentage of all genes that are part of an operon with 2–5 members was calculated. Operons with 2 members account for more than 10 % of all genes. **b** The relative percentage of genes as part of an operon varied greatly between the 1334 species analyzed. On average 60 % of genes are organized in an operon. The *error bars* represent the standard deviation (SD)

Fig. 2.

Gene fusions. Gene fusion events were identified from the *H. pylori* genome using BLAST and the RefSeq database. The distribution of fused and single gene variants was calculated for the paired genes

Fig. 3.

Cloning strategies. **a** In the 'Operon-strategy' neighboring genes of an operon are amplified via gene-specific LIC primers and cloned into an N-terminal His₆ vector, pMCSG7 as shown in **b**. **c** In the 'Eps-RBS-fusion-strategy' the individual genes are amplified via gene-specific and an 'Eps-RBS'-specific primers in the first step. The individual PCR products are extended and further amplified via gene-specific LIC primers and cloned into an N-terminal His₆ vector, pMCSG7. **d** Similarly, a C-terminal His₆ vector, pMCSG28, can be used to generate C-terminal variants

Fig. 4.

Petrobactin synthesis. **a** Petrobactin is produced by the *B. anthracis* siderophore biosynthetic (*asb*) operon. The unusual precursor, the 3,4 isomer of dihydrobenzoic acid is synthesized by AsbF, while the condensation of spermidine and 3,4-dihydrobenzoic acid is performed by the concerted action of AsbCDE proteins. The synthesis of the larger intermediate and the final product are performed by AsbA and AsbB, respectively. **b** Gene fusion of AsbDE genes in *Rhodopseudomonas palustris*. **c** Co-elution of co-expressed *B. anthracis* His-tagged

AsbC and untagged AsbD genes (*left*) and co-elution of co-expressed *B. anthracis* His-tagged AsbD and untagged AsbE genes (*right*)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 5.

Small-scale expression using ‘Operon-strategy’ and IMAC I analysis of native operons. **a** Representative SDS-PAGE gels of expression analysis of clones expressing native operons. Bicistronic orthologous operons show similar stoichiometries and expression levels in spite of sequence divergence: [1, 3] 2-oxoglutarate synthase from *Bacteroides* and [2] acetolactate synthase from *Escherichia*. A representative tricistronic clone expressing xanthine dehydrogenase from *Clostridium* [4] displays an average expression level. **b** A selected set of clones expressing bicistronic or tricistronic operons at high level were further analyzed by

small-scale IMAC I. Most proteins remained complexed during the analysis and were tested in large-scale purification

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 6.

Large-scale SEC purification of native operon complexes. The large-scale purification of glycyl-tRNA synthetase and acetyl-CoA carboxylase from *K. pneumoniae* subsp. *pneumoniae* MGH 78578 was carried out with the semiautomated method. The IMAC-purified complexes were treated with TEV to remove the tag and subjected to size-exclusion chromatography (SEC) prior crystallization setup. **a** The SEC of the glycyl-tRNA synthetase (*top*) and acetyl-CoA carboxylase (*bottom*) samples. **b** The native gels of the IMAC and the SEC peaks of glycyl-tRNA synthetase and acetyl-CoA carboxylase. The main peaks (peak 2

for both) contain both subunits as verified via SDS-PAGE. **c** Representative crystals of the acetyl-CoA carboxylase complex

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 7.
The structures of *H. pylori* 3-oxoadipate CoA-transferase and molybdopterin converting factor complexes. **a** Cartoon showing the heterotetramer of 3-oxoadipate CoA-transferase formed by the interactions of YxjD and YxjE subunits. **b** The two subunits of the molybdopterin converting factor, MoaE and MoaD, interact via loop swapping. Crystals were grown using the ANL-1, ANL-2, and the MCSG1–4 screens (crystals used for data collection are shown)

Table 1
Comparison of cloning approaches for the production of interacting pairs of proteins

Method	Vectors	LIC primers	PCR reactions	Transformations	Transcripts	Eps-RBS sites
Dual vectors ^a	2	4	2	2 ^b	2	2
Operon ^c	1	2	1	1	1	1
Eps-RBS ^d	1	4	3	1	1	2

^aTwo vectors with different origins of replication
^bThe transformations can be performed simultaneously, but at lower efficiency
^cThe amplification generates a portion of a natural operon, with the original intergenic sequences
^dThe sequential amplifications create a synthetic operon. Each gene is preceded by its own eps-RBS

Table 2
The expression and solubility summary of *H. pylori* PP-CX expressed by the Eps-RBS strategy

gi	Annotation	gi	Annotation	Notes
2313492	Purine-binding chemotaxis protein (cheW)/AE000511	2313493	Histidine kinase (cheA)/AE000511	Poor solubility
2314361	Ribosomal protein L7/L12 (rpL7/L12)/AE000511	2314362	Ribosomal protein L10 (rpL10)/AE000511	Poor solubility
2313815	3-Oxoadipate CoA-transferase subunit A (yxjD)/AE000511	2313816	3-Oxoadipate CoA-transferase subunit B (yxjE)/AE000511	Soluble
2313153	Urease beta subunit (urea amidohydrolase) (ureB)/AE000511	2313154	Urease alpha subunit (ureA) (urea amidohydrolase)/AE000511	Soluble
2314120	Cell division protein (ftsA) protein/AE000511	2314121	Cell division protein (ftsZ)/AE000511	Cell did not grow
2313932	Molybdopterin converting factor, subunit 2 (moaE)/AE000511	2313939	Molybdopterin converting factor, subunit 1 (moaD)/AE000511	Soluble

Table 3

Data collection and refinement statistics

Protein name	Molybdopterin converting factor subunit 1 (MoaD) and 2 (MoaE)	3-Oxoadipate CoA-transferase subunit A (YxjD) and B (YxjE)
<i>Data collection</i>		
Space group	<i>F</i> 222	<i>C</i> 21
Cell dimensions (Å)	<i>a</i> = 88.2, <i>b</i> = 127.6, <i>c</i> = 187.9	<i>a</i> = 156.4, <i>b</i> = 65.7, <i>c</i> = 104.7, β = 109.8°
Radiation source	APS, 19-ID	APS, 19-ID
Wavelength (Å)	0.9794	0.9794
Resolution (Å) ^a	28.00–1.90	39.3–2.29
Unique reflections	41,239	44,755
Redundancy	3.7 (3.7)	3.7(3.6)
<i>R</i> _{merge} ^b	0.082 (0.65)	0.127 (0.83)
Mean <i>I</i> /sigma(<i>I</i>)	24.9 (2.35)	17.05(1.80)
Completeness (%)	99.8 (99.1)	99.9 (100)
Refinement		
<i>R</i> _{work} / <i>R</i> _{free} ^c	0.202/0.242	0.176/0.217
No. of protein residue/water molecules	866/71	435/120
Average B factor (Å ²) protein/water	45.18/40.55	44.51/43.29
RMSD from ideal		
Bond lengths (Å)	0.007	0.005
Bond angles (°)	0.985	1.019
Ramachandran plot (%)		
Favored/allowed/outliers	99.05/0.95/0	97.32/2.68/0

^aValues in parentheses correspond to the highest resolution shell^b $R_{\text{merge}} = \sum h \sum j |I_{hij} - \langle I_h \rangle| / \sum h \sum j I_{hij}$, where I_{hij} is the intensity of observation *j* of reflection *h*^c $R_{\text{free}} = \sum h |F_o - F_c| / \sum h |F_o|$ for all reflections, where *F_o* and *F_c* are observed and calculated structure factors, respectively. *R_{free}* is calculated analogously for the test reflections, randomly selected and excluded from the refinement