Integron-associated Mobile Gene Cassettes Code for Folded Proteins: The Structure of Bal32a, a New Member of the Adaptable $\alpha + \beta$ Barrel Family

Andrew Robinson$^1$, Peter S.-C. Wu$^2$, Stephen J. Harrop$^3,4$
Patrick M. Schaeffer$^2$, Zsuzsanna Dosztányi$^5$, Michael R. Gillings$^6$
Andrew J. Holmes$^6$, K. M. Helena Nevalainen$^6$, H. W. Stokes$^6$
Gottfried Otting$^2$, Nicholas E. Dixon$^2$, Paul M. G. Curmi$^3,4$ and
Bridget C. Mabbutt$^1*$

$^1$Department of Chemistry
Macquarie University, NSW 2109, Australia
$^2$Research School of Chemistry
Australian National University
ACT 0200, Australia
$^3$School of Physics, University of New South Wales, NSW 2052 Australia
$^4$Centre for Immunology
St Vincent’s Hospital, Sydney NSW 2010, Australia
$^5$Institute of Enzymology
Biological Research Center
Hungarian Academy of Sciences
Budapest, Hungary
$^6$Department of Biological Sciences, Macquarie University
NSW 2109 Australia

*Corresponding author

The wide-ranging physiology and large genetic variability observed for prokaryotes is largely attributed not to the prokaryotic genome itself, but rather to mechanisms of lateral gene transfer. Cassette PCR has been used to sample the integron/gene cassette metagenome from different natural environments without laboratory cultivation of the host organism, and without prior knowledge of any target protein sequence. Since over 90% of cassette genes are unrelated to any sequence in the current databases, it is not clear whether these genes code for folded functional proteins. We have selected a sample of eight cassette-encoded genes with no known homologs; five have been isolated as soluble protein products and shown by biophysical techniques to be folded. In solution, at least three of these proteins organise as stable oligomeric assemblies. The tertiary structure of one of these, Bal32a derived from a contaminated soil site, has been solved by X-ray crystallography to 1.8 Å resolution. From the three-dimensional structure, Bal32a is found to be a member of the highly adaptable $\alpha + \beta$ barrel family of transport proteins and enzymes. In Bal32a, the barrel cavity is unusually deep and inaccessible to solvent. Polar side-chains in its interior are reminiscent of catalytic sites of limonene-1,2-epoxide hydrolase and norgalonic acid methyl ester cyclase. These studies demonstrate the viability of direct sampling of mobile DNA as a route for the discovery of novel proteins.

Introduction

Recent analyses of complete bacterial and archaeal genomes have highlighted the importance of lateral gene transfer as a mechanism for generating genetic diversity among these organisms. Originally discovered for its role in the spread of antibiotic resistance genes, the integron/gene cassette system is now recognised as both an abundant resource and a significant participant in lateral gene transfer events. Integrons consist of a gene encoding a site-specific integrase, IntI, and an attachment site attI into which gene cassettes are inserted. Gene cassettes are the simplest mobile elements known, and typically contain a...
promoterless gene and a recombination site known as a 59 base element.\textsuperscript{9,11} Gene cassettes are inserted into (or excised from) integrons through the action of IntI1 to form a cassette array downstream of an integron-associated promoter that controls expression of the newly inserted genes.\textsuperscript{2,10} In this way, the cassettes are responsible for the clustering of a significant proportion of novel genes in bacteria.

We recently demonstrated the technique of cassette PCR for the amplification of gene cassettes, irrespective of their encoded gene sequences.\textsuperscript{12} The PCR-based technique utilizes primers targeting common features of the 59 base elements that flank each gene within an integron cassette array.\textsuperscript{12} Cassette PCR thus provides the opportunity to sample the integron/gene metagenome present in a multitude of environmental ecosystems without the need for laboratory cultivation of a host organism, and without prior knowledge of any target protein sequence. The technique also allows the recovery of complete genes in a single step, e.g. ready for insertion into a protein expression vector for further characterisation.\textsuperscript{13} From within our cassette-associated gene pool recovered to date from both pristine and industrially contaminated sites, we have identified genes showing some sequence homology to known enzymes of secondary metabolic function such as phosphotransferases, DNA glycosylases, methyl transferases and thiotransferases.\textsuperscript{7} Initial studies targeting cassette-encoded genes with identifiable homologs have demonstrated that some at least code for folded and functional proteins.\textsuperscript{13}

The vast majority (over 90\%) of our collection of amplified gene cassettes, however, contain orphan ORFs (ORFans),\textsuperscript{14} i.e. genes with no identifiable homolog in the sequence databases. This high percentage emphasises the degree to which cassette-PCR is able to tap into a wide pool of genetic diversity; generally ORFans have been found to account for 25–30\% of a newly sequenced genome.\textsuperscript{15} The question remains as to whether these novel cassette-encoded genes code for functional proteins. The predicted proteins are in general quite small, with 60\% containing 70–140 amino acid residues, and only 10\% being over 200 amino acid residues.\textsuperscript{5} The novel genes may thus correspond to single-domain proteins, and potentially could include protein fold classes not isolated previously. It is clear that information beyond sequence is required to ascertain the structural integrity of the gene products and to infer their functions, which presumably reflect biochemical processes that relate to evolutionary advantage.

Given that little can be learnt about ORFans via homology relationships, one route to functional annotation is through elucidation of three-dimensional structure, allowing subsequent identification of functional relationships with identified structural homologs.\textsuperscript{16,17} Here, we describe a systematic program to structurally sample a panel of cassette-associated genes whose sequences appear highly likely to encode fully folded proteins. At the time of cloning, none of the eight selected targets had any significant sequence identity with any functionally or structurally characterised protein. Of this initial group, five genes were successfully isolated as soluble proteins using a heterologous \textit{Escherichia coli} host, and each showed biophysical characteristics consistent with a folded, native structure. One protein produced crystals suitable for structure determination at 1.8 Å resolution by X-ray crystallography. The resulting structure belongs to a fold class associated with diverse functions, often in highly stressed natural environments.

\section*{Results}

\subsection*{Selection of cassette-encoded genes and recombinant production}

Over 100 genes derived from environmental gene cassettes were potentially available for study.\textsuperscript{12} Targets were selected from this field (of 106) based on several criteria designed to maximise the chances of defining a novel domain and at the same time the chances of successful structural studies. First, all protein sequences were removed which contained possible transmembrane helical regions according to prediction servers, leaving 57 genes. Of those, only sequences coding for fewer than 160 amino acid residues and <3 Cys residues were retained (31 genes). Subsequently, sequences for which globular fold features were predicted were selected, leaving 13 targets. Three genes were eliminated on the basis of codon usage likely to compromise expression in \textit{E. coli} and, finally, two further genes were removed as sequence homologs were identified. The eight target sequences selected for structural study are characterised in Table 1.

The gene encoding each target was amplified and inserted into an \textit{E. coli} expression vector for production as an N-terminal Met-His\textsubscript{6}-fusion protein. Due to difficulties with transformation, one target (SM57c2) was dropped from the study at this early cloning stage. For three of the remaining targets (Bal2-1, Bal32a and Bal50b), soluble recombinant products were obtained in high yield (>15 mg/l of culture). A fourth protein, HB1-2, was produced in limited yield as a soluble product using slow induction of expression. Of the proteins expressed into insoluble inclusion bodies, one (Pu17) was able to be refolded and purified.

\subsection*{Solution characteristics of protein products}

Following purification by affinity chromatography, solution-based techniques were used to confirm the presence of tertiary folds for the five isolated proteins. Solubility ranged from <1 mg/ml (HB1-2 and Pu17) to over 25 mg/ml (Bal32a and Bal50b) in phosphate and Tris-based buffers. One-dimensional NMR spectra showed well-dispersed features (see Figure 1(a)), confirming the
presence of secondary structure elements for all products. The line-widths of the NMR signals, while consistent with single fold domains, were broad enough to suggest the presence of oligomeric species, particularly for Bal2-1 and Bal32a samples.

For three of the protein solutions, quaternary structures were more clearly revealed by size-exclusion chromatography (Figure 1(b) and Table 1). Under reducing conditions, Bal2-1 was identified as a 40 kDa tetramer, Bal32a as a 35 kDa dimer, and Bal50b as a mixture of predominantly dimeric and trimeric species (25 kDa and 38 kDa). When isolated, the Bal50b trimers were found to reorganise as dimers over time, suggesting the dimer to be the more stable form in solution. Circular dichroism (CD) spectra were used to characterise the secondary fold features and thermal stability of the targets. Both the Bal2-1 tetramer and Bal32a dimer appear to be rich in β-sheet secondary structure, yielding CD data consistent with 35–40% sheet and 5% helix content (Figure 1(c)). The CD spectra of Bal50b dimers and trimers, while containing an unusual feature at 230 nm, are both consistent with approximately 40% β-sheet structure. These secondary structure assessments indicate less helical but greater β-sheet content for our targets than initially predicted from sequence (see Table 1). Clear melting temperatures were determined for Bal2-1 and Bal32a, 81°C and 65°C, respectively, confirming the stability of their tertiary folds. The melting temperature of the Bal32a dimer is increased to 80°C in the absence of a reducing agent, suggesting extra stability due to the formation of a disulphide bond. This bond is likely to be intermolecular, as the only two Cys residues of Bal32a are immediately adjacent in sequence.

Crystal structure of Bal32a

Three of the isolated proteins, Bal2-1, Bal32a and Bal50b, were screened using commercially available crystallisation kits. Diffraction-quality crystals of Bal32a were grown directly from wells containing 0.1 M sodium acetate (pH 4.6), 2 M ammonium sulfate. The structure of Bal32a was determined by X-ray diffraction at 1.8 Å resolution using phases obtained from a single Xe heavy-atom derivative (Table 2). Electron density was continuous from Asn19 to Arg149 of the sequence.

Bal32a forms a dimeric structure, featuring two α+β barrel subunits related through a crystallographic 2-fold symmetry axis (Figure 2(a)). The two subunits interact at the exposed “back” face of each β-sheet, primarily through hydrophobic interactions. Although the crystallised protein was in the reduced state, the two Cys121 side-chains are within close proximity at the dimer interface, suggesting an intermolecular disulphide bridge may readily form in the absence of reducing agent (supporting CD data above). The preceding residue, Cys120, is unpaired and is oriented towards the protein interior.
The overall shape of each α + β barrel fold is that of a cone, dominated by a six-stranded β-sheet that is highly curved (secondary structure elements are numbered in Figure 2(a)). In particular, four strands (named S3, S4, S5 and S6) are bent about their midpoints and provide a trough-shaped scaffold. The two N-terminal strands, together with three helical portions, serve to close the barrel, forming a large central cavity (Figure 2(a)). The short C-terminal helix D contains a number of hydrophobic residues and, together with the S1/S2 loop, forms a cap that completely blocks the larger end of the cone (the “top” of the cavity), isolating it from the external solvent.

The cavity encloses a volume of 280 Å³, with maximum dimensions approximately 17 Å × 5 Å × 5 Å. Located at its top are residues of helix D (Leu141, Trp144, Phe147, Trp148) and the S1/S2 loop (Pro56–Ala62). The S1/S2 loop contains a largely polar sequence (-PGRSRIA-) whose residues show moderate to high B-factors, indicating some degree of flexibility or static disorder. This loop appears to be held in place by only two hydrogen bonds: between Asp145 and the backbone N of Arg58, and between Gln74 and the backbone O of Ile61. These features are suggestive of the S1/S2 loop being part of a removable cap to the barrel cavity. Nearby residues at the C-terminal ends of helices C and D also show elevated B-factors.

The cavity is largely lined with hydrophobic side-chains (Thr27; Val28 and Val131; Gly31; Tyr32 and Tyr78; Phe35, Phe47, Phe75, Phe86, Phe124 and Phe138; Leu43 and Leu46; Trp53; Ala88; Ile122) (Figure 2(b)). However, a distinct ring of more polar side-chains (His105, Asn107, Glu136) and a water molecule form an extensive hydrogen bonding

**Figure 1.** Biophysical characterisation of recombinant products of integron-encoded genes. (a) The 600 MHz 1H NMR spectra of Bal2-1, Bal32a, Bal50b and HB1-2 at 25 °C, pH 7.3–7.5. Intense narrow lines are from low molecular mass impurities. (b) Size-exclusion chromatography of recombinant preparations of Bal2-1 (above) and Bal32a (below) on Superose-12, eluting in Tris/DTT buffer (pH 7.9) at 60 µl/minute. Elution times were calibrated as indicated. Bal2-1 elutes with an apparent size of 40 kDa (tetramer), and includes a minor species corresponding to 20 kDa (dimer). Bal32a elutes as a primary species of 35 kDa (dimer), with a minor component visible as 22 kDa (monomer). (c) CD spectra of Bal2-1 in Tris/DTT buffer (pH 7.9) at 25 °C and 95 °C. The inset shows the first derivative of the CD signal at 220 nm against temperature, indicating a single, sharp transition at 81 °C.
network around one portion of the cavity, towards its top (Figure 3(a)). This clustered arrangement of polar groups is highly suggestive of a catalytic site. When screened against an enzyme active-site template database, however, the side-chain geometry did not match any currently known spatial arrangements of residues.

Electron density consistent with a molecule of water above a small organic molecule (about six to ten non-H atoms) is clearly visible within the cavity, detailed in Figure 2(b) and (c). The ring of polar residues surrounding the top of the cavity suggests that one end of the unknown small molecule is polar in nature. The remaining portion of the organic species is likely to be non-polar, as surrounding side-chains are hydrophobic. In an attempt to identify this small molecule, several target species were modelled into the electron density, which was then refined. Features of the modified density maps are consistent with the polar end of the small molecule containing a carboxyl group or similar. However, little improvement was seen in the refinement statistics (crystallographic R-values) with the inclusion of any ligand. In any case, the electron density observed does not occupy the entire central cavity, thus the unidentified small molecule is unlikely to be the natural ligand of Bal32a.

**Structural homologs of Bal32a**

A search for structural homologs of the Bal32a tertiary structure identified a family of very close fold relatives, several showing rmsd agreement to within 2 Å for main chain Cα atoms (ranked in Table 3 and overlayed in Figure 2(d)). These proteins are all members of the α + β barrel fold superfamily, known to be highly diverse in both sequence and function. Like Bal32a, many α + β barrel fold proteins are dimeric, with the β-sheet forming the dimer interface. These close structural relatives of Bal32a include the enzymes ketosteroid isomerase (KSI), nogalonic acid methyl ester cyclase (known as SnoaL) and scytalone dehydratase (SD). The overlay of Figure 2(d) emphasises the strong conservation across the family of β-strand elements forming the barrel scaffold, with most variations from Bal32a occurring at the top of the cone. The aligned regions of secondary structure identified by this overlay were used as a basis for the construction of a sequence alignment of these proteins and their α + β barrel relatives (Figure 4).

A notable feature of most members of the superfamily is a solvent-accessible ligand-binding pocket accommodated at the larger end of the cone-shaped α + β barrel (represented schematically in Figure 3(b)). The pocket is lined primarily with hydrophobic residues, with the catalytic residues (where present) residing at the innermost point of the pocket. In all these cases, a conserved bulky residue (most commonly aromatic) forms the bottom of the cavity, e.g. Tyr14 in KSI, Trp31 in SD. Bal32a differs considerably from its structural homologs in this region, with the equivalent position being occupied by Gly31. This creates an unusual extension of the cavity, which protrudes deep into the Bal32a barrel fold. The bulky aromatic residues of helix D and the extended S1/S2 loop capping the cavity in Bal32a also appear to be unique to this new member of the α + β barrel family. As a result, the deep central cavity of Bal32a appears to be unusually inaccessible to solvent.

The structure-based alignment (Figure 4) shows that while sequence homology is weak overall (all <25% identity), several positions are strongly conserved. These conserved residues tend to be of structural importance, e.g. an Asn/Asp side-chain anchors the turn between helices A and B; a Gly

---

**Table 2. Crystallographic data collection and refinement statistics for Bal32a structure determination**

<table>
<thead>
<tr>
<th>Data set</th>
<th>Resolution (Å)</th>
<th>Total reflections</th>
<th>Unique reflections</th>
<th>I/σ</th>
<th>Completeness (%)</th>
<th>Multiplicity</th>
<th>Rmerge*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native (outer shell)</td>
<td>1.85 (1.95–1.85)</td>
<td>116,213 (16,423)</td>
<td>15,010 (2133)</td>
<td>20.9 (3.2)</td>
<td>99.9 (100)</td>
<td>7.7 (7.7)</td>
<td>0.063 (0.586)</td>
</tr>
<tr>
<td>Native (outer shell)</td>
<td>2.2 (2.32–2.20)</td>
<td>59,199 (8319)</td>
<td>8982 (1268)</td>
<td>17.8 (4.3)</td>
<td>99.8 (99.9)</td>
<td>6.6 (6.6)</td>
<td>0.085 (0.427)</td>
</tr>
<tr>
<td>Xenon (outer shell)</td>
<td>2.4 (2.53–2.40)</td>
<td>94,732 (13,675)</td>
<td>6956 (990)</td>
<td>34.0 (12.8)</td>
<td>99.9 (100)</td>
<td>13.6 (13.8)</td>
<td>0.060 (0.200)</td>
</tr>
</tbody>
</table>

**Refinement statistics**

<table>
<thead>
<tr>
<th>Water molecules</th>
<th>75</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rmerge/Rfree</td>
<td>0.187/0.212</td>
</tr>
<tr>
<td>Reflections in Rfree/Rfree</td>
<td>14,962/758</td>
</tr>
<tr>
<td>Resolution range (Å)</td>
<td>1.85–22.47</td>
</tr>
<tr>
<td>Isotropic mean R-factors overall/main-chain (Å²)</td>
<td>35.0/33.3</td>
</tr>
<tr>
<td>rmsd bond lengths (Å)</td>
<td>0.020</td>
</tr>
<tr>
<td>rmsd bond angles (deg)</td>
<td>1.752</td>
</tr>
<tr>
<td>Ramachandran plot</td>
<td>Most favoured region (%)</td>
</tr>
<tr>
<td>Additionaly allowed (%)</td>
<td>8.1</td>
</tr>
</tbody>
</table>

---

* From PROCHECK.
residue allows a change of chain direction just N-terminal to helix C. Residues involved in the catalytic functions of the different enzymes (shaded in black in Figure 4 where known) are generally varied in location across the sequences. In KSI, catalytic residues lie on helix A (Tyr14), strands S1 (Asp38) and S5 (Asp99), yet in SnoaL, residues of C-terminal strands S5 (Gln105) and S6 (Asp121) are functional.

At one point on β-strand S6, however, a nucleophilic Asp residue is common to two members of the family, SnoaL (Asp121) and limonone-1,2-epoxide hydrolase (LEH) (Asp132). The structurally equivalent position in Bal32a maintains this homology, being occupied by Glu136. This side-chain forms part of the polar region identified earlier as a putative catalytic site within the Bal32a structure, and is hydrogen bonded to the water molecule within the cavity (Figure 3(a)). On close inspection, many of the bonding and spatial features of the active site of LEH (detailed in Figure 3(c)) correspond closely with this specific.

Figure 2. The 1.8 Å crystal structure of Bal32a. (a) Ribbon diagram of the Bal32a dimer coloured by β-factor values, ranging from 21.9 Å² (blue) to 70.1 Å² (red). The cone shaped subunits are formed by means of a highly curved β-sheet, closed by three α-helices. Secondary structure encompasses: Glu20–Phe35 (helix A), Asp39–Glu45 (helix B), Ala51–Thr55 (S1), Gly63–Gly67 (S2), Arg68–Gly80 (helix C), Lys100–Arg111 (S4), Lys114–Lys127 (S5), Arg130–Tyr139 (S6), Leu141–Trp148 (helix D). (b) Part of the 1.8 Å 2Fₒ−Fᵣ map (contoured at a level of 1σ) showing some of the residues that line the central cavity and an area of electron density attributed to an unidentified small molecule. (c) Stereo view of the Bal32a monomer showing the location of the central cavity (blue transparent surface) and the region of the electron density (2Fₒ−Fᵣ map contoured at 1σ) arising from an unknown small molecule (black mesh). (d) Stereo view showing a Cα trace of Bal32a (red) overlayed with D5-3-ketosteroid isomerase, PDB 1OH0 (blue), nogalonic acid methyl ester cyclase, PDB 1SJW (cyan) and Apc116, PDB 1S5A (green). Figures were created using PyMol (DeLano, W. L. (2002). The PyMOL Molecular Graphics System, San Carlos, CA.).

Figure 3. Comparison of the putative catalytic site of Bal32a with active sites of close structural relatives. (a) In the structure of Bal32a, a ring of polar residues form a hydrogen bonded network surrounding the central cavity. (b) Schematic diagram contrasting features of the Bal32a cavity to those of other α+β barrel proteins. The position of the ring of polar residues is maintained, but in Bal32a the barrel is capped by side-chains of helix D. (c) Detailed view of the known active site of limonone-1,2-epoxide hydrolase (LEH), and (d) nogalonic acid methyl ester cyclase (SnoaL).
<table>
<thead>
<tr>
<th>PDB file</th>
<th>Z-score&lt;sup&gt;a&lt;/sup&gt;</th>
<th>rmsd (Å)&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Sequence identity (%)</th>
<th>Name</th>
<th>Ascribed function</th>
<th>Quaternary structure</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1oh0</td>
<td>15.9</td>
<td>1.9</td>
<td>21</td>
<td>Δ&lt;sup&gt;2&lt;/sup&gt;-3-Ketosteroid isomerase</td>
<td>Binds Δ&lt;sup&gt;2&lt;/sup&gt;-3-ketosteroids, cleaves C–H bond adjacent to carbonyl, enol formation</td>
<td>Homodimer (α+β barrel)</td>
<td>19</td>
</tr>
<tr>
<td>1sjw</td>
<td>14.6</td>
<td>2.6</td>
<td>13</td>
<td>Nogalonic acid methyl ester cyclase</td>
<td>Binds polyketides, ring closure</td>
<td>Homotetramer (α+β barrel)</td>
<td>20</td>
</tr>
<tr>
<td>1s5a</td>
<td>13.8</td>
<td>2.4</td>
<td>25</td>
<td>Hypothetical protein apc1116</td>
<td>Acetate and glycerol bound, function unknown</td>
<td>Unknown</td>
<td>51</td>
</tr>
<tr>
<td>1oun</td>
<td>13.7</td>
<td>2.3</td>
<td>10</td>
<td>Nuclear transport factor 2</td>
<td>Binds RanGDP and FxFG nucleoporins, transport</td>
<td>Homodimer (α+β barrel)</td>
<td>25</td>
</tr>
<tr>
<td>1nu3</td>
<td>13.5</td>
<td>2.4</td>
<td>22</td>
<td>Limonene-1,2-epoxide hydrolase</td>
<td>Binds limonene-1,2-epoxide, cleaves C–O epoxide bond</td>
<td>Homodimer (α+β barrel)</td>
<td>22</td>
</tr>
<tr>
<td>3std</td>
<td>13.3</td>
<td>2.5</td>
<td>8</td>
<td>Scytalones dehydratase</td>
<td>Binds scytalone, cleaves C–H bond adjacent to carbonyl, dehydration</td>
<td>Homotrimer (α+β barrel)</td>
<td>21</td>
</tr>
<tr>
<td>1pn5</td>
<td>13.2</td>
<td>1.9</td>
<td>14</td>
<td>p15</td>
<td>mRNA transport (complex with Tap)</td>
<td>Heterodimer (α+β barrels)</td>
<td>52</td>
</tr>
<tr>
<td>1m98</td>
<td>11.8</td>
<td>1.8</td>
<td>14</td>
<td>Orange carotenoid protein</td>
<td>Binds carotenoids, photoprotection</td>
<td>Homodimer (α+β barrel plus α-helical domain)</td>
<td>53</td>
</tr>
<tr>
<td>1hkx</td>
<td>11.8</td>
<td>3.4</td>
<td>9</td>
<td>Cam-kinase II, α subunit</td>
<td>Possible protein docking, interacts with kinase domain</td>
<td>Tetradecamer (α+β barrel)</td>
<td>54</td>
</tr>
<tr>
<td>1pn5</td>
<td>11.3</td>
<td>2.5</td>
<td>10</td>
<td>Tap</td>
<td>FG repeat recognition, mRNA transport (complex with p15)</td>
<td>Heterodimer (α+β barrels)</td>
<td>52</td>
</tr>
<tr>
<td>1of5</td>
<td>10.8</td>
<td>2.9</td>
<td>10</td>
<td>mtr2</td>
<td>mRNA transport (complex with Mex67)</td>
<td>Heterodimer (α+β barrels)</td>
<td>55</td>
</tr>
<tr>
<td>1ndo</td>
<td>10.1</td>
<td>2.8</td>
<td>4</td>
<td>Naphthalene-1,2-dioxygenase, β subunit</td>
<td>Associates with α subunit to form active enzyme</td>
<td>Heterodimer (α+β barrels)</td>
<td>55</td>
</tr>
<tr>
<td>1of5</td>
<td>9.7</td>
<td>2.9</td>
<td>8</td>
<td>Mex67</td>
<td>FG repeat recognition, mRNA transport complex with mtr2</td>
<td>Heterodimer (both α+β barrels)</td>
<td>55</td>
</tr>
</tbody>
</table>

Three-dimensional structures were identified by DALI<sup>33</sup> as structural relatives of Bal32a.

<sup>a</sup> Statistical significance of best alignment.

<sup>b</sup> Positional deviation calculated over main-chain C<sup>a</sup> atoms as output by DALI.
region of the Bal32a structure. In LEH, Asp101 on strand S4 is known to be vital for activity, and is extensively hydrogen bonded both to the substrate and to Arg99 (adjacent on strand S4). In Bal32a, the equivalent strand S4 position is occupied by Asn107, and is connected via analogous hydrogen bonding interactions to the cavity water molecule and to the strand S4 residue His105 (Figure 3(a)). The active site within the cyclase SnoaL also contains a similar array of bonding interactions (Figure 3(d)). It centres around Gln105, which is hydrogen bonded to the reaction product and to the side-chain of His107. (Gln105 resides on strand S5 of SnoaL, but spatially overlays the strand S4 Asn107 position of Bal32a). The structural similarity of the polar region encircling the Bal32a cavity to the active sites of LEH and SnoaL suggests that Bal32a is likely related to an ancestral form of these enzymes, in which the Asn107 position may have played an active functional role.

During the course of this study (due to increases in sequence databases), an iterative PSI-BLAST search using the sequence of Bal32a produced sequence matches with members of the “KSI-related” family (with 20–33% identity, expectation E-values <10^-46) and to the consensus sequence COG3631. Note that key active site residues of KSI (namely Tyr14, Tyr30, Asp38, Asp99 and Tyr55) are, however, absent in the Bal32a sequence.

Discussion

One of the key questions regarding the
integron/cassette system is: do the cassette genes code for functional proteins with a stable, folded native structure? Previous studies have answered this question in the affirmative for cassette genes coding for antibiotic resistance proteins and proteins with known homologs. Our results examine this question for a sample of proteins from the 90% of cassette genes with no known homologs. Of eight protein products we were able to express and purify, four showed significant characteristics of folded proteins (size-exclusion chromatograms, CD spectra, melting curves and NMR spectra), whilst a fifth indicated some globular features upon refolding. Thus, our study extends support for the hypothesis that the integron/cassette system actively mobilises functional genes that code for active, folded proteins.

In the case of Bal32a, we have obtained a crystal structure showing that it is a member of the \(\alpha + \beta\) barrel superfamily, a well-described group (e.g. CATH 2.20.25.70)\(^{24}\) of diverse sequence and function.\(^{21,22,25}\) The members of this protein family include enzymes, nuclear transport factors, RNA transport proteins and numerous binding proteins, most of which are dimeric. The common feature of all of these proteins is the central cavity of the barrel, used to bind either a hydrophobic small molecule or an extended polypeptide chain. Thus, the \(\alpha + \beta\) barrel fold provides a robust scaffold whose function can be tuned by altering the properties of the residues lining the interior cavity. This correlates with the observed functional diversity, but, in turn, makes it difficult to predict the precise function of Bal32a from its structure.

The Bal32a structure we have defined is a new member of the \(\alpha + \beta\) barrel structural family, with yet another arrangement of putative catalytic residues within the preserved cone-like cavity. Bal32a is unique in that the central cavity is unusually deep and isolated from the solvent by a lid formed from helix D and the S1/S2 loop. A potential catalytic site has been identified within the barrel scaffold that is in an equivalent position to the active sites of LEH and SnoaL. It might therefore be suggested that these enzymes share a common evolutionary ancestry, with the position of the active site remaining preserved while specific catalytic groups have been altered to meet particular demands.

Whether Bal32a is in fact an enzyme or a binding/transport protein remains to be established. Assaying the activity of Bal32a may indeed be rendered difficult by the closed nature of its cavity. The gene cassette containing Bal32a was originally recovered from a site contaminated with a number of organochlorine pesticides, as well as high levels of \(\text{Cu}, \text{Zn}\) and \(\text{Mg}\).\(^{26}\) Further to explore any connection of mobilised gene to environment type, primers internal to the Bal32a gene (encoding the segment Val28–Leu90) were used in PCR on total environmental DNA from a variety of sources (S. Krishnan & M.G., unpublished results). Products were obtained from only a pair of independent soil sites, the only two subject to industrial contamination. The two sites were remote from one another, and included the power station site from which Bal32a was originally isolated. On sequencing, the products showed minor variations, in all cases involving conservative variations of surface residues. Residues within our proposed active site of Bal32a (Tyr32, Trp53 and Tyr78) were invariant. To isolate variants of Bal32a from environmental DNA samples from two distinct sites confirms that the gene is likely to be active. That both of the sites were contaminated by industrial waste leads to the speculation that Bal32a might be associated with waste compound degradation.

Finally, the process of cassette PCR most commonly results in the amplification of a single gene cassette, and hence sequence information concerning the associated cassettes is lost. In the case of the cassette PCR product from which Bal32a was derived, however, two gene cassettes were obtained from the one product. The second gene cassette codes for a theoretical protein, termed Bal32b,\(^{12}\) with a predicted transmembrane helix at the N-terminus. Bal32b currently shows no significant sequence homology with any protein of known structure or function. It is, however, reasonable to speculate that these two gene products may act together in a concerted fashion.

### Materials and Methods

#### Environmental source

The target genes were the products of environmental gene cassettes sourced from both contaminated and pristine sites, as described.\(^{75}\) The soil from which Bal2-1, Bal32a and Bal50b were derived was chemically analysed (Pacific Power, AWT EnSight, Sydney and Biological and Chemical Research Institute, Sydney), and found to be contaminated with several organochlorine pesticides, including Arochlor 1248 (polychlorobiphenyl compounds), aldrin and dieldrin. In addition, metals (Cu, Zn and Mg) were concentrated up to 50,000-fold over levels associated with pristine soils.\(^{26}\)

#### Bioinformatics methods

The sequences were characterized using TMHMM,\(^{27}\) HMMTOP\(^{28}\) and DAS\(^{29}\) programs for transmembrane helix prediction and fold characteristics assessed with tools from the PredictProtein server.\(^{26}\) Sequence homology searching was performed using PSI-BLAST\(^{31}\) against the non-redundant database as at December 2002. Searching of databases with the Bal32a sequence in mid-2004 included a sequence search profile built using an acceptance threshold of \(E\)-value \(< 10^{-9}\) and iterated until convergence (five rounds). Results included identification of the COG3631 domain by the Conserved Domain database.\(^{32}\)

DALI\(^{33}\) was used for protein structure comparisons against the PDB database, and the PROCAT\(^{34}\) database was searched for enzyme active-site templates (March 2004). Pair-wise least-squares structure alignments were generated between Bal32a and each of the identified structural homologs using LSQMAN as implemented in...
the program O35 and based on C* atoms. Following the superposition of backbone structures, those regions for which C* atoms overlaid within 3.8 Å were output by O as fragments of aligned sequence. These fragments were compiled to produce a pair-wise sequence alignment for each of the homologs against the complete sequence of Bal32a (Figure 4).

**Gene expression**

Target genes were amplified by PCR from the original pGEM-T Easy (Promega) constructs.37 An NdeI site was introduced at the ATG start codon and an EcoRI site was provided beyond a TAA stop codon. The genes were inserted between these restriction sites in the phage T7 promoter-based vector pETMCSIII36 and transformed into the E. coli strain BL21(DE3)lysS for protein expression.37 Nucleotide sequences were determined using an ABI 3730 sequencer (Biomolecular Resource Facility, ANU), following the recommendations of the manufacturer (Applied Biosystems). The sequences verified correct insertion of each gene, preceded in the vector by codons for a Met-(His)6, bag.

Typically, bacterial cells were grown at 37 °C in Luria–Bertani broth containing ampicillin (50 μg/ml) and chloramphenicol (50 μg/ml). Gene expression was induced by addition of isopropyl-β-D-thiogalactoside (0.5 mM) when cultures reached absorbance (A595) of 0.5–1.0 (or 1.3, in the case of Bal2-1). After three hours, cells from two to five litres of culture were pelleted by centrifugation, resuspended in 15 ml/l of culture of buffer A (50 mM sodium phosphate (pH 7.7), 300 mM NaCl, 2 mM β-mercaptoethanol) containing 10 mM spermidine (15 mM/g of cells) and lysed in a French press. Solubility was determined by comparing the amount of protein in the cell lysate with that in the supernatant fraction on an SDS–PAGE gel. Strains from which the integron-encoded proteins did not partition to the soluble cell fraction were also grown for two days at lower temperature (−20 °C) in an autoinduction medium (Overnight Express, Novagen). For recovery of Pu17, inclusion bodies were purified from the supernatant by Ni-NTA affinity chromatography in buffer A, eluting purified proteins did not partition to the soluble cell fraction were also grown for two days at lower temperature (−20 °C) in an autoinduction medium (Overnight Express, Novagen).36

The His-tagged soluble recombinant proteins were purified from the supernatant by Ni-NTA affinity chromatography in buffer A, eluting purified proteins by batch-wise applications of buffer A containing increasing concentrations of imidazole (over the range 50–175 mM). For recovery of Pu17, inclusion bodies were solubilised in buffer B (8 M urea, 10 mM Tris–HCl, 100 mM sodium phosphate (pH 8.0)). After centrifugation, the supernatant was diluted fourfold with buffer B and loaded onto a Ni-NTA column equilibrated with buffer B (pH 6.3). Pu17 was eluted using buffer B (pH 5.9) and refolded by rapid dilution (1–5 ml) into 11 of PBS buffer (10 mM sodium phosphate, 138 mM NaCl, 2.7 mM KCl (pH 7.4)). After two hours, the pH was slowly adjusted to 7.4 and the protein solution re-loaded onto the Ni-NTA column for concentration. After washing with 10 mM imidazole solution a (pH 8.0), Pu17 was eluted with 0.3 M imidazole–HCl (pH 3.7) and dialysed into PBS buffer (pH 7.4).

**Protein characterisation**

Electrospray ionization mass spectrometry was used to confirm the sequence of recombinant products. Samples were exchanged into high-purity water on a desalting column and analysed with a triple quadrupole mass spectrometer (Micromass Quattro II) operating in positive-ion mode. Molecular mass values obtained for the Met-His6-tagged proteins were within 0.5 unit of calculated masses: Bal2-1 (10,116), Bal32a (17,736) and Bal50b (12,405). Attempts to identify the trapped ligand within Bal2a by LC-MS were unsuccessful; a low molecular mass species with absorbance at 254 nm was observed in an acetone extract, but failed to ionise in the electrospray source.

H NMR spectra were acquired on a Varian INOVA 600 MHz NMR spectrometer using the Watergate water suppression sequence.38 NMR samples were prepared in 90% H2O/10% 2H2O containing 10 mM sodium phosphate, 50 mM NaCl, 0.02% (w/v) sodium azide. A pH value of pH 7.5 was used for all samples except Bal2-1 (pH 7.3). CD spectra were recorded from 300 nm to 200 nm in Tris/DTT buffer (10 mM Tris (pH 7.9), 2 mM dithiothreitol) on a JASCO 810 spectropolarimeter. Deconvolution of spectra was performed in the CDPro software package using a 43 protein reference set.39 Thermal melting of protein samples was achieved by monitoring the CD signal at its maximum value whilst ramping the temperature at 3 deg./C/minute. Size-exclusion chromatography was performed in a 30 ml l/g of precipitant) using Crystal Screen, Crystal Screen 2 and PEG/Ion Screen kits (Hampton Research).

**Crystallisation and structure determination of Bal32a**

Diffraction-quality crystals of Bal32a grew at room temperature over a two-week period from 0.1 M sodium acetate (pH 4.6) with 8% (w/v) PEG 4000. Larger crystals grew from 0.1 M sodium acetate (pH 4.6) with ammonium sulphate (2 M) and these were used for X-ray diffraction experiments. Crystals belonged to the space group P412121 (a = 60.728 Å, c = 89.910 Å). A native data set (1.8 Å) was recorded at 100 K, using a crystal previously soaked in 15% (w/v) glucose, 15% (w/v) PEG 400 prior to flash-freezing in liquid N2. For a second native data set (2.2 Å), a crystal was prepared by dragging twice through Paratone N (Hampton) prior to flash-freezing. For the recording of a Xe derivative data set, a crystal was drogged twice through Paratone N, treated in a Xe pressure cell (Hampton) at 3.5 MPa for five minutes, and flash-frozen in liquid N2.

Diffraction data were collected at 100 K on a Mar345db imaging plate detector mounted on a Nonius Cu rotating anode generator. Data were processed using MOSFLM40 SCALa41 and CCP4 software.42 MULTAN43 was used to locate isomorphous and anomalous scattering, producing comparable three-site solutions with occupancies of approximately 0.37, 0.18 and 0.07. Experimental phases were obtained using MLPHARE44 to 2.4 Å resolution using both isomorphous difference and anomalous scattering data from the Xe derivative (overall calculated figure of merit for the phases was 0.51 (outer shell (2.72–2.40 Å) 0.37) with an overall phasing power of 1.5). Phase extension was carried out through DM,45 using solvent flattening and histogram matching to the 1.8 Å native data set. ARP/wARP46 was used for automated model building. The model was refined using REFMAC47 with five cycles of refinement, using individual isotropic restrained thermal factors, a weighting term of 0.5 and
both Babinet bulk solvent and anisotropic scaling (Table 2). The model was built and visualised using O59 and analysed with PROCHECK.48 VOIDOOG59 was used to detect cavities within the structure, using a 1.4 Å rolling probe. The cavity surface and volume were calculated based on the probe-occupied region.

In an attempt to identify the small molecule observed within the cavity electron density, a variety of likely candidates were built, and the coordinates refined. The quality of the resulting protein–ligand models was assessed by inspection of the electron density maps, and by monitoring changes in the crystallographic and free R-factors. The models trialled contained pentanal, pentanol, pentanoic acid, hexanoic acid, heptanoic acid, m-propylphenol, p-amino benzoic acid and 1-carboxy-4-amino-cyclohexane. Crystallographic and free R-factors were impacted by no more than 0.004 and 0.002, respectively, with inclusion of these ligands.

**Protein Data Bank accession code**

Coordinates of the Bab32a structure were deposited in the RCSB Protein Data Bank as 1TUH.

**Acknowledgements**

We thank Mr Paul Worden (Macquarie University) for DNA sequence determination, and Drs Daniel Jardine (now of Flinders University) and Robert Willows for assistance with mass spectrometry. M.G., A.H., K.N., H.S. and B.M. acknowledge Macquarie University funding through its Innovation Program. Work at the Australian National University was supported by grants from the Australian Research Council, including a Federation Fellowship (to G.O.). A.R. is supported by a Macquarie University Postgraduate (RAACE) Scholarship, and P.W. by an Australian Postgraduate Research Award. Access to the NMR facility at the Australian National University is gratefully acknowledged.

**References**

24. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T,
Piotto, M., Saudek, V. & Sklenar, V. (1992). Gradient-
Studier, F. W., Rosenberg, A. H., Dunn, J. J. &
Neylon, C., Brown, S. E., Kralicek, A. V., Miles, C. S.,
Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, G. J.
Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J.,
Yeates, C. (1999). Microbial diversity and the degra-

Folded Proteins from the Mobile Metagenome

protein secondary structure from circular dichroism
tailored excitation for single-quantum NMR spectro-
studies of mutant proteins by surface plasmon
replication terminator protein (Tus)

Edited by J. Karn

(Received 20 September 2004; received in revised form 12 December 2004; accepted 15 December 2004)