

Research article

Open Access

## Combined protein construct and synthetic gene engineering for heterologous protein expression and crystallization using Gene Composer

Amy Raymond<sup>1,2</sup>, Scott Lovell<sup>1,2</sup>, Don Lorimer<sup>1,2,3</sup>, John Walchli<sup>3,4</sup>, Mark Mixon<sup>3,4</sup>, Ellen Wallace<sup>1,3</sup>, Kaitlin Thompkins<sup>1,2</sup>, Kimberly Archer<sup>1,2</sup>, Alex Burgin<sup>1</sup> and Lance Stewart\*<sup>1,2,3</sup>

Address: <sup>1</sup>deCODE biostructures Inc, 7869 NE Day Road West, Bainbridge Island, WA 98110, USA, <sup>2</sup>Seattle Structural Genomics, Center for Infectious Disease, Bainbridge Island, WA 98110, USA, <sup>3</sup>Accelerated Technologies, Center for Gene to 3D Structure, Bainbridge Island, WA 98110, USA and <sup>4</sup>Emerald BioSystems Inc, 7869 NE Day Road West, Bainbridge Island, WA 98110, USA

Email: Amy Raymond - araymond@decode.com; Scott Lovell - slovell@decode.com; Don Lorimer - dlorimer@decode.com; John Walchli - jwalchli@decode.com; Mark Mixon - mmixon@decode.com; Ellen Wallace - ewallace@decode.com; Kaitlin Thompkins - kthompkins@decode.com; Kimberly Archer - karcher@decode.com; Alex Burgin - aburgin@decode.com; Lance Stewart\* - lstewart@decode.com

\* Corresponding author

Published: 21 April 2009

Received: 15 October 2008

BMC Biotechnology 2009, 9:37 doi:10.1186/1472-6750-9-37

Accepted: 21 April 2009

This article is available from: <http://www.biomedcentral.com/1472-6750/9/37>

© 2009 Raymond et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** With the goal of improving yield and success rates of heterologous protein production for structural studies we have developed the database and algorithm software package Gene Composer. This freely available electronic tool facilitates the information-rich design of protein constructs and their engineered synthetic gene sequences, as detailed in the accompanying manuscript.

**Results:** In this report, we compare heterologous protein expression levels from native sequences to that of codon engineered synthetic gene constructs designed by Gene Composer. A test set of proteins including a human kinase (P38 $\alpha$ ), viral polymerase (HCV NS5B), and bacterial structural protein (FtsZ) were expressed in both *E. coli* and a cell-free wheat germ translation system. We also compare the protein expression levels in *E. coli* for a set of 11 different proteins with greatly varied G:C content and codon bias.

**Conclusion:** The results consistently demonstrate that protein yields from codon engineered Gene Composer designs are as good as or better than those achieved from the synonymous native genes. Moreover, structure guided N- and C-terminal deletion constructs designed with the aid of Gene Composer can lead to greater success in gene to structure work as exemplified by the X-ray crystallographic structure determination of FtsZ from *Bacillus subtilis*. These results validate the Gene Composer algorithms, and suggest that using a combination of synthetic gene and protein construct engineering tools can improve the economics of gene to structure research.

## Background

The gene to structure endeavor faces intrinsic challenges at several steps in the experimental pathway. The primary challenge is expressing the great quantities of protein required to support structural studies, and this challenge can be magnified by the use and necessity of heterologous expression systems. While many research pursuits begin with protein purification, the extent of material required is a burden somewhat unique to protein crystallography. A second major problem of protein crystallography is identifying a suitable protein construct, given that small changes in the protein can have profound effects on crystallization [1,2]. These two problems are related since small changes in the protein construct can also have profound effects on the level of protein expression, solubility, and diffraction quality of protein crystals obtained. One recent report notes that construct engineering can double the number of targets expressing soluble protein in a heterologous system, and can provide a fourfold increase in the likelihood of obtaining well-diffracting crystals [2]. These aggregate improvements can enable extensive crystallization trials.

It is well known that codon utilization is highly biased and varies considerably among various organisms and codon usage is considered a key determinant of eventual heterologous protein expression [3]. Codon bias in genomes can be defined as the unequal usage of synonymous codons in known or predicted open reading frames (ORFs). Codon-usage patterns are related to the relative abundance of tRNA isoacceptors, and genes encoding highly expressed proteins show differences in their codon usage frequencies [4]. Extensive research by Karlin and Sharp have demonstrated that highly expressed genes in *E. coli* and other bacteria have a significant bias towards certain subsets of codons [5,6]. Moreover, bacteria and unicellular eukaryotic organisms seem to have codon biases that are highly correlated with measured isoaccepting tRNA levels [7-9]. In general, where it has been measured, higher eukaryotes also appear to have codon biases that complement the abundance of isoaccepting tRNAs [10,11].

It is generally thought that codon usage alters peptide elongation rates, however codon-usage patterns can also improve the fidelity and kinetic efficiency of translation [12]. Since the level of isoaccepting tRNAs in *E. coli* are correlated with codon bias, it is likely that the presence of rare codons has a negative impact on expression of heterologous genes. It is also likely that the kinetics of translation are improved by more closely matching the codon usage of a recombinant gene product to the usage of the expression host [3,13-17]. Apart from the nonrandom use of codons, it has also been noted that codon/anticodon recognition is influenced by sequences outside the codon

itself, a phenomenon termed codon context [18-20]. There is, for example, an occurrence bias between specific adjacent codon pairs and these biases are different for highly expressed versus low expressed proteins in *E. coli* [21-23]. Clearly, there are still a number of mysteries in the subject of codon bias and codon context. Taken together these results also emphasize how difficult it is to create "codon optimized" genes for expressing proteins since the rules are complex and not yet fully defined.

As noted recently, more empirical evidence about protein expression from engineered genes is still needed, and efforts are being made to database the outcomes of gene engineering [24,25]. There remain a variety of other factors that should be considered when engineering synthetic genes and the details of "optimal" gene design continue to evolve [26]. For example, the synthetic gene sequence should not only contain the proper codon bias and codon context, but should also not result in mRNA secondary structures that may inhibit translation [27]. Additional considerations such as unusual sequence repeats, and cryptic regulatory sequences (*e.g.* Shine-Dalgarno, splice sites, RNase cleavage sites, *etc.*) can impact the fidelity and yield of protein expression [28].

Large scale projects in genomic sequencing and protein structure determination are producing enormous quantities of data on the relationships between 2D gene sequence and 3D protein structure. Moreover, such efforts are providing experimental data on success factors at every step in the gene to structure research endeavor. This wealth of information should be used in a feedback cycle to facilitate the design and production of genes and protein constructs that are engineered for the successful production of functional protein samples for structural studies. Fundamentally, this goal represents a bioinformatics software challenge.

To better address this issue in the context of synthetic gene design, we have created a Protein Construct Design interface for Gene Composer which distills protein structure information from PDB files and comparative sequence information into a graphical user interface that allows the user to simultaneously analyze known protein structures, together with homologous target protein sequences for which derivative constructs can be designed. This interface allows the user to simultaneously understand sequence conservation, ligand contacts, and known or predicted structural elements to define candidate protein constructs for crystallization and then design harmonized nucleic acid sequences to express those proteins.

To demonstrate the utility of this approach, we have used Gene Composer to design protein constructs, as well as the synthetic gene sequences to express those proteins. We

compared the resulting protein expression level to full length proteins expressed from native genes. We initially chose to examine proteins from three different sources (human, viral and eubacterial) in two different heterologous expression systems (*E. coli* and wheat germ *in vitro* cell free expression). The results show that Gene Composer designed proteins and their corresponding engineered genes can be expressed at significantly higher levels than native genes. In addition, the results show that using Gene Composer to design protein constructs can significantly improve the likelihood of obtaining protein crystals.

## Results and discussion

### Construct Design Using Gene Composer™ Software

In order to compare protein expression from engineered genes vs. native genes we chose to examine genes with significantly different G:C content and codon usage bias. We examined three globular proteins from three phylogenetically disparate sources including: FtsZ, the cell division protein from *B. subtilis*; NS5B, the RNA-dependent RNA polymerase from Hepatitis C Virus; and P38 $\alpha$ , a human protein kinase. Using Gene Composer software, we made an alignment of each protein with published structural information, as well as the sequences of several homologous proteins for which structural information is not available. These alignments provided insights into the conservation of amino acid sequences and the known structure of homologous proteins that can be combined to engineer superior protein constructs for expression testing and crystallization trials. As an example, The Protein Alignment viewer of FtsZ (Figure 1) shows that the first eleven N-terminal amino acids of *B. subtilis* FtsZ are not conserved among several closely related species (*Clostridium sp.*, *Listeria sp.*, *Staphylococcus sp.*, and *Streptomyces sp.*). In addition, the homologous N-terminal region of *Aquifex* FtsZ (PDB ID 2R6R) is present but disordered in the crystal lattice (compare 2R6R chain sequence to 2R6R model sequence). Together, this information suggests that this N-terminal region is not important for the function and structure of *B. subtilis* FtsZ, and that removal of this region could improve protein expression. A similar analysis of the C-terminal region of FtsZ showed that a C-terminal region (Glu317-Gly382 in *B. subtilis* FtsZ) was not conserved and not present in several other constructs used to generate FtsZ crystals (including 2R6R). Based on this analysis we defined a truncated form of *B. subtilis* FtsZ spanning residues Ala12-Ile316 (termed "FtsZ truncated").

A similar analysis was performed with NS5B and showed that C-terminal residues were not conserved and not visible within the electron density maps of several crystal structures. From this analysis, we identified a C-terminal truncation spanning residues 1–568 for expression test-

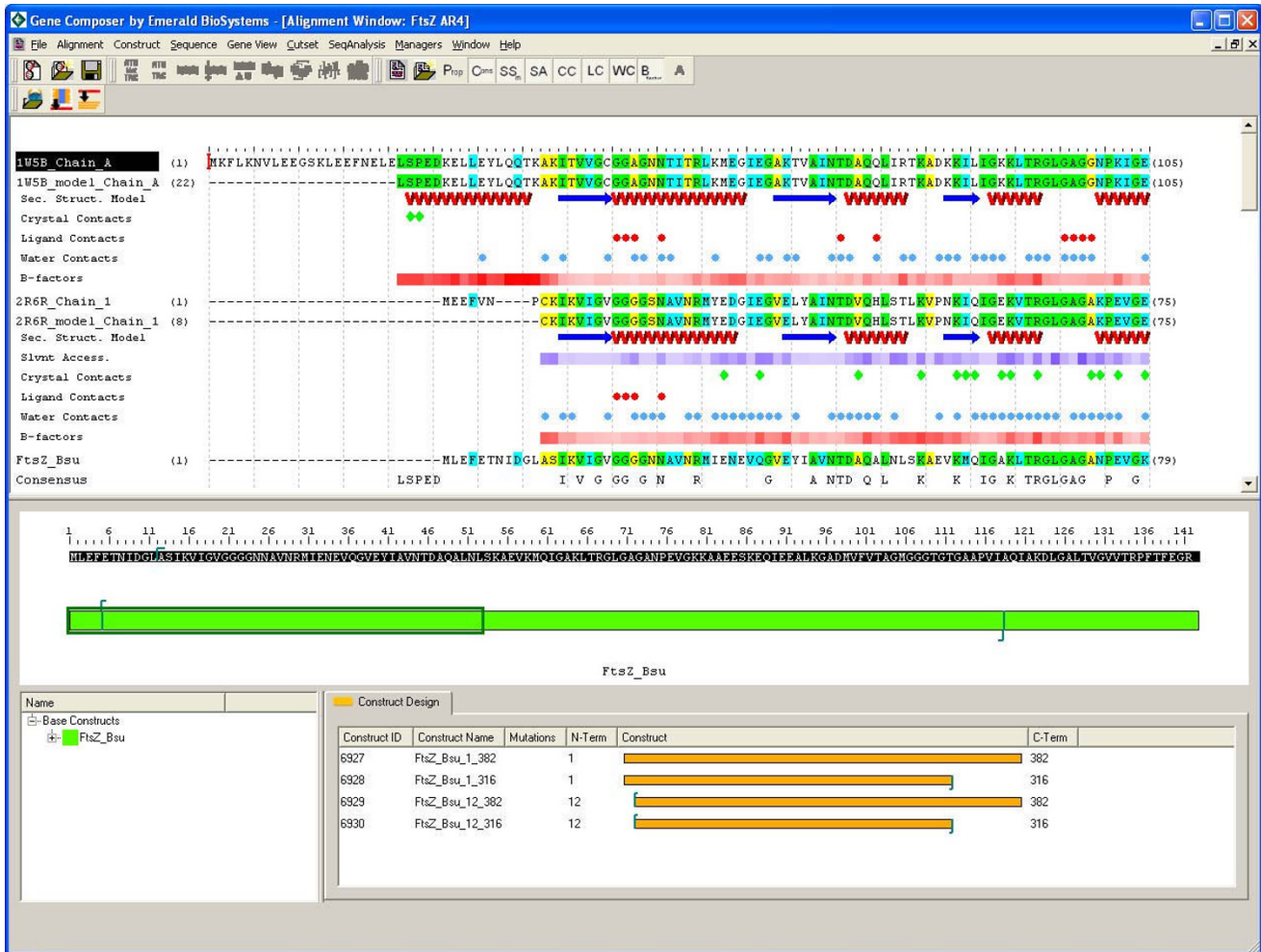
ing. Finally, an analysis of P38 $\alpha$  showed that residues across the entire protein were relatively highly conserved and almost all residues could be visualized in published crystal structures. Based on these results, we did not create a truncated form and only attempted expression of the full-length protein.

At this point, each target was moved into the Gene Design module (Protein-to-DNA) of Gene Composer™. For engineered genes, the nucleic acid sequence was determined by back-translation from the amino acid sequence. As previously described [29], multiple factors were incorporated in the final design including *E. coli* codon usage, overall G:C content, potential mRNA secondary structures, presence of out-of-frame stop codons, removal of cryptic Shine-Dalgarno sequences, removal of repeated strings of the same nucleic acid sequence, as well as the silent introduction and removal of select restriction sites to facilitate cloning. For native genes, the wild-type sequences were input with no sequence alteration. The resulting full length engineered genes and native genes express identical proteins, and were cloned into identical expression vectors.

### Protein Expression in *E. coli*

In order to determine if the use of phylogenetic and structural information to design modified protein constructs and engineering of the nucleic acid sequence resulted in improved protein expression, we compared the level of protein expressed from the engineered synthetic genes to proteins (constructs listed in Table 1) expressed from native sequences under identical conditions. It has been well demonstrated that expression conditions (media, temperature, inducer concentration, length of induction, etc.) can have a profound effect of the level of protein expression and it is certainly possible that some or all of these variables could be optimized to alter the relative expression levels between the engineered proteins; however we first chose to compare the expression under a single condition. This condition was based on an analysis of expression conditions used by multiple high throughput protein structure initiative centers and we believe a typical representative of a common "first pass" expression condition (0.4 mM IPTG inducer at mid-log phase at 30°C for four hours in 2YT media).

The results in Figure 2 show that the amount of soluble protein produced by the Gene Composer engineered genes for all three targets are either equal to or greatly improved over the level of protein expressed from the native gene. This improvement was most dramatic in the case of FtsZ where gene engineering resulted in a 12-fold increase in the amount of soluble protein from the same weight of cell paste (full-length native vs. full-length engineered). The deletion of N-terminal and C-terminal resi-



**Figure 1**  
**Gene Composer software.** Comparative sequence alignments and corresponding structural information are organized by the Alignment Viewer of Gene Composer (top portion of panel). Amino acid sequence of *Aquifex aeolicus* FtsZ used to obtain FtsZ crystal structure (2R6R) is shown; Chain sequence indicates amino acid residues present in the protein constructs and Model sequence indicates amino acid residues that could be modeled from the electron density maps. Protein secondary structural regions are represented by red helices for alpha helical regions, blue arrows for beta-sheet regions, and green squiggles for turns, as defined the respective PDB file format. Thermal B-factors from the PDB file are normalized and represented in red shading levels (light is low and dark is high). The crystal, ligand, and water contacts are all extracted from the PDB file and listed pictorially. The *bacillus subtilis* FtsZ sequence (target) is shown on the bottom line and is aligned with FtsZ sequences from multiple species (accession codes indicated). Conserved residues are highlighted (green = high; blue = medium; yellow = low), and a consensus sequence is listed in the bottom row of the alignment.

dues did not significantly improve expression relative to the full length FtsZ protein (15-fold increase relative to expression from full-length native gene). The expression of P38α from the synthetic gene was also significantly improved (6-fold) relative to expression from the wild-type P38α gene. It is important to emphasize that synthetic gene engineering using Gene Composer does not always result in improved protein expression, and there was no clear improvement of expression of NS5B compared to native NS5B. We also compared the expression of

four different human GPCR membrane proteins from engineered and native genes: human adenosine A2a receptor (Adora2a), human adrenergic, beta-2 receptor (b2AR), human tachykinin receptor 1 isoform long (NK1), human Galanin receptor type 1 (GalRI). These GPCRs were expressed as N-terminal Mystic fusions [30,31], and in all cases the level of expression from the engineered gene was the same or slightly higher when compared to expression from the native gene (data not shown).

**Table 1: Protein constructs used in this test set.**

| Target                        | Construct   | aa     | MW (kDa) |
|-------------------------------|-------------|--------|----------|
| <b>FtsZ</b>                   | Full Length | 1–382  | 42.8     |
|                               | truncated   | 12–316 | 34       |
| <b>NS5B</b>                   | Full Length | 1–589  | 66.4     |
|                               | truncated   | 1–568  | 64       |
| <b>P38<math>\alpha</math></b> | Full Length | 1–370  | 43.5     |

### Crystallography and Structure Determinations

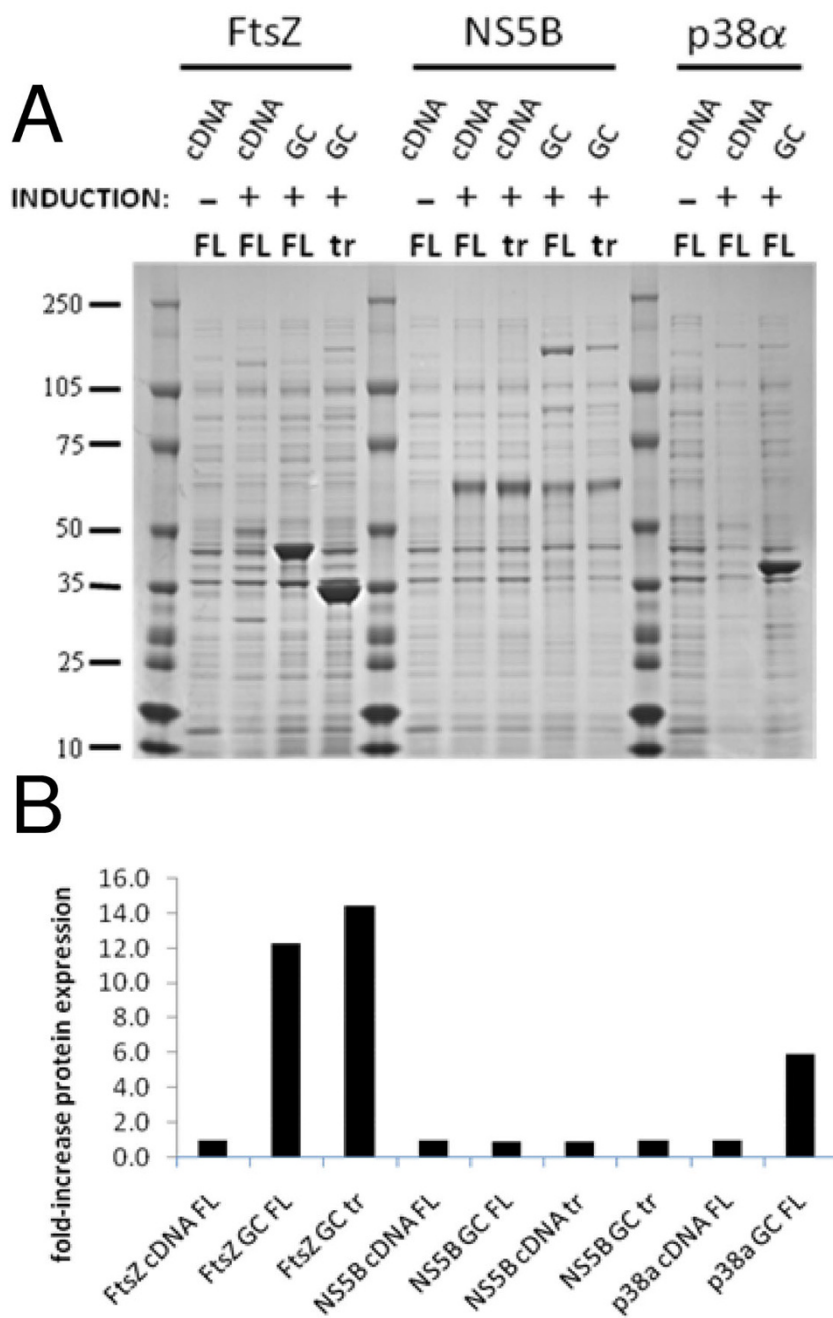
In order to determine if proteins expressed from the synthetic genes could be used successfully in protein crystallization experiments, we scaled up the expression and purified full length and truncated forms of FtsZ. The purified proteins were concentrated and used in a series of random sparse matrix crystallization screens. For the full-length FtsZ protein, we established 768 random sparse matrix sitting drop crystallization trials, but did not observe any crystallization hits after monitoring trials for one month. 192 different random sparse matrix crystallization trials were established with the truncated construct (FtsZ truncated) and 32 different crystallization hits were observed following 1 week of monitoring. Representative examples of these crystal hits are shown in Figure 3. Based on the quality of crystal morphology and initial diffraction screening experiments, we focused on a single condition (30% PEG-200, 100 mM MES pH 6.0, 5% PEG-3000) and harvested crystals that diffracted X-rays to 2.45 Å resolution. The structure [PDB code: [2RHLL](#)] was solved by molecular replacement as described in the Materials and Methods.

FtsZ formed orthorhombic crystals and contained two molecules in the asymmetric unit related by a non-crystallographic 2-fold axis (Fig. 4A). Non-bonded contacts at the NCS dimer interface are formed between the backbone oxygen of Lys 64 and the side chain of Asn 74 as well as Glu 76 and Glu 84. The latter are formed via water mediated contacts. The structure was traced from Ala 12 to Phe 315 of FtsZ for both subunits. Although guanosine diphosphate (GDP) was not added during the purification or crystallization, analysis of the nucleotide binding pocket following refinement revealed prominent difference density (Fig. 4B) consistent with the presence of GDP. The presence of GDP in the active site of FtsZ has been observed before [32,33] and suggests that the crystal-

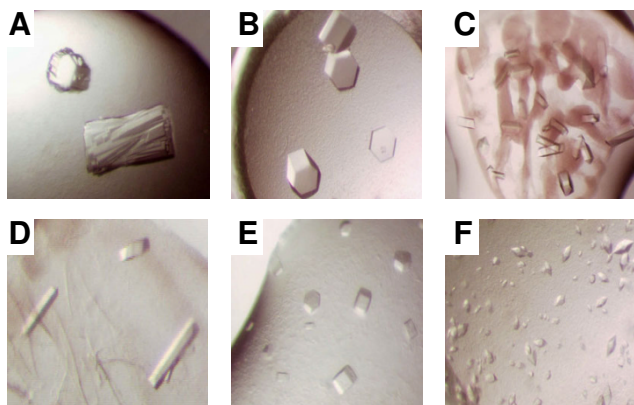
lized protein is functional. GDP forms direct non-bonded contacts with the T1 loop (Gly 21-Gly 22) and T4 loops (Gly 108-Gly 110) (Fig. 4C). The latter corresponds to the tubulin signature motif (GGGTGTG) which is involved in phosphate binding [34]. These domains are in similar orientations to domains previously described [35,36].

It has been shown previously that nucleotide exchange occurs readily and has negligible conformational effects in certain FtsZ structures[33]. To determine if this was the case for the *B. subtilis* FtsZ structure, crystals of FtsZ:GDP were soaked in the presence of GTP- $\gamma$ -S. The crystals (FtsZ:GSP) tolerated the soaks well and diffracted to the same resolution (2.45) as FtsZ:GDP crystals. Similar to FtsZ:GDP, this structure [PDB code: [2RHQO](#)] was traced from Ala12 to Phe315. However, the residual electron density at the C-terminus of subunit B was weak and the purification tag could only be traced to Leu316. Also, residues Lys221 and Gly222 on chain B were disordered and were not modeled. Examination of the difference electron density revealed that subunit A contained GTP- $\gamma$ -S as there was clear density for the  $\gamma$ -phosphate. However, the GDP in subunit B was not exchanged as evident by difference density which was consistent with GDP. Interactions with GTP- $\gamma$ -S and FtsZ are similar to those observed in the FtsZ:GDP structure except that the  $\gamma$ -phosphate forms contacts with Ala71 and Ala73 of the T3 loop.

Two additional crystal forms were obtained in the presence of lithium sulfate. A primitive tetragonal form, FtsZ:SO<sub>4</sub>, was found to contain a single sulfate bound in the nucleotide pocket [PDB code: [2RHH](#)]. This sulfate ion occupies position similar to that of the  $\beta$ -phosphate of GDP and GTP- $\gamma$ -S observed for FtsZ:GDP and FtsZ:GSP. Not surprisingly, the sulfate ion forms non-bonded contacts with residues in the T1 and T4 loops. The sulfate ion is positioned in a nearly identical manner as that observed in a recently published full length *Bacillus* FtsZ structure [37]. A C-centered orthorhombic crystal form (FtsZ:2SO<sub>4</sub>) was found to contain two sulfate ions in the nucleotide binding pocket [PDB code: [2RHI](#)]. Additionally, prominent difference density was observed between the two sulfate ions which we clearly observed at a 6 $\sigma$  contour level. Initially, a water molecule was placed at this site however this water was surrounded by a large amount of positive difference electron density following refinement. The unidentified atom appeared to be coordinated by two water molecules and two oxygen atoms of the sulfates in a distorted square planar arrangement. We considered the possibility that this was a magnesium ion. However, although four coordinate magnesium is a possibility it typically forms six coordinate interactions[38]. We then used the program Wasp[39] to examine the solvent structure for potential metal ions. The water positioned at this site of positive difference density was flagged as a possible



**Figure 2**  
**Expression of full length and truncated native and Gene Composer engineered genes in *E. coli*.** **A**, Total soluble protein was resolved by SDS-PAGE and the resulting comassie blue stained gel is shown. Lanes containing *B. subtilis* FtsZ (full length: 42.8 kDa, truncated: 34.0 kDa), Hepatitis C Virus NS5B (full length: 66.4 kDa, truncated: 64.0 kDa); and human P38α (full length: 43.5 kDa) are indicated. Lanes from Gene Composer engineered genes are indicated (Eng) and compared to native genes (Native). Expression of full length (FL) and truncated forms (tr) proteins is also compared. Negative control experiments showing expression in the absence of inducer (-) is also shown. **B**, For each sample, the amount of protein observed was divided by the amount of full length native protein and plotted. By definition, FtsZ native FL, NS5B native FL, and P38α native samples have values of 1 and are shown as controls.



**Figure 3**  
**Crystallization of FtsZ.** Protein crystals of FtsZ obtained under different conditions are shown: **A**, 1.26 M  $(\text{NH}_4)_2\text{SO}_4$ , 100 mM MES pH 6.5. **B**, 1.26 M  $(\text{NH}_4)_2\text{SO}_4$ , 100 mM Hepes pH 7.5. **C**, 30% (v/v) PEG-200, 100 mM MES pH 6.0, 5% (w/v) PEG-3000. **D**, 20% (w/v) PEG-8000, 100 mM Hepes pH 7.5. **E**, 1.26 M  $(\text{NH}_4)_2\text{SO}_4$ , 100 mM Hepes pH 7.5. **F**, 50% (v/v) PEG-200, 100 mM Tris pH 7.0, 0.05 M  $\text{Li}_2\text{SO}_4$ .

sodium ion. Therefore, a sodium ion was subsequently positioned in the difference density and refined. The refined metal-oxygen non-bonded distances and coordination geometry were consistent with those expected for a sodium ion[40]. In this structure, the sulfate ions adopt similar positions as the  $\alpha$  and  $\gamma$  phosphates of GTP- $\gamma$ -S found for FtsZ:GSP. Similarly, non-bonded contacts are formed between the T1, T3, and T4 loops of FtsZ and the sulfate ions.

These results clearly show that proteins expressed from a synthetic Gene Composer engineered gene can be successfully used in protein crystallization experiments. More importantly, the results also show that protein construct engineering using structural and phylogenetic information can be critical for successful crystallization of a protein target since only the truncated form of FtsZ yielded protein crystals.

#### Protein Expression in *E. Coli* Under High-Throughput Conditions

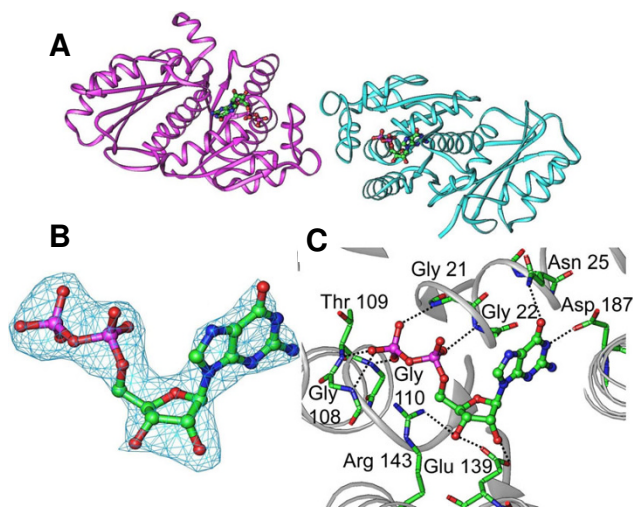
In order to begin to understand how often engineering of the nucleic acid sequence will improve protein expression, we randomly chose eleven different proteins from three different bacterial species on the National Institute of Allergy and Infectious Disease priority pathogens list, and compared protein expression from full length wild type genes and full length Gene Composer engineered genes. The proteins were expressed from identical expression vectors, were amended with identical purification tags, and the final protein product pairs had identical

amino acid sequences. All of the expressed proteins contained 6 $\times$  His tags to facilitate quantitation of soluble protein expression. In these examples the proteins were expressed using autoinduction media [41].

The results in Figure 5 show that the level of soluble protein expression is nearly identical for the randomly selected eleven pairs of constructs (see Additional file 1). As a control, the engineered and native FtsZ genes that had previously demonstrated significant differences in the level of protein expression under standard expression testing (see Figure 2) were included in the experiment. A similar result was obtained with FtsZ (see pair 12) with the engineered gene expressing at significantly higher levels than the wild-type gene, however there was only modest differences between the level of expression from the engineered gene vs. the native genes for the randomly selected constructs. This result demonstrates that redesign of the nucleic acid sequence does not always result in improved levels of protein expression; however it also demonstrates that gene engineering does not lead to decreased levels of protein production.

#### Protein Expression in CFS

The improved expression of FtsZ and P38 $\alpha$  from synthetic genes in *E. coli* may result from the biased codon usage of the native genes leading to abortive heterologous protein expression when rare tRNA pools become depleted *in vivo*. It is difficult to test this possibility *in vivo* because of the many factors that could potentially modulate levels of aminoacylated tRNA pools; however it is possible to control tRNA pools *in vitro*. To this end, we expressed several native and engineered gene sequences (optimized for *E. coli* expression) in a wheat germ extract-based Cell Free System (CFS, <http://www.cfsciences.com>). The cell free system reagents include an excess of molecular translation machinery, resulting in a protein expression strategy that is not limited by depletion of tRNA pools. We cloned the same engineered and native FtsZ, NS5B and P38 $\alpha$  genes into a CFS expression vector (Materials and Methods) ensuring that the full length proteins had identical amino acid sequences. The proteins were expressed under identical conditions and all proteins present in the final expression extract (soluble and insoluble) were resolved by SDS PAGE. The resulting Commassie blue stained gel is shown in Figure 6A. The results show that because the recombinant expressed proteins FtsZ, NS5B and P38 $\alpha$  only account for a small percentage of the total proteins present in the CFS extracts, an accurate comparison of expression levels of the proteins of interest cannot be made from the SDS-PAGE analysis of the total CFS extracts. Only the NS5B protein bands of  $\sim$  60 KDa in lanes 3 and 4 of Figure 6B could be reliably discerned above the background proteins, and it appears that the



**Figure 4**  
**Structure of *Bacillus subtilis* FtsZ.** **A**, Ribbon diagram of the non-crystallographic dimers viewed approximately parallel to the NCS 2-fold axis is shown. Bound GDP within the active site is represented in stick drawing. **B**, The  $F_0-F_c$  omit electron density map contoured at  $3\sigma$  is shown in blue. Stick representation of bound GDP is modeled in the omit density. **C**, Hydrogen bond interactions between FtsZ active site residues (stick) and GDP (ball and sticks) are shown as black dashed lines. Amino acids residues are numbered and the FtsZ backbone is shown in grey.

native gene expressed NS5B protein in lane 3 is made in slightly larger amounts than its counterpart in lane 4 produced by the synthetic NS5B gene.

In order to more accurately compare the protein expression levels, the soluble 6X-His tagged recombinant proteins were purified from the CFS extracts by nickel chelate bead capture and resolved SDS PAGE with Coomassie blue staining as shown in Figure 6B. The NS5B proteins were not captured by nickel chelate and were shown to be located entirely in the insoluble fraction of the expression extract (data not shown). Quantification of the full length purified FtsZ proteins shows an approximate 2-fold increase in protein expressed from the engineered gene relative to the native gene, while expression of full length P38 $\alpha$  from both sources is equal. Thus, the difference in protein expression levels from synthetic engineered P38 $\alpha$  and FtsZ genes relative their native counterparts is modest in the CFS experiment, and do not recapitulate the outcomes observed when these genes were expressed in *E. coli* cells (see Figure 2). Taken together, these results suggest that the rare codons present in the native gene sequences of human P38 $\alpha$  and *B. subtilis* FtsZ can significantly hamper protein expression of these proteins in *E. coli*, but this can be overcome either by codon engineering for *E. coli* or

by providing an excess of molecular translation machinery in the wheat germ CFS system. Moreover, the synthetic codon engineered versions of these genes do not appear to offer any significant advantage for protein production in the CFS wheat germ extract.

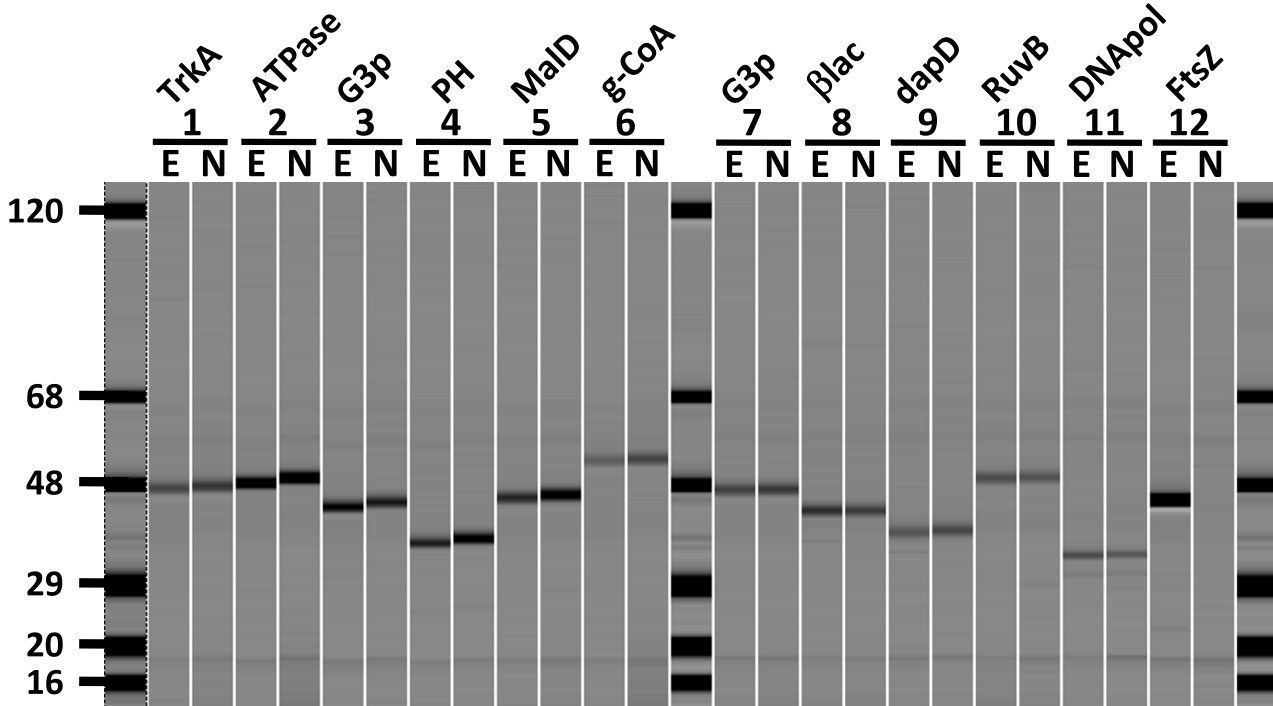
In order to explore the possibility that codon engineering of synthetic genes for high level *E. coli* expression could affect protein production in a wheat germ extract, we compared the CFS expression of synonymous pairs of genes from *Arabidopsis thaliana* whose codon usage is comparable to that of wheat, wherein the native *A. thaliana* gene is compared to that of a synthetic codon engineered gene designed with codon usage of wheat genes. In all cases, the level of CFS expression from the wheat codon engineered gene was the same or slightly higher than expression from the native *Arabidopsis* gene (data not shown). This data is consistent with the suggestion that translation machinery is not limiting for protein production in the CFS wheat germ extract.

## Conclusion

Obtaining sufficient quantities of soluble protein for structural studies often requires expressing proteins in heterologous expression systems. Unfortunately, the native nucleic sequence of any gene is likely to be biased from the evolution of that gene for expression in the native host, and in some cases, the gene may have selective pressures to be naturally expressed at very low levels. This bias may affect codon usage, codon context, mRNA secondary structures, regulatory elements, *etc.* which will likely impact the level of expression in a heterologous system. In order to overcome this fundamental problem, we have developed Gene Composer software to engineer any nucleic acid sequence for any given expression system (sister manuscript). In this report, we have systematically compared protein expression from native genes to protein expression from genes specifically engineered for high level expression in *E. coli*.

The majority of genes tested did not result in improved levels of protein expression. There are several possible explanations for this result. First, the yield of soluble protein may be dominated by the balance of protein turnover. If a protein is extremely unstable, increases in the rate of translation or fidelity of translation will have only small effects on the steady concentration of that protein in the cell. Future pulse-chase experiments designed to access rates of protein translation are necessary to address this possibility. Another possibility is that the results reflect our lack of understanding of the many factors that are important for protein expression and the likelihood that the "rules" are different for individual genes and proteins. For example, the engineered synthetic genes were harmonized for codon usage, absence of sequence





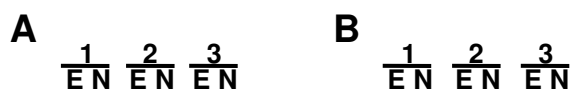
**Figure 5**

**Comparative analysis of protein expression from full length synthetic genes.** Proteins were expressed and purified in parallel from Gene Composer engineered (E) or native (N) genes as described in Materials and Methods (see Additional File 1). Proteins from *Brucella melitensis*, str. biovar Abortus 2308 were compared in pairs: **Pair 1**) Lactate/malate dehydrogenase:L-lactate dehydrogenase:TrkA potassium uptake protein (44.8 kDa); **Pair 2**) CbxX/CfqX superfamily:Disease resistance protein:ATP/GTP-binding site motif A (P-loop):AAA ATPase:AAA ATPase, central region (49.4 kDa); **Pair 3**) Glyceraldehyde 3-phosphate dehydrogenase:TrkA potassium uptake protein:Glyceraldehyde-3-phosphate dehydrogenase, type I (47.4 kDa); **Pair 4**) 3' exoribonuclease:Ribonuclease PH (37.0 kDa). *Burkholderia pseudomallei*, str. 1710b: **Pair 5**) malate dehydrogenase (46.2 kDa); **Pair 6**) glutaryl-CoA dehydrogenase (54.3 kDa); **Pair 7**) glyceraldehyde-3-phosphate dehydrogenase, type I (47.3 kDa); **Pair 8**) beta-lactamase (44.0 kDa); *Rickettsia prowazekii*, str. Madrid E: **Pair 9**) 2,3,4,5-Tetrahydropyridine-2-Carboxylate N-Succinyltransferase (dapD) (41.3 kDa); **Pair 10**) Holliday junction DNA helicase RuvB (49.6 kDa); **Pair 11**) DNA polymerase III subunit epsilon (37.2 kDa). *Bacillus subtilis*, subsp. Subtilis, str. 168; **Pair 12**) FtsZ (full length, 42.8 kDa). Molecular weight markers are present in lanes 1, 14, 27.

repeats, absence of repeating codons within a sequence, reduction of local mRNA secondary structures within the transcript, and removal of cryptic Shine Dalgarno sequences; however, other factors may also be important for expression. It is not possible to test these unknown factors at this point; but synthetic gene design offers the ability to incorporate and test new gene design features as they become known.

Because we are still in the early days of discovering all of the rules that govern gene optimization, it is important to determine if Gene Composer designed genes reduce the level of protein expression compared to native genes. Engineered genes did not result in a decrease in the level of soluble protein from the 14 different genes tested, and

in two cases, gene engineering resulted in a significant increase in soluble expression (FtsZ from *Bacillus subtilis* and human P38 $\alpha$ ). The underlying cause for the dramatic increased expression of FtsZ from the engineered synthetic gene is not immediately clear and likely results from a combination of factors. For example, comparison of expression of FtsZ in *E. coli* vs. cell free expression may suggest that codon usage resulted in a decrease in expression from the native gene since both engineered and native FtsZ genes expressed at relatively high levels in this heterologous system. The native FtsZ gene contained a total of 19 rare codons (1.58% of all codons) which were eliminated in the engineered gene (see Table 2), however NS5B contained a significantly larger number of rare codons (62 or 3.46% of all codons) but elimination of



**Figure 6**  
**Expression of full length native and Gene Composer engineered genes in Cell Free System, SDS-PAGE analyses.** **A**, Total protein from CFS expression of *B. subtilis* FtsZ (42.8 kDa), Hepatitis C Virus NS5B (66.4 kDa); and human P38 $\alpha$  (43.5 kDa) were resolved by SDS PAGE and the resulting coomassie blue stained gel is shown. Odd lanes contain expression from native genes, even lanes contain expression from engineered genes. **B**, Soluble proteins from the total expression were purified by Ni affinity chromatography and resolved by SDS PAGE. Odd lanes contain expression from native genes, even lanes contain expression from engineered genes. Data shown is a representative from three independent experiments.

rare codons from NS5B had no effect on the level of expression. The total number of rare codons may not be important if the location of rare codons relative to individual folding protein domains or mRNA secondary structure elements could also affect protein expression. There are clearly a large number of different factors that could be the cause of the significant increase in FtsZ expression.

The ability to express significant quantities of FtsZ enabled the crystallization of this target and demonstrates the utility of Gene Composer for protein design and construct engineering for structural studies. Specifically, we used Gene Composer to define structure-guided protein constructs for crystallization studies and showed that we were unable to obtain protein crystals of full length FtsZ, but did obtain multiple crystal hits from the truncated, engineered FtsZ construct (see Figures 1 and 3).

In conclusion, these results demonstrate the utility Gene Composer for the design of protein constructs and synthetic genes to express those proteins. The use of protein engineering to improve crystallization can have a significant impact on the gene to crystal structure process since very complex protein construct engineering principles

(fusions, rearrangements, rationale surface variants, etc.) can be rapidly implemented without multiple molecular biology steps that can require significant amount of time and effort. Because engineering of the nucleic acid sequence did not decrease the level of protein expression and in some cases dramatically improved expression, synthetic gene design can also significantly improve the gene to crystal structure process. The utility of Gene Composer for engineering nucleic acid sequences will improve as more design principles are identified, especially as the cost of gene synthesis continues to decrease and robust protocols for rapidly synthesizing genes become more available.

## Methods

### Gene Composer engineering

Gene Composer engineering considerations included codon usage frequency minimum threshold of 2%, elimination of undesired restriction sites, elimination of five or more nucleotide repeats, placement of stop codons every 100 bp in second and third reading frames, elimination of cryptic Shine-Delgarno sequences, and optimized  $T_m$  and  $\Delta G$ . Detailed information about engineering with Gene Composer can be found in (Sister Manuscript).

### Cloning

All amplifications were performed in 50  $\mu$ l reactions using KOD Hi-Fi polymerase (Novagen) according to the manufacturer's instructions. Resulting PCR products were resolved on 1% agarose gels in  $1 \times$  TBE + EtBr, and subsequently gel-purified using Qiagen gel extraction purification protocol. Amplification products not requiring gel purification were purified using the microfuge PCR purification kit protocol. DNA yields were assessed with a Beckman Coulter DU 500 spectrophotometer. Ligation reactions were performed with T4 DNA Ligase (NEB), using a 1:3 molar ratio of vector and insert DNA. Chemically competent OneShot TOP10 cells were transformed with 6  $\mu$ l of each ligation using a standard heat shock protocol. Cells were pelleted, resuspended in 50  $\mu$ l, and spread on 10 cm 2XYT plates containing the appropriate antibiotic. Following 16 hour incubation at 37°C, transformant colonies were screened by PCR and positive clones were sequence verified.

### Bacterial expression of P38 $\alpha$ , NS5B, and FtsZ

Gene Composer engineered and native full length FtsZ genes were purchased from DNA2.0, and subcloned into pET28 using by AflIII/BamHI digestion. The engineered genes were optimized for expression in *E. coli* using *E. coli* codon usage tables and design algorithms (Sister Manuscript). For the eight other constructs in this test set, sequence verified, pCRBluntII Topo clones were amplified by PCR. Truncated FtsZ native and Gene Composer engineered inserts were amplified with primers to incorporate

a 5' NcoI site before the first Met ATG, and adding two stops with a 3' BamHI site following the affinity tags. Native P38 $\alpha$  was amplified using primers incorporating a flanking 5' NdeI site such that after subcloning, the 5' ATG would fall in-frame with the vector derived N-terminal tags. The 3' primer added two stop codons and a BamHI site. The resulting PCR product was purified using the Qiagen PCR purification kit, and sequentially digested by first NdeI (Fermentas), ethanol precipitated and digested with BamHI (Fermentas). Due to an internal NdeI site, two digestion fragments were gel-purified: a 5' 238 bp fragment, flanked on both sides by NdeI, and a 3' 857 bp fragment, with 5' NdeI and 3' BamHI termini. A 3-fold molar excess of pooled fragments was combined with NdeI/BamHI-digested pET15b in a standard ligation. Gene Composer engineered P38 $\alpha$  was cloned in the same fashion, except that no internal NdeI site was present. The four NS5B constructs were amplified with primers which add a 5' NcoI site flanking the initiating methionine ATG, followed by a 6 $\times$  His tag. Two stop codons were added after the last amino acid, followed by a 3' BamHI site for subcloning into NcoI/BamHI-digested pET15b. Cloning of full length and truncated wild-type NS5B was complicated by the presence of two internal NcoI restriction sites; desired constructs were the product of multiway ligation into pET15b. All transformant colonies were screened by PCR, and sequence verified.

#### *Cell free expression of P38 $\alpha$ , NS5B, and FtsZ*

Sequence verified whole gene synthesis products of native and Gene Composer engineered versions of P38 $\alpha$  and NS5B were used as PCR templates for cloning into vector pEU-E01. Full length native and Gene Composer engineered FtsZ ORFs were ordered from DNA 2.0, and served as template for cloning of the FtsZ expression constructs. Amplification primers for all six constructs added a 5' SpeI site and 3' XhoI site for cloning into SpeI/XhoI-cut pEU-E01 (CFS). Following spin purification of PCR products, approximately 1  $\mu$ g PCR product was digested with SpeI/XhoI (Roche) and gel purified. Ligations were set up as described previously, and transformation reactions were spread on 2XYT ampicillin 100  $\mu$ g/ml plates.

#### **Expression**

##### *E. coli expression of pET P38 $\alpha$ , NS5B, and FtsZ constructs*

Cultures of 2YT supplemented with antibiotic were inoculated with single colonies of BL21(DE3) cells that had been transformed with each sequence verified pET28 or pET15 construct. Following overnight growth at 37 $^{\circ}$ C, cultures were diluted 100-fold in the same medium to a final volume of 1L, shaken at 235 rpm at 37 $^{\circ}$ C until OD<sub>600</sub> reached approximately 0.8. Cultures were induced with 0.4 mM IPTG and incubated at 30 $^{\circ}$ C for 4 hours. Bacterial cells were harvested by centrifugation and frozen at -20 $^{\circ}$ C. From each expression, 1.0 g of cell paste was

lysed in 20 mM Tris-HCl (pH 8.0), 500 mM NaCl, 20% (v/v) Glycerol, Complete-EDTA free protease inhibitor (Roche), 25 units Benzamide hydrolyase, and 0.1 mg/mL lysozyme and sonicated on a Branson Sonifier using a microtip, on 70% duty cycle, output 7, for 45 seconds each. Equal amounts of total protein, as quantified by A<sub>280</sub> measurement on NanoDrop ND-1000 spectrophotometer were analyzed by SDS-PAGE. Samples were resolved on 4–12% Bis-Tris NuPage 12 well 1 mm gels in 1 $\times$  MES running buffer, using reducing SDS-PAGE sample buffer and heating samples to 95 $^{\circ}$ C 5 mins prior to loading and visualized by Coomassie blue staining, following manufacturer instructions. Protein bands were visualized on a Kodak Imager and quantified with the software ImageQuant 5.2 (Molecular Devices).

##### *E. coli expression of pET constructs of Brucella, Burkholderia, Rickettsia targets and pET FtsZ constructs*

Pre-cultures (1.2 mL/well, in 96-well deep-well, round bottom block) of 2YT supplemented with antibiotic and 0.5% glucose were inoculated with 15% glycerol stocks made from isolated colonies of BL21(DE3) cells that had been transformed with each sequence verified pET28 construct. Pre-cultures were grown at 37 $^{\circ}$ C, 220 rpm for 16 hours. Cultures (1.2 mL/well, in 96-well deep-well, round bottom block) of 2YT supplemented with antibiotic and Autoinduction System 1 (Novagen) were inoculated with 40  $\mu$ L of pre-culture and grown for 40 hours at 20 $^{\circ}$ C, 220 rpm. Cultures were harvested by centrifugation and stored at -20 $^{\circ}$ C for at least 1 hour. His-tagged recombinant proteins were purified with magnetic Ni-NTA beads (Qiagen) according to manufacturer protocols on a BioRobot 3000 (Qiagen). Purified recombinant products were analyzed by capillary electrophoresis on Lab Chip 90 (Caliper) using an HT Protein Expression Assay Kit, according to manufacturer protocols.

##### *Cell Free Systems*

All sequence verified constructs for wheat germ cell free expression (CFS) were purified with an ultra pure plasmid miniprep kit (Marligen) and set to 1  $\mu$ g/ $\mu$ L. Three expression campaigns were carried out on all FtsZ, NS5B, and P38 $\alpha$  constructs. Due to the presence of His tags on all constructs, an H-kit was used. The H-kit contains all components necessary for transcription and translation. All steps were performed according to the manufacturer's instructions for small scale batch mode expression (WEPRO1240H Expression Kit\_E\_ver. 2.0, Aug. 13, 2007) [42,43]. The RNA was quantified (A<sub>260</sub> absorbance measurements) and normalized to provide the same amounts of each mRNA sample to the cell free translation mix.

In brief, mRNA encoding each protein was synthesized by *in vitro* transcription (20  $\mu$ L) of 2  $\mu$ g the pEU plasmid carrying the gene under control of the SP6 RNA polymerase

**Table 2: Summary of rare codon usage in full length native genes.**

| <b>FtsZ FL native</b>                   |              |                   |
|---|--------------|-------------------|
| <b>AA</b>                               | <b>codon</b> | <b>number (%)</b> |
| Arg                                     | AGG          | 0                 |
| Arg                                     | CGA          | 1 (6%)            |
| Arg                                     | CGG          | 0                 |
| Gly                                     | GGA          | 12 (31%)          |
| Gly                                     | GGG          | 3 (8%)            |
| Leu                                     | CTA          | 2 (7%)            |
| Pro                                     | CCC          | 0                 |
| Ser                                     | AGT          | 1 (6%)            |
| <b>NS5B FL native</b>                   |              |                   |
| <b>AA</b>                               | <b>codon</b> | <b>number (%)</b> |
| Arg                                     | AGG          | 8 (21%)           |
| Arg                                     | CGA          | 7 (18%)           |
| Arg                                     | CGG          | 7 (18%)           |
| Gly                                     | GGA          | 6 (17%)           |
| Gly                                     | GGG          | 11 (31%)          |
| Leu                                     | CTA          | 9 (14%)           |
| Pro                                     | CCC          | 13 (42%)          |
| Ser                                     | AGT          | 1 (2%)            |
| <b>P38<math>\alpha</math> FL native</b> |              |                   |
| <b>AA</b>                               | <b>codon</b> | <b>number (%)</b> |
| Arg                                     | AGG          | 6 (30%)           |
| Arg                                     | CGA          | 1 (5%)            |
| Arg                                     | CGG          | 3 (15%)           |
| Gly                                     | GGA          | 4 (22%)           |
| Gly                                     | GGG          | 3 (17%)           |

promoter (37°C 6 hrs). To quantitate RNA, transcription samples were diluted 1:50 and quantitated against a blank of 1:50 dilution of transcription buffer at A<sub>260</sub> on a Beckman Coulter DU500 spectrophotometer. Thirty-three ug RNA was added to 10.8 ul WEPRO 1240H containing 80 ng/ul creatine kinase and this mixture was laid under 208 ul 1× SUB-AMIX (24 mM HEPES/KOH (pH 7.8), 1.2 mM ATP, 0.25 mM GTP, 16 mM creatine phosphate, 250 units of RNasin (ribonuclease inhibitor), 1 mM dithiothreitol, 0.4 mM spermidine, 0.3 mM each of the 20 amino acids, 2.7 mM magnesium acetate and 100 mM potassium acetate). Translation was carried out overnight at 15–22°C. The following day the bilayer was mixed to homogeneity, and then 100 ul of crude extract was removed, and spun 13,000 rpm 4°C 15 minutes. The supernatant was removed as "soluble" fraction and the small remaining pellet resuspended in 100 ul 10 mM Tris pH 7.5 ("insoluble" fraction). For large scale expression and purification performed by the Desktop II robot (CFS), an H-kit was used, running 6 mls × 6 reaction cups, one construct per cup (six constructs total, NS5B FL, FtsZ FL, and P38 $\alpha$ ; native and engineered versions of each). The DT II program transcribes, translates, and Ni-IMAC purifies targets with a final elution in 0.5 M imidazole. It was carried out using default parameters (6 hours transcription, 16 hours translation at 15°C) and according to the manufacturer's instructions with the following exception. The robot was paused after the transcription step but before addition of wheat germ to the RNA, so that the RNA for each sample could be manually adjusted to 2600 ug/ml (0.3 ml RNA/sample). After this adjustment, automation was resumed through purification. Protein concentrations of fractions from all were estimated with a detergent compatible Protein Assay (Bio-Rad) according to manufacturer's protocols

### Crystallization

Coordinates and structure factors for a FtsZ double deletant construct ( $\Delta$ 1–11,  $\Delta$ 316–382) with GDP bound have been deposited to the Protein Data Bank with accession code 2RHL. Crystals were obtained at 20°C in Compact Jr. (Emerald BioSystems) sitting drop plates using the CryoF crystallization screen (Emerald BioSystems) and 1  $\mu$ L of protein with 1  $\mu$ L of reservoir solution. The following conditions yielded the FtsZ crystals: 30% PEG 200, 100 mM MES pH 6.0, 5% PEG 3000. All crystals were frozen in a fresh drop of their respective crystallant which also served as a cryoprotectant.

### Data Collection, Structure Solution and Refinement

Data were collected at 100 K at the Advanced Photon Source Structural Biology Center beamline 19BM (Argonne, IL) and were integrated and scaled with HKL2000 [44]. Prior to creating the construct described above, we had crystallized full length *B. subtilis* FtsZ con-

**Table 2: Summary of rare codon usage in full length native genes. (Continued)**

|     |     |         |
|-----|-----|---------|
| Leu | CTA | 3 (7%)  |
| Pro | CCC | 4 (22%) |
| Ser | AGT | 1 (4%)  |

Codons rarely found in *E. coli* highly expressed genes are listed on the left. The number of occurrences of each rare codon in the wild type FtsZ, NS5B and P38 $\alpha$  full length genes are listed. The percentage of occurrence (number of occurrences/total number of codons for each amino acid) is listed in parentheses.

struct (unpublished work). This construct crystallized in the trigonal space group  $P3_221$  with  $a = 87.1 \text{ \AA}$ ,  $c = 88.1 \text{ \AA}$  and diffracted to  $\sim 2.8 \text{ \AA}$  resolution. The initial structure was solved by molecular replacement with Molrep [45] using data collected at the Advanced Photon Source COMCAT beamline sector 32 and 1FSZ [36] as the search

model. The search model contained only homologous residues. All other residues of the search model were replaced with alanine. The structure was built manually with O [46] and refined with Refmac [45]. After successive steps of building and refinement, the *R*-factor converged at  $\sim 24\%$ . The structure spanned residues Ala 12 to Phe 315 as the N-terminal (1–11) and C-terminal (316–382) residues were disordered. An isomorphous *B. subtilis* FtsZ structure has been deposited to the protein data bank (PDB ID: [2VAM](#)) [47]. The model was improved using Arp/Warp [48] and additional building was carried out with Coot [49]. This refined model was used for molecular replacement searches against the additional data sets. Residues modeled beyond Phe 315 correspond to the rTEV-His-Glu purification tag. All refinements were performed with Refmac. Data collection and refinement statistics are provided in Table 3. Figures were created using the Ribbons software package [50].

**Table 3: Crystallographic data for *Bacillus subtilis* FtsZ structures.**

| Data Collection                                     | 2RHH                | 2RHJ                 | 2RHL                               | 2RHO                               |
|---|---------------------|----------------------|------------------------------------|------------------------------------|
| Unit Cell <i>a</i> , <i>b</i> , <i>c</i> (Å)        | 66.37, 66.37 152.09 | 74.14, 82.06, 117.51 | 82.29, 97.58, 135.45               | 82.30, 97.16, 134.72               |
| Space group   | $P4_12_12$          | $C222_1$             | $P2_12_12_1$                       | $P2_12_12_1$                       |
| Resolution (Å) range/highest resolution shell       | 50-2.0/(2.07-2.0)   | 50-1.76/(1.82-1.76)  | 30-2.45/(2.54-2.45)                | 50-2.45/(2.54-2.45)                |
| Wavelength (Å)                                      | 0.97934             | 0.97934              | 0.97934                            | 0.97934                            |
| Observed Reflections                                | 267,428             | 236,328              | 262,233                            | 228,703                            |
| Unique Reflections                                  | 23,413              | 34,949               | 41,246                             | 40,114                             |
| Average $I/(\sigma I)^a$                            | 19.8 (1.4)          | 27.4 (2.1)           | 29.0 (2.6)                         | 24.7 (2.0)                         |
| $R_{\text{sym}}$ (%) <sup>a,c</sup>                 | 8.7 (44.9)          | 7.1 (64.2)           | 6.9 (47.6)                         | 5.0 (56.8)                         |
| Completeness (%) <sup>a</sup>                       | 98.3 (88.6)         | 98.0 (88.8)          | 99.2 (98.2)                        | 98.1 (94.9)                        |
| Redundancy <sup>a</sup>                             | 11.4 (6.0)          | 6.8 (4.6)            | 6.4 (5.5)                          | 5.7 (4.7)                          |
| <b>Refinement</b>                                   |                     |                      |                                    |                                    |
| Resolution (Å)                                      | 50 - 2.0            | 50 - 1.76            | 30 - 2.45                          | 50 - 2.45                          |
| Reflections (working/test)                          | 22,153/1,194        | 33,173/1,752         | 38,318/2,039                       | 37,673/2,005                       |
| $R_{\text{cryst}}/R_{\text{free}}$ (%) <sup>d</sup> | 21.2/25.7           | 18.0/20.8            | 22.7/27.0                          | 22.7/27.0                          |
| Number of atoms (protein/ligand/water)              | 2191/5/114          | 2279/64/172          | 2199, 2290 <sup>b</sup> /<br>28/97 | 2199, 2178 <sup>b</sup> /32, 28/99 |
| r.m.s.d. bond lengths (Å)                           | 0.019               | 0.017                | 0.015                              | 0.014                              |
| r.m.s.d. bond angles (°)                            | 1.580               | 1.520                | 1.521                              | 1.508                              |
| Average B factor all (Å <sup>2</sup> )              | 34.2                | 27.3                 | 67.8                               | 54.8                               |
| Average B factor protein (Å <sup>2</sup> )          | 36.8                | 25.7                 | 68.0/67.8 <sup>b</sup>             | 55.2/54.7 <sup>b</sup>             |
| Average B factor ligand (Å <sup>2</sup> )           | 31.7                | 43.6                 | 59.1/53.4 <sup>b</sup>             | 44.4/43.1 <sup>b</sup>             |
| Average B factor water (Å <sup>2</sup> )            | 37.1                | 39.5                 | 67.7                               | 52.4                               |
| Coordinate Error (Å) <sup>e</sup>                   | 0.170               | 0.104                | 0.235                              | 0.235                              |
| <b>Ramachandran Analysis (%)</b>                    |                     |                      |                                    |                                    |
| Most Favored  | 95.8                | 96.6                 | 94.8/95.4 <sup>b</sup>             | 96.5/96.1 <sup>b</sup>             |
| Additionally Allowed                                | 4.2                 | 3.4                  | 5.2/4.6 <sup>b</sup>               | 3.5/3.5 <sup>b</sup>               |

<sup>a</sup>Values in parenthesis are for the highest resolution shells.

<sup>b</sup>Values for chains A and B respectively when applicable.

<sup>c</sup> $R_{\text{sym}} = (\sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl) \rangle|) / \sum_{hkl} \sum_i I_i(hkl)$ ;  $I_i(hkl)$  is the intensity measured for the  $i^{\text{th}}$  reflection and  $\langle I(hkl) \rangle$  is the mean intensity of reflections with indices  $hkl$ .

<sup>d</sup> $R_{\text{cryst}} = (\sum_{hkl} ||F_{\text{obs}}(hkl)| - |F_{\text{calc}}(hkl)||) / \sum_{hkl} F_{\text{obs}}(hkl)$ ;  $R_{\text{free}}$  is calculated in an identical manner using 5% of randomly selected reflections.

<sup>e</sup>Estimated standard uncertainty based on  $R_{\text{free}}$ .

## Availability and requirements

Gene Composer software can be downloaded from <http://www.genecomposer.net>.

## Authors' contributions

AR contributed to experimental strategy, analyzed data, conducted experiments, and drafted the manuscript. DL contributed to experimental strategy, analyzed data, and drafted the manuscript. SL managed all aspects of the protein crystallography and solved the reported structures. JW and MM wrote the Gene Composer database software package. EW and KT cloned the expression constructs. KT and KA purified and analyzed recombinant proteins. AB and LS conceptualized experimental strategy and guided the manuscript preparation. All authors read and approved the final manuscript.

## Additional material

### Additional file 1

Gene and protein sequences for all targets.

Gene and protein sequences for all targets. Nucleic acid and amino acid sequences for all targets.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1472-6750-9-37-S1.doc>]

## Acknowledgements

Special thanks to Kai Post and Kathryn Hjerrild for contributions to Gene Composer and gene synthesis. This work was supported in part by the NIGMS-NCRR co-sponsored PSI-2 Specialized Center Grant U54 GM074961 for the Accelerated Technologies Center for Gene to 3D Structure, and by the Seattle Structural Genomics Center for Infectious Disease funded by the NIAID Contract HHSN266200700057C.

## References

- Malawski GA, Hillig RC, Monteclaro F, Eberspaecher U, Schmitz AA, Crusius K, Huber M, Egner U, Donner P, Muller-Tiemann B: **Identifying protein construct variants with increased crystallization propensity – a case study.** *Protein Sci* 2006, **15**:2718-2728.
- Graslund S, Sagemark J, Berglund H, Dahlgren LG, Flores A, Hammarstrom M, Johansson I, Kotenyova T, Nilsson M, Nordlund P, Weigelt J: **The use of systematic N- and C-terminal deletions to promote production and structural studies of recombinant proteins.** *Protein Expr Purif* 2008, **58**:210-221.
- Lithwick G, Margalit H: **Hierarchy of sequence-dependent features associated with prokaryotic translation.** *Genome Res* 2003, **13**:2665-2673.
- Osawa S: *Evolution of the Genetic Code* Oxford: Oxford University Press; 1995.
- Karlin S, Mrazek J, Campbell A, Kaiser D: **Characterizations of highly expressed genes of four fast-growing bacteria.** *J Bacteriol* 2001, **183**:5025-5040.
- Sharp PM, Devine KM: **Codon usage and gene expression level in Dictyostelium discoideum: highly expressed genes do 'prefer' optimal codons.** *Nucleic Acids Res* 1989, **17**:5029-5039.
- Sharp PM, Tuohy TM, Mosurski KR: **Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes.** *Nucleic Acids Res* 1986, **14**:5125-5143.
- Bennetzen JL, Hall BD: **Codon selection in yeast.** *J Biol Chem* 1982, **257**:3026-3031.
- Ikemura T: **Codon usage and tRNA content in unicellular and multicellular organisms.** *Mol Biol Evol* 1985, **2**:13-34.
- Moriyama EN, Powell JR: **Codon usage bias and tRNA abundance in Drosophila.** *J Mol Evol* 1997, **45**:514-523.
- Duret L: **tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes.** *Trends Genet* 2000, **16**:287-289.
- Bulmer M: **The selection-mutation-drift theory of synonymous codon usage.** *Genetics* 1991, **129**:897-907.
- Kane JF: **Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli.** *Curr Opin Biotechnol* 1995, **6**:494-500.
- Gustafsson C, Govindarajan S, Minshull J: **Codon bias and heterologous protein expression.** *Trends Biotechnol* 2004, **22**:346-353.
- Cormack BP, Bertram G, Egerton M, Gow NA, Falkow S, Brown AJ: **Yeast-enhanced green fluorescent protein (yEGFP) a reporter of gene expression in Candida albicans.** *Microbiology* 1997, **143**:303-311.
- Collins K, Gandhi L: **The reverse transcriptase component of the Tetrahymena telomerase ribonucleoprotein complex.** *Proc Natl Acad Sci USA* 1998, **95**:8485-8490.
- Burgess-Brown NA, Sharma S, Sobott F, Loenarz C, Oppermann U, Gileadi O: **Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study.** *Protein Expr Purif* 2008, **59**:94-102.
- Hagervall TG, Bjork GR: **Undermodification in the first position of the anticodon of supG-tRNA reduces translational efficiency.** *Mol Gen Genet* 1984, **196**:194-200.
- Bouadloun F, Srichaiyo T, Isaksson LA, Bjork GR: **Influence of modification next to the anticodon in tRNA on codon context sensitivity of translational suppression and accuracy.** *J Bacteriol* 1986, **166**:1022-1027.
- Bjork GR, Durand JM, Hagervall TG, Leipuviene R, Lundgren HK, Nilsson K, Chen P, Qian Q, Urbonavicius J: **Transfer RNA modification: influence on translational frameshifting and metabolism.** *FEBS Lett* 1999, **452**:47-51.
- Gutman GA, Hatfield GW: **Nonrandom utilization of codon pairs in Escherichia coli.** *Proc Natl Acad Sci USA* 1989, **86**:3699-3703.
- Boycheva S, Chkodrov G, Ivanov I: **Codon pairs in the genome of Escherichia coli.** *Bioinformatics* 2003, **19**:987-998.
- Boycheva SS, Bachvarov BI, Berzal-Heranz A, Ivanov IG: **Effect of 3' terminal codon pairs with different frequency of occurrence on the expression of cat gene in Escherichia coli.** *Curr Microbiol* 2004, **48**:97-101.
- Wu G, Dress L, Freeland SJ: **Optimal encoding rules for synthetic genes: the need for a community effort.** *Mol Syst Biol* 2007, **3**:134.
- Wu G, Zheng Y, Qureshi I, Zin HT, Beck T, Bulka B, Freeland SJ: **SGDB: a database of synthetic genes re-designed for optimizing protein over-expression.** *Nucleic Acids Res* 2007, **35**:D76-79.
- Villalobos A, Ness JE, Gustafsson C, Minshull J, Govindarajan S: **Gene Designer: a synthetic biology tool for constructing artificial DNA segments.** *BMC Bioinformatics* 2006, **7**:285.
- de Smit MH, van Duijn J: **Control of prokaryotic translational initiation by mRNA secondary structure.** *Prog Nucleic Acid Res Mol Biol* 1990, **38**:1-35.
- Jin H, Zhao Q, Gonzalez de Valdivia EI, Ardell DH, Stenstrom M, Isaksson LA: **Influences on gene expression in vivo by a Shine-Dalgarno sequence.** *Mol Microbiol* 2006, **60**:480-492.
- Lorimer D, Raymond A, Walchli J, Mixon M, Barrow A, Wallace E, Grice R, Burgin A, Stewart L: **Gene Composer: Database Software for Protein Construct Design, Codon Engineering, and Gene Synthesis.** *BMC Biotechnology* 2009, **9**:36.
- Roosild TP, Vega M, Castronovo S, Choe S: **Characterization of the family of Mistic homologues.** *BMC Struct Biol* 2006, **6**:10.
- Kefala G, Kwiatkowski W, Esquivies L, Maslennikov I, Choe S: **Application of Mistic to improving the expression and membrane integration of histidine kinase receptors from Escherichia coli.** *J Struct Funct Genomics* 2007, **8**:167-172.
- Cordell SC, Robinson EJH, Löwe J: **Crystal structure of the SOS cell division inhibitor SulA and in complex with FtsZ.** *Proc Natl Acad Sci USA* 2003, **100**:7889-7894.
- Oliva MA, Cordell SC, Löwe J: **Structural insights into FtsZ protofilament formation.** *Nat Struct Mol Biol* 2004, **11**:1243-1250.

34. Mukherjee A, Dai K, Lutkenhaus J: **Escherichia coli cell division protein FtsZ is a guanine nucleotide binding protein.** *Proc Natl Acad Sci USA* 1993, **90**:1053-1057.
35. Erickson HP: **Atomic structures of tubulin and FtsZ.** *Trends Cell Biol* 1998, **8**:133-137.
36. Hale RS, Thompson G: **Codon optimization of the gene encoding a domain from human type I neurofibromin protein results in a threefold improvement in expression level in Escherichia coli.** *Protein Expr Purif* 1998, **12**:185-188.
37. Löwe J, Amos LA: **Crystal structure of the bacterial cell-division protein FtsZ.** *Nature* 1998, **391**:203-206.
38. Harding MM: **Geometry of metal-ligand interactions in proteins.** *Acta Crystallogr D Biol Crystallogr* 2001, **57**:401-411.
39. Nayal M, Di Cera E: **Valence screening of water in protein crystals reveals potential Na<sup>+</sup> binding sites.** *J Mol Biol* 1996, **256**:228-234.
40. Harding MM: **Metal-ligand geometry relevant to proteins and in proteins: sodium and potassium.** *Acta Crystallogr D Biol Crystallogr* 2002, **58**:872-874.
41. Studier FW: **Protein production by auto-induction in high density shaking cultures.** *Protein Expr Purif* 2005, **41**:207-234.
42. Madin K, Sawasaki T, Ogasawara T, Endo Y: **A highly efficient and robust cell-free protein synthesis system prepared from wheat embryos: plants apparently contain a suicide system directed at ribosomes.** *Proc Natl Acad Sci USA* 2000, **97**:559-564.
43. Sawasaki T, Hasegawa Y, Tsuchimochi M, Kamura N, Ogasawara T, Kuroita T, Endo Y: **A bilayer cell-free protein synthesis system for high-throughput screening of gene products.** *FEBS Lett* 2002, **514**:102-105.
44. Otwinowski Z, Minor W: **Processing of x-ray diffraction data collected in oscillation mode.** *Methods Enzymol* 1997, **276**:307-326.
45. Murshudov GN, Vagin AA, Dodson EJ: **Refinement of macromolecular structures by the maximum-likelihood method.** *Acta Cryst D* 1997, **53**:240-255.
46. Jones TA, Zou JY, Cowan SW, Kjeldgaard M: **Improved methods for building protein models in electron density maps and the location of errors in these models.** *Acta Crystallogr A* 1991, **47**:110-119.
47. Oliva MA, Trambaiolo D, Löwe J: **Structural insights into the conformational variability of FtsZ.** *J Mol Biol* 2007, **373**:1229-1242.
48. Lamzin VS, Wilson KS: **Automated refinement for protein crystallography.** *Methods Enzymol* 1997, **277**:269-305.
49. Emsley P, Cowtan K: **Coot: model-building tools for molecular graphics.** *Acta Crystallogr D Biol Crystallogr* 2004, **60**:2126-2132.
50. Carson M: **Ribbons.** *Methods in Enzymology* 1997, **277**:493-505.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

