Journal of Structural and Functional Genomics **4:** 141–157, 2003.

Structural genomics of highly conserved microbial genes of unknown function in search of new antibacterial targets

Chantal Abergel*, Bruno Coutard, Deborah Byrne, Sabine Chenivesse, Jean-Baptiste Claude, Céline Deregnaucourt, Thierry Fricaux, Celine Gianesini-Boutreux, Sandra Jeudy, Régine Lebrun[1], Caroline Maza, Cédric Notredame, Olivier Poirot, Karsten Suhre, Majorie Varagnol & Jean-Michel Claverie

*Structural and Genomic Information Laboratory, UMR 1889 CNRS-AVENTIS, [1]Institute for Structural Biology and Microbiology, 31 Chemin Joseph Aiguier, 13402 Marseille cedex 20, France; *Author for Correspondence: E-mail: Chantal.Abergel@igs.cnrs-mrs.fr*

Received 22 November 2002; Accepted in revised form 27 April 2003

Key words: Escherichia coli, Structural Genomics, Comparative Genomics, Bioinformatics, anti-bacterial

Abstract

With more than 100 antibacterial drugs at our disposal in the 1980's, the problem of bacterial infection was considered solved. Today, however, most hospital infections are insensitive to several classes of antibacterial drugs, and deadly strains of *Staphylococcus aureus* resistant to vancomycin – the last resort antibiotic – have recently begin to appear. Other life-threatening microbes, such as *Enterococcus faecalis* and *Mycobacterium tuberculosis* are already able to resist every available antibiotic. There is thus an urgent, and continuous need for new, preferably large-spectrum, antibacterial molecules, ideally targeting new biochemical pathways. Here we report on the progress of our structural genomics program aiming at the discovery of new antibacterial gene targets among evolutionary conserved genes of uncharacterized function. A series of bioinformatic and comparative genomics analyses were used to identify a set of 221 candidate genes common to Gram-positive and Gram-negative bacteria. These genes were split between two laboratories. They are now submitted to a systematic 3-D structure determination protocol including cloning, protein expression and purification, crystallization, X-ray diffraction, structure interpretation, and function prediction. We describe here our strategies for the 111 genes processed in our laboratory. Bioinformatics is used at most stages of the production process and out of 111 genes processed – and 17 months into the project – 108 have been successfully cloned, 103 have exhibited detectable expression, 84 have led to the production of soluble protein, 46 have been purified, 12 have led to usable crystals, and 7 structures have been determined.

Introduction

Despite the exponential growth of sequence information from a large diversity of organisms, each newly sequenced bacterial genome continues to reveal up to 50% of genes without significant similarity to protein of characterized function [1]. By focusing a structural genomics project on such *anonymous* proteins, our goal is twofold. On one hand, we expect a sizable fraction of these proteins to exhibit a recognizable 3-D structure similarity with previously characterized protein families. Structure determination is thus used as a technique of functional genomics, allowing common functional attributes to be recognized beyond the twilight zone of sequence similarity. On the other hand, focusing a structural genomics effort on anonymous proteins should also enhance the probability of discovering original folds that are highly valuable by-products for the academic community.

The first antibacterial drug was made available in 1936 (sulfonamides), and the list regularly expanded during the following 30 years (beta-lactam (1940), tetracyclines (1949), chloramphenicol (1949), aminoglycosides (1950), macrolides (1952), strepto-

gramins (1962), quinolones (1962), rifampin (1963)). With more than 100 different drugs available in the 1980's, the problem of bacterial infection was considered to be definitely solved. However, resistant strains started to appear and spread quickly. Today, most hospital-acquired infections are insensitive to one or several classes of antibiotics, and deadly strains of *Enterococcus faecalis* or *Mycobacterium tuberculosis* have been found resistant to all available drugs. Until the recent approval of an oxazolidinone, Linezolid, and ketolide, telithromycin, by the FDA (2000, 2003 respectively), no new class of antibacterial drug had been approved for more than 25 years, and there are very few other classes of compound currently in clinical development [2, 3] besides some promising inhibitors of peptide deformylase [4].

Linezolid and telithromycin [5, 6] are active against a broad spectrum of gram-positive pathogens, including multidrug resistant *Staphylococci, Streptococci* and *Enterococci.* However, Linezolid resistant *Enterococci* strains have already been reported [7]. The utility of existing antimicrobial agents is thus rapidly eroding, and the need for research directed toward development of new antibiotics has never been greater.

It is remarkable that less than 20 distinct evolutionary conserved macromolecules – belonging to 5 essential pathways – constitute the targets of all antibiotics available today. Given that pathogenic bacteria have a thousand genes or more, it is realistic to expect that the rational analysis of the abundant genomic information – more than 60 complete bacterial genomes sequences – that have become recently available could unravel a sizable set of entirely novel target genes and inspire the design of original classes of antibacterial molecules. Our approach is geared toward the identification of genes and proteins involved in essential biochemical pathways not yet targeted by existing antibacterial molecules.

The work presented in this paper concerns the two first steps of a rational drug development program: i) the identification of candidate target genes and ii) the determination of the three-dimensional structure of the corresponding proteins (production, purification, characterization, crystallization, and crystallographic analysis). We first describe the comprehensive bioinformatic analysis used to define the subset of the most promising candidate genes based on their evolutionary conservation across a large panel of bacterial species (Table 1), and their further prioritization using properties computed from their amino-acid sequence.

Table 1: Genome sequence data. The 'reference genomes' correspond to complete genomes used to determine the most conserved genes. Genomes in bold correspond to the one used to determine the genes conserved both in *E. coli* and at least one gram-positive bacteria. 'Other genomes' correspond to complete genomes of less biomedical relevance or incomplete genomes near completion. These genomes are accessible for further analysis.

Reference Genomes	Other Genomes
Aquifex aeolicus VF5	Archaeoglobus fulgidus
Bacillus subtilis 168	Aeropyrum pernix
Borrelia burgdorferi B31	Bacillus halodurans C–125
Campylobacter jejuni NCTC 11168	Buchnera sp.Clostridium acetobutylicum
Chlamydia muridarum	Deinococcus radiodurans R1
Chlamydia pneumoniae AR39	Halobacterium sp.
Chlamydia pneumoniae AR39 plasmid	Methanococcus jannaschii
Chlamydia pneumoniae CWL029	Mycobacterium leprae
Chlamydia trachomatis serovar D	Mycoplasma pulmonis
Chlamydophila pneumoniae J138	Mesorhizobium loti
Escherichia coli K–12 MG1655	Methanobacterium thermoautotrophicum
Escherichia coli O157:H7	Neisseria meningitidis
Haemophilus influenzae KW20	Pyrococcus abyssi
Helicobacter pylori J99	Pyrococcus horikoshii
Helicobacter pylori 26695	Sinorhizobium meliloti
Lactococcus lactis IL1403	Staphylococcus aureus strain Mu50
Mycobacterium tuberculosis H37Rv	Staphylococcus aureus N315
Mycoplasma genitalium G–37	Streptococcus pneumoniae
Mycoplasma pneumoniae M129	Sulfolobus solfataricus
Neisseria meningitidis MC58	Synechocystis sp.
Pasteurella multocida PM70	Thermoplasma acidophilum
Pseudomonas aeruginosa PA01	Thermotoga maritima
Rickettsia conorii	Treponema pallidum
Rickettsia prowazekii Madrid E	Xylella fastidiosa
Streptococcus pyogenes	
Treponema pallidum Nichols	
Ureaplasma urealyticum serovar 3	
Vibrio cholerae chromosome I & II	

We then present the various experimental techniques integrated into our 3-D structure determination pipeline.

Finally, we show how the mapping of paradoxical conservation patterns revealed by the multiple alignment of orthologous sequences can help the interpretation of anonymous 3-D structures (prediction of potential active and cofactor-binding sites) and suggest functional hypotheses for further experimental validation.

Materials and methods

Target identification

The complete genome sequences were collected for a number of bacterial species covering a wide range of evolutionary distances (Gram-negative to Gram-positive) and life-style (free-living, parasitic and/or intra-cellular) (see Table 1). For a subset of these genomes ('Reference Genomes' in Table 1) all potential coding regions (ORFs) of length larger than 150 nucleotides (with ATG, GTG or TTG as initiator codon) were extracted and their conceptual translation stored. This represented a total of 275,791 putative ORFs, many of them probably not corresponding to actual proteins.

All large-scale sequence comparisons were performed using a distributed processing protocol using a cluster of 48 PC ('gigablaster' [8]) under the linux operating system. Each node included a 500 MH Pentium III, 256 Mb of RAM and a 13 Gb disk. Using the sequence alignment program Blastp [9], each of these putative ORFs were compared to i) the set of all *E. coli K12* ORFs and ii) the sets of *B. subtilis*, *L. lactis*, *S. pyogenes*, and *M. tuberculosis* ORFs. We only retained the ORFs with a significant match in both sets. Each retained ORF was then mapped back to its *E. coli* orthologous gene, using Ecogene [10] as our primary reference database. We then further reduced our target gene set by only retaining the genes of *unknown* or *hypothetical* function ('Y' genes according to the *E. coli* nomenclature). A prediction of membrane spanning segments [11] was then performed on each of the previously selected genes to determine which were most likely to be expressed as soluble globular proteins.

Cloning

In order to allow batches of ORFs to be processed in parallel, directional cloning was performed using the GATEWAY system (Invitrogen, [12]). Oligonucleotide primers were designed for each of the *E. coli* ORF with AttB1 and AttB2 overhangs added for recombination cloning. PCR was performed on *E. coli* K12 purified genomic DNA using Pfx proof reading polymerase from Invitrogen. Alternatively, Ex Taq (Takara) or Platinum High fidelity Taq (Invitrogen) were used to relieve amplification problems. PCR products were purified as described in the GATEWAY manual (Invitrogen) using Poly Ethylene Glycol precipitation. This procedure was further automatized using Multi screen-PCR 96 wells purification from Millipore. The corresponding PCR products were inserted by homologous recombination using the 'one tube reaction' in the pDEST17 expression plasmid in frame with a N-terminal His_6-tag, under the control of a T7 promoter. The reaction was performed using 1 µl BP reaction buffer, 25–125 ng PCR products in a volume of 2.5 µl, 0.5 µl pDONR vector and 1 µl BP clonase enzyme mix. The reaction mix was incubated at 25 °C for 4 hours before adding 0.25 µl NaCl 0.75 M, 0.75 µl pDEST17 and 1.5 µl LR clonase and then incubated 2 hours at 25 °C. 0.5 µl proteinase K was added 10 minutes at 37 °C before transformation in DH5α. 3 colonies per cloning were picked for overnight cultures using 2ml LB amp medium. Plasmids were then purified using Qiagen miniprep kit (spin or racks) and positive plasmids were screened by restriction profiling using NdeI and HindIII enzymes (Biolab). Positive clones were sequenced over a length of 600 nucleotides (GenomExpress) using attB1 primer (5′ ACA AGT TTG TAC AAA AA AGC 3′)

Protein expression

The purified plasmids were used for the over-expression of the recombinant proteins both in *E. coli* BL21(DE3) cells and in a cell-free system (Roche Applied Science), (First pass, standard protocol, [13] for details). Incomplete factorial experimental design (as implemented by the SamBa software [14], http://igs-server.cnrs-mrs.fr/samba/) was used to define a set of 12 conditions corresponding to the combination of 3 variables (Expression strain, Medium type and Temperature). This set of conditions was used as a starting point for the second pass screening/optimization protocols. A partially automated procedure (including liquid handling, dilutions, bacterial lysate normalization via a connection to a Packard/µquant spectrophotometer (Bio-Tek), and dot-blot spotting) was developed using a Multiprobe II (Perkin elmer)

automated platform allowing us to screen the expression of 12 ORFs in parallel. The cloned ORFs were transformed in four *E. coli* expression strains: BL21 (DE3) pLysS, Rosetta (DE3), Origami (DE3) (Novagen), and C41 (Avidis). For each transformed strain, a colony was picked for overnight preculture in 2 ml 2YT with the suitable antibiotic. The Multiprobe II automated platform dispensed on 24 wells plates 2.5 ml of one of the 3 selected media (2YT (Difco), Superior Broth (SB) and Turbo Broth (TB) (Athena Enzyme Systems)) – each with the suitable antibiotic –, and inoculated them with 2.5% precultures. Culture plates were then incubated at 37 °C with 250 rpm agitation for one hour before being temperature regulated at one of the 3 defined temperatures (25 °C, 37 °C and 42 °C). Induction was performed on the Packard platform with 0.5 mM IPTG when OD_{600nm} reached 0.4–0.8. Cultures were stopped 3 hours after induction. The measured OD_{600nm} value is used to normalize the amount of bacteria used for protein quantification. The 12 resulting aliquots were then dispensed in a 96 well plate and centrifuged 10 minutes at 4000 rpm. Pellets were resuspended in 250 µl BugBuster (Novagen), 0.1 mg/ml lysozyme and 0.1 mg/ml DnaseI (Euromedex) 30 min 4 °C under agitation. 100 µl of each bacterial lysate was used for estimating the total amount of expressed protein, and 150 µl was used to assay the level of soluble expression.

Assessment of soluble expression

A clearing filter plate (MultiScreen NucleicA from Millipore), a 0.22 µm hydrophilic filter plate (MultiScreen GV from Millipore) and a 96 well plate were assembled. The lysates were dispensed on the first plate and filtered in one step with one minute centrifugation at 4000 rpm. Lysates were conserved at –20 °C. The detection of recombinant protein was performed using a Dot Blot procedure. 4 µl of both the soluble and total protein was resuspended in 100 µl of a denaturant buffer (urea 8 M, Sodium Phosphate 100 mM Tris-HCl 10 MM pH 8). 25 µl of this mix were transferred onto nitrocellulose membrane Protan 45 (Schleicher and Schuell) using a blotter adapted for the Multiprobe station. The membrane was washed twice in TBS Buffer (Tris-HCl 10 mM, NaCl 150 mM pH 7.5), blocked in blocking reagent 1X (Qiagen) with 0.1% Tween−20 for 1 hour, washed 2 times 10 minutes in TBS-TT buffer (Tris HCl 20 mM NaCl 500 mM, 0.02% Tween−20, 0.2% Triton X−100), and washed once 10 minutes in TBS before incubation with a penta-His HRP conjugated antibody (Qiagen) diluted 1/2000 in blocking reagent 1X with 0.1% Tween−20 for 1 hour. Subsequent washes were performed twice for 15 minutes in TBS-TT, and once 10 minutes in TBS. His-tagged protein expression was revealed by detecting HorseRadish Peroxidase activity conjugated to the anti-His_5 antibody, using chemiluminescence with ECL RPN2106 kit (Amersham Biosciences) on Biomax films (Kodak) . Each of the 12 conditions received a score according to the Dot Blot spot intensity. The influence of each variable on the soluble expression was then extracted by adding the scores of the experiments sharing the same variable value and comparing this total score to the others. For instance, if we consider the variable 'temperature', the sum of the scores of the four experiments performed at 37 °C is compared to the sum of the scores of the four experiments done at 25 °C and at 42 °C. In the case these total scores are similar, temperature is not considered as a significant factor for protein expression and solubility. If the scores are different, temperature is retained as a pertinent variable. The temperature corresponding to the highest total score was then used for the optimized expression protocol.

Protein purification and characterization

Protein purifications were performed using the AKTÄ explorer 10S (Amersham-Pharmacia) using standard affinity chromatography on HiTrap chelating (Amersham-Pharmacia) or NiNTA (Qiagen) columns charged with Ni^{2+}, eluted with an imidazole gradient after removal of non-specific low interacting proteins by a preliminary wash with 50mM imidazole buffer. The recombinant proteins were usually eluted within a 70−150 mM imidazole range. A detailed protocol has been published elsewhere [15]

A microdialysis step is applied to each protein to identify a suitable buffer for desalting and concentration. We use 250 µl Dialysis Buttons (Hampton Research: Calafornia, USA) with MWCO 10,000 Spectra/Por CE membrane (Spectrum Labs, California). 250 µl of each sample are loaded into 6 dialysis buttons, and dialyzed overnight at 4 °C on a rotating wheel against 50 ml of 6 different buffers in the presence or absence of ionic strength (NaCl 0.2 M). We used the following buffers: Tris 10 mM pH (7–9), Mes 10 mM pH (5.5–6.7), Mops 20 mM pH (6.9–8.3), Hepes 10 mM pH (6.8–8.2), Imidazole

10 mM pH (6.2–7.8) and Sodium Acetate 10 mM pH (3.6–5.6). The various buffers were adjusted to at least one pH unit above or below the pI of each protein. To estimate the amount of precipitated *versus* soluble protein, 125 µl of each sample were loaded onto a 96 microtitre plate for subsequent UV scanning (DO_{280}total). The remaining 125 µl were centrifuged at 12,000 rpm for 20 minutes and the supernatant loaded onto a 96 microtitre plate for soluble protein evaluation (DO_{280}soluble). The comparison of the two UV scans using the respective buffer as the blank with the KC4 software on a µquant spectrophotometer (Bio-tek instruments inc.) was used to select the optimal desalting buffer as the one corresponding to the smallest $\Delta(DO_{280}\text{total} - DO_{280}\text{soluble})$. The buffers were then exchanged using the Hiprep 26/10 desalting column (Amersham-pharmacia) on the AKTÄ explorer 10S.

Quality control of the samples

Before entering crystallization trials, purified recombinant proteins were first characterized using circular dichroïsm spectroscopy to assess the folding status of the expression products, and its consistency with secondary structure predictions [11]. Monodispersity being a key factor in the crystallization process, the aggregation state of the protein solution was systematically monitored by dynamics light scattering (Dynapro MS800-TC). These measurements were also used to define the most suitable buffer in which to perform concentration, phase diagrams, and crystallization trials. Each sample was also controlled on iso-electro-focusing, native and SDS gels both in the absence and presence of reducing agent and 2 M urea. The results were compared with the theoretical molecular weight of the target, as well as with its predicted isoelectric point and cysteine content. The stability of the protein over time was assessed by storing purified samples at 3 temperatures: 20 °C, 4 °C and –80 °C. Every two weeks, they were controlled by electrophoresis on SDS gels to reveal an eventual degradation process.

Finally, the purified proteins were all characterized by mass spectroscopy (MALDI-TOF, Voyager DE-RP, Perceptive Biosystem) and by 30 cycles of amino-acid N-terminal Edman sequencing (Applied Biosystem 473A).

Crystallization

The screening for crystallization conditions was performed on 3x96-well crystallization plates (Greiner) loaded by an 8-needle dispensing robot (Tecan, WS 100/8 workstation modified for our needs), using one 1-µl sitting drop per condition. Each protein was initially tested against 480 different conditions at a concentration determined by the phase diagram analysis. The tested crystallization conditions include in-house designed [14] and commercially available solution sets (CrystalScreen-Hampton research, Wizards-Emerald BioStructures). After analyzing the results of this initial screen, conditions were refined using the incomplete factorial design approach [14 and ref. herein].

Data collection and structure determination

Diffraction data for proteins with structural homologues in the PDB [16] are interpreted using the molecular replacement procedure using the AMoRe software [17]. Alternatively, the Multiwavelength Anomalous Diffraction (MAD) technique [18 and ref. herein] was used for interpreting the diffraction data of protein with no obvious structural homologues, using crystals of the seleno-methionine substituted proteins [19]. Detailed protocols can be found in [20].

Results and discussion

Target selection

Because of the detrimental effect of their mutations, genes belonging to essential pathways exhibit low evolution rates. Moreover, the reliable identification – using sequence similarity – of homologous genes in very distant organisms is only possible for slowly evolving genes. Thus, one expect the subsets of ubiquitous or 'wide-spectrum' genes to be enriched in *essential* genes for the same panel of microorganisms.

The first step of the target selection process thus consisted into a comprehensive cross-comparison of a large panel of complete genomes from Gram-negative and Gram-positive bacteria (separated by 2 billions years of divergent evolution), to identify the most conserved genes among those with clear homologues both in *E. coli* and at least one Gram-positive bacteria. This first step, using a BlastP score threshold of 150 (with Blosum62), resulted into a set of

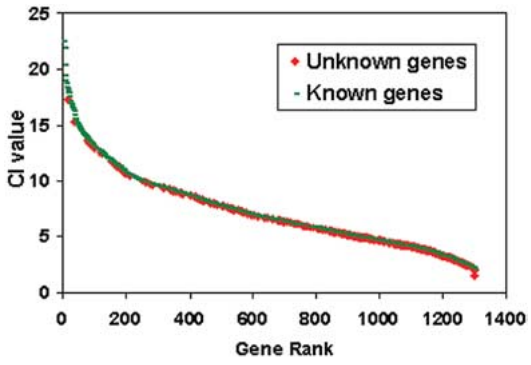

Figure 1. Mandelbrot plot (rank vs. score) of the initial 1300 putative target genes. The genes (X-axis) are ranked according to their associated CI value (Y-axis). The first 200 genes with CI value >10 are ubiquitous genes involved in the most essential cellular pathways. The rest of the distribution is smoothly decreasing suggesting a fairly homogeneous selection pressure among all these candidate target genes. Genes of unknown functions are in red, gene of known functions are in green.

1300 different homologous ORF sets, each with a different representative in *E. coli*. For each of these *E. coli* ORFs a conservation index was computed as:

$$\mathrm{CI} = \sum_{all} \frac{BestScore}{E.coli.SelfScore}$$

Where E.coli.SelfScore is the BlastP [9] score of each *E.coli* ORF aligned with itself, and BestScore corresponds to the score of the best matching ORFs within each genome (counting 0 for matches scoring below the threshold), the sum being performed over all tested genomes. The CI value is thus quantifying the sequence conservation of the gene (and its presence) across a large number of evolutionary diverse microbial genomes, including important pathogens. For the initial 1300 selected genes, CI values ranged from 22.5 (for the elongation factor subunit tufB) to 1.4 for the putative sugar hydrolase ybgG gene (Figure 1).

For all 1300 selected genes (as well as for all 4222 *E. coli* genes) a number of properties were precomputed and stored in a master database ('Master DB'). In this database, each gene is represented by a specific interactive identity card (Figure 2). The ID card contains the gene nucleotide and amino-acid sequence, its precise position in the *E. coli* genome, its significant protein matches within the Reference Genome set, some theoretical properties of its protein product (molecular weight, pI, extinction coefficient, number of cysteines, methionines, and charged residues). Each gene ID card also gives access to additional features such as a restriction map of the gene, the amino-acid translations derived from alternative reading frames (for error checking), a map of the rare codons, as well as the protein hydropathy profile, predicted signal peptide, membrane spanning segments and secondary structure. The 'Blast' button (Figure 2) gives access to interactive similarity searches against additional genomes listed in Table 1 ('Other Genomes'). This set includes the complete genomes of microorganisms of lesser biomedical relevance (e.g. hyperthermophilic archebacteria) or interesting bacteria the genomic sequence of which is nearing completion. This list is submitted to periodical changes. The 'extended alignment' button generate a multiple alignment (using the T-coffee [21] program) of the *E. coli* ORF with the homologous ORFs identified in the other bacterial genomes. The 'Pdb alignment' button does the same with homologues of known 3-D structures. Master DB and the results of these various sequence analyses are frequently consulted to guide the subsequent experimental steps of the project such as cloning, gene expression, protein purification, and crystallization trials.

The set of 1300 well-conserved gene was further reduced to 317 by only retaining the genes of *unknown* or *hypothetical* function. The distribution of conservation index (CI values) in this subset ranges from 17.2 to 1.4. Figure (3A) presents a graph of known *vs*. unknown genes ranked according to their evolutionary conservation (CI value). This graph indicates that the most conserved genes (likely to correspond to essential pathways) have been historically characterized in priority. However, it is comforting to see that a fraction of well-characterized genes are found at lower CI values where genes of yet unknown functions begin to dominate the distribution. This shows that our protocol of ORF selection based on the evolutionary conservation across Gram-positive/Gram-negative genomes does actually converge onto actual genes.

The set of 317 highly conserved anonymous genes was further reduced to enhance our probability of success in the experimental phase of the project, that is the over expression and determination of the 3-D structure of the largest possible number of these *E. coli* gene products. Although the structural determination of membrane-bound proteins will eventually be attempted later on (e.g. by expressing the ORF as separate domains), we chose to first focus on the

easier soluble targets (the so-called 'low hanging fruits'). A prediction of membrane spanning segments [11] was then performed to determine which of the above 317 genes were most likely to be expressed as soluble globular protein. This final bioinformatic filter resulted into a final set of 221 candidate target genes. The distribution of CI values within this gene set does not exhibit any significant bias, ranging from 17.2 to 1.4. (Figure 3B). Out of the 221 corresponding proteins, only 17 exhibit a strong sequence similarity (defined as >40% identity over a segment of at least 130 residues) to entries in the PDB. Our 221-candidate gene set is thus likely to correspond to a sizable number of interesting new structures, and possibly some original folds. 110 of these target genes were attributed to the AFMB laboratory and are not further described.

Additional bioinformatic analyses

Additional sequence analyses were performed at the various stage of the experimental process, in particular for genes and proteins for which difficulties were encountered. For instance, putative cofactor-binding motifs were searched to help expressing, purifying or crystallizing recalcitrant proteins. This was performed by combining the recognition of usual sequence signatures (e.g. Prosite motif [22]) with an analysis of their conservation within multiple alignments.

Multiple alignment is also useful to identify the proper N-terminus of the ORF (for cloning purpose) or delineating detrimental flexible N- or C-terminal segments of the candidate proteins. An analysis of the confidence measure at the heart of the T-coffee multiple alignment program [21] indicated a correlation between the least reliably aligned protein segments, the regions of higher sequence variability and those of greater flexibility [23]. As we work with well-conserved/ubiquitous genes, the situation is also optimal to take advantage of multiple alignments to identify key-residues in these 'anonymous' proteins, and use the conserved patterns to link them with protein families of known functions, despite a lack of overall significant sequence similarity (see 24 as an example).

Phylogenomics analyses, aiming at identifying subset of genes co-segregating or co-evolving across multiple bacterial genomes, are also extensively used to predict protein-protein interaction [25, 26] eventually leading to the design of co-expression, co-purification or co-crystallization experiments. Phylogenomics is used to suggest links between ORFs of unknown functions and identified pathways [25]. These hints might help the functional interpretation of the 3-D structures of our anonymous targets. Phylogenomics analyses are performed with our in-house software PhyDBac [26, http://igs-server.cnrs-mrs.fr/phydbac/].

On the negative side, the priority of candidate genes was sometimes downgraded due to their absence in important pathogens with reduced genomes (such as *H. influenzae* or *Mycoplasma*), or for being too similar to their human homologues. Ideally, only bacteria-specific biochemical pathways should be targeted for the development of antibiotics (e.g. penicillin blocking peptidoglycan biosynthesis). However, this constraint is probably not acceptable on the long run, given the constant need for new antibiotics and the limited number of strictly bacterial specific targets. A good counter example is the success of the quinolones that are targeting DNA gyrase, one of the most conserved proteins (26% identical amino acids between human and *E. coli*).

Finally, the amino acid conservation patterns derived from multiple alignments become even more informative when analyzed in the context of a representative 3-D structure (from the *E. coli* gene product). For instance, the structural (hydrophobic core or stabilizing residues) *vs.* functional (e.g. catalytic residues) nature of the conserved positions can be assessed (as illustrated in [20]). Paradoxical features – such as well-exposed surface loops exhibiting strong residue conservation –, are highly informative for the functional interpretation of anonymous 3-D structures. Protein cavities lined with conserved residues are also good candidates for 'active sites'. Previously unnoticed (i.e. misaligned) conserved isolated residues (dispersed in the sequence) might be found close together in the 3-D structure and suggest the presence of a catalytic center. The spatial distribution of conserved residues can be exhaustively searched within proteins of known function, eventually resulting in the identification of catalytic sites [27]. Finally, the drugability of the regions of the protein pointed out by these analyses can be assessed, and large libraries of putative ligands can be computationally screened [28, 29].

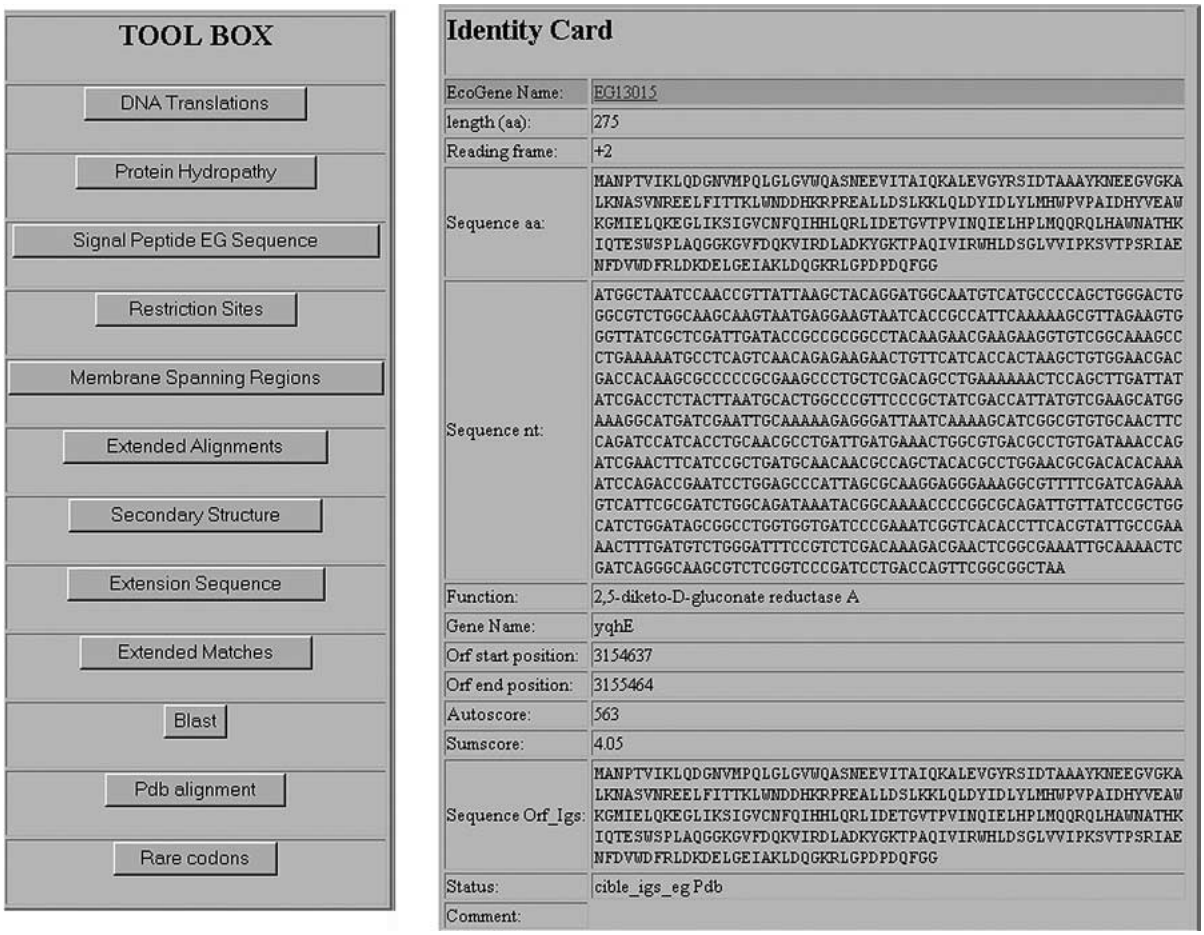

Figure 2. The yqhE gene identity card in Master DB. The various buttons (top left column) provide interactive access to pre-computed additional information as well as to 'on the fly' sequence analysis.

Gene expression and protein purification pipeline: First pass

The general principle governing our large-scale protein expression approach is to quickly identify the 'easy' ORFs by passing our whole set of 111 through a common first pass standard protocol. Further expression trials are then performed on the 'recalcitrant' proteins. These trials involve lowering the temperature, trying other *E. coli* strains, adding predicted co-factors, ..., etc. After about 17 months, the first pass has been completed on the 111 targeted ORFs. The results are as follows: 108 ORFs have been successfully cloned, 104 have been tested, 89 have exhibited a detectable expression (either *in vivo* or using the Roche cell free system), 54 in a soluble form in a first pass screening. In parallel, our industrial partner (Infectious Disease Division, Aventis Pharma) is assessing the essentiality of these putative target genes in major pathogens using proprietary approaches.

Second pass: addressing the bottleneck of soluble expression

The above results identified the expression of the targets in a *soluble form* as a major bottleneck. A second pass protocol was thus designed using a statistical screening of various variables in order to determine the best conditions for soluble expression. We first determined that the three main variables influencing protein solubility and/or expression yield were 1) the bacterial strain 2) the culture temperature and 3) the growth medium. A set of 12 conditions corresponding to the combination of these 3 variables was thus

Properties

MW:	31109.6
pI:	6.43
Cys #:	1
Met #:	5
Ext coef:	1.7139
aa +:	41
aa -:	37

Blast Results

Organism	Position	PValue	Score	Identity	Begin Q	End Q	Begin S	End S
Escherichia coli O157:H7	3964697_3965521	1e-160	563	0.99	1	275	1	275
Salmonella typhi	3181088_3181609	1e-89	327	0.95	1	174	1	174
Bacillus subtilis 168	3426653_3425814	2e-68	257	0.7	7	272	13	277
Mycobacterium tuberculosis H37Rv	3326098_3326943	2e-66	250	0.64	7	275	13	281
Bacillus subtilis 168	2969001_2969951	1e-65	247	0.64	9	273	48	315
Lactococcus lactis IL1403	333512_332610	3e-62	236	0.64	7	273	30	299
Streptococcus pyogenes	1406474_1405635	1e-59	228	0.63	5	271	4	276
Lactococcus lactis IL1403	1267992_1268831	9e-59	225	0.64	9	273	9	276
Lactococcus lactis IL1403	266240_267082	2e-58	224	0.62	9	272	9	278
Lactococcus lactis IL1403	1944439_1943597	8e-54	208	0.63	7	262	5	258
Escherichia coli O157:H7	359399_360280	9e-51	198	0.6	15	263	17	270
Salmonella typhi	1340834_1341700	2e-49	194	0.59	15	263	12	265
Salmonella typhi	292873_293694	4e-46	183	0.59	6	266	1	263
Escherichia coli O157:H7	232497_233297	5e-45	179	0.6	15	266	3	256
Pseudomonas aeruginosa PA01	4663510_4662695	9e-40	162	0.56	15	268	8	264
Escherichia coli O157:H7	353368_352796	6e-37	152	0.6	7	194	4	191
Salmonella typhi	3181688_3181912	7e-36	149	0.96	201	275	1	75
Salmonella typhi	1746728_1745874	4e-22	103	0.5	15	254	15	266
Escherichia coli O157:H7	2542848_2541997	2e-20	97	0.49	15	259	14	270
Pseudomonas aeruginosa PA01	881910_882776	1e-19	95	0.48	7	254	20	270
Borrelia burgdorferi B31	539275_538331	1e-17	88	0.52	13	192	9	198
Bacillus subtilis 168	1029745_1030737	6e-14	76	0.56	18	145	23	162
Pseudomonas aeruginosa PA01	1220664_1219624	6e-14	76	0.44	17	260	35	326

Figure 2. (Continued.)

defined using an incomplete factorial experimental design (see Materials and Methods). A typical example of such a screen applied to 6 different targets is given in Figure 4.

At this point, out of the 108 ORFs cloned, 94 have been passed through this screening, 93 have been successfully expressed, including 68 in a soluble form. The incomplete factorial approach demonstrated its power since out of the 94 ORFs screened, only 1 was not expressed and 25 failed the soluble expression assay. This systematic screening allowed 72.3% of the targets to be recovered in a soluble form. This is to be compared to the 52% yield obtained through the first pass standard protocol.

The scale up bottleneck

For a certain number of proteins, we observed that directly scaling up the culture from 4 ml to 1 liter could lead to insoluble expression. This problem was usually solved by simply replacing the 1-liter culture by several 200-ml cultures.

Protein purification

Protein purifications are not yet performed through an automatic procedure. The buffer most suitable for protein concentration is determined through dialysis experiments of the purified material, monitored by UV spectra and taking into account the theoretical isoelectric point of each protein. The systematic check for N-terminal sequence and mass of each purified protein, proved extremely useful for debunking various problems (including the mishandling of samples). We noticed frequent cases (15/31) of spontaneous removal of the extended His-tagged N-terminus inherent to the use of the GATEWAY system (17 to 24 first residues removed). The final statistics on this observation will only be available once all soluble proteins will have been purified and sequenced.

A prerequisite to successful crystallization was also the determination of appropriate storage conditions for each purified targets, using gel electrophoresis over time (see Materials and Methods). For unstable proteins, crystallization trials were per-

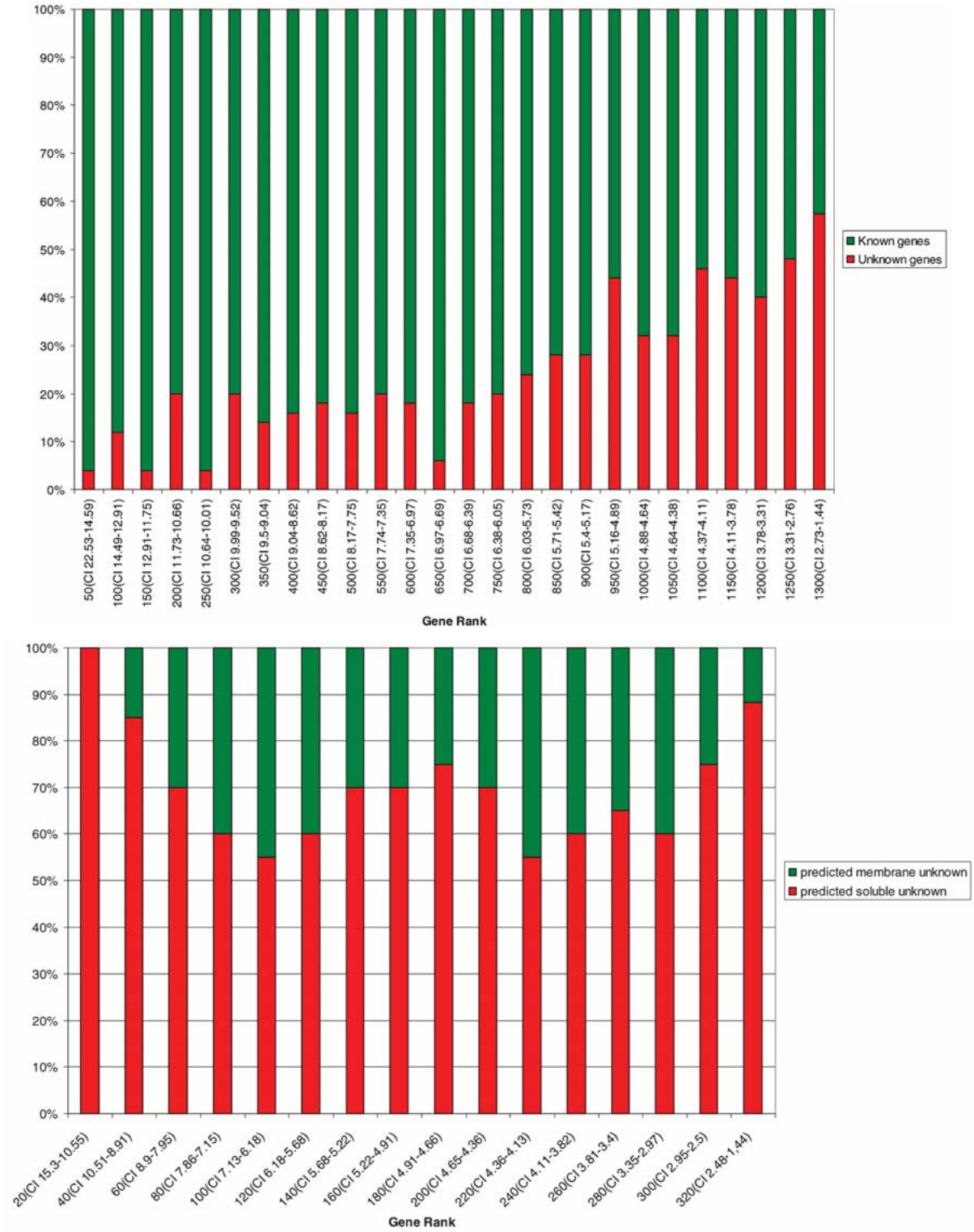

Figure 3. (A) Distribution of known vs. unknown genes according to their evolutionary conservation. The 1300 most-conserved E. coli genes are ranked according to their decreasing CI value (x-axis, in parenthesis) and grouped in successive boxes of 50 (Gene rank). For each box, the number of genes of unknown vs. known function are shown in red or green, respectively. Functionally characterized genes dominate the distribution for the high CI values, but many anonymous genes also exhibit comparable evolutionary conservations.
(B) Distribution of predicted membrane vs. soluble proteins according to their evolutionary conservation (decreasing CI value, x-axis, in parenthesis). The 317 highly conserved ORFs of unknown function have been divided into membrane-bound (green boxes) vs. soluble (red boxes) after sequence analysis. There is no overall bias, although proteins predicted soluble dominate the distribution of the most conserved candidates.

Table 2: Experimental matrix used for expression screening. The first column is the experiment number. The second column corresponds to the 4-state *strain* variable: BL21(DE3) pLysS, Rosetta, Origami and C41. The third column corresponds to the 3-state *growth medium* variable: 2YT, SB and TB media. The fourth column corresponds to the 3-state *temperature* variable: 25 °C, 37 °C and 42 °C.

Condition number	Strain	Medium	Temperature
1	Bl21 (DE3) pLysS	SB	25
2	Rosetta (DE3)	TB	25
3	Origami (DE3)	2YT	25
4	C41 (DE3)	SB	25
5	Bl21 (DE3) pLysS	TB	37
6	Rosetta (DE3)	SB	37
7	Origami (DE3)	TB	37
8	C41 (DE3)	2YT	37
9	Bl21 (DE3) pLysS	2YT	42
10	Rosetta (DE3)	2YT	42
11	Origami (DE3)	SB	42
12	C41 (DE3)	TB	42

formed on freshly purified samples, or using material stored at −80 °C.

Screening of alternative cloning contexts using linear PCR products

To quickly explore alternative cloning contexts (Tag type and position) for recalcitrant targets, we developed an *in vitro* expression screening based on linear templates, involving a two steps PCR and universal primers. Tested on 24 different *E. coli* ORFs, the linear templates mimicking pDEST17 vector context (N-terminal His-tag), performed as efficiently as the corresponding circular plasmid. This approach can thus be applied to modify the standart pDEST17 ORF context into a tag/target combination more suitable for soluble expression. Possible alternatives include C-terminal His-tag, GST, MBP, Thioredoxin, Strep tag, and GFP.

Overview of the structures obtained to date

YecD: molecular replacement using homology modeling

The interpretation of yecD diffraction data (collected on the ID29 beam line at the ESRF synchrotron in Grenoble-France) was first attempted by molecular replacement using AmoRe [17] with the 1NBA and 1IM5 PDB entries. These proteins share less than 25% identical residues with the yecD sequence and failed to produce a solution. We then build a yecD structure model with MODELER [30] using the known structures as templates. The yecD model was used for molecular replacement and produced a good solution with 4 molecules per asymmetric unit. The yecD structure was then refined to 1.3 Å (PDB 1J2R). YecD and 1NBA are annotated in Swiss-Prot as belonging to the isochorismatase protein family. However, the superposition of active site region of the yecD, 1NBA and 1IM5 structures reveals that a cysteine residue essential to the isochorismatase enzymatic activity is replaced by a glycine in yecD. In addition, yecD lacks a cis-peptide conserved in 1NBA and 1IM5. It is thus likely that yecD has a different enzymatic activity. Detailed sequence analyses are being performed to predict its function.

YggV: a nucleoside triphosphate phosphohydrolase?

The yggV structure was solved using the MAD technique with one seleno-methionine. Data were collected on the BM30 beam line at ESRF. The structure was refined to 1.8 Å and was found to be strongly similar to PDB entries 2MJP (Nucleoside triphosphate phosphohydrolase from *Methanocaldococcus jannaschii*) and 1EX2 (MAF protein from *Bacillus subtilis*). YggV crystals were thus grown in presence of dGTP. They diffract to 2.4 Å resolution. The structure of yggV was also solved in the absence of dGTP by the Midwest Center for Structural Genomics (PDB 1K7K). A multiple alignment of the homologous sequences is currently analyzed in the context of the various available structures (free forms and complexed with nucleotides). Our phylogenomic analyses pointed out putative partners to yggV, suggesting that its function in *E. coli* might not be entirely understood.

YhbO: an intracellular protease?

The yhbO structure was solved by molecular replacement using AmoRe with the PDB structure 1G2I (an intracellular protease from *Pyrococcus horikoshii*). The yhbO structure was subsequently refined to 2.0 Å resolution. A nucleophile elbow motif as well as a cysteine residue essential for the protease activity of the *P. horikoshii* intracellular protease is conserved in yhbO. The multimeric state of the 1G2I structure was also described as essential for the catalytic activity, the active site involving residues contributed by two monomers. Interestingly, the interface between the

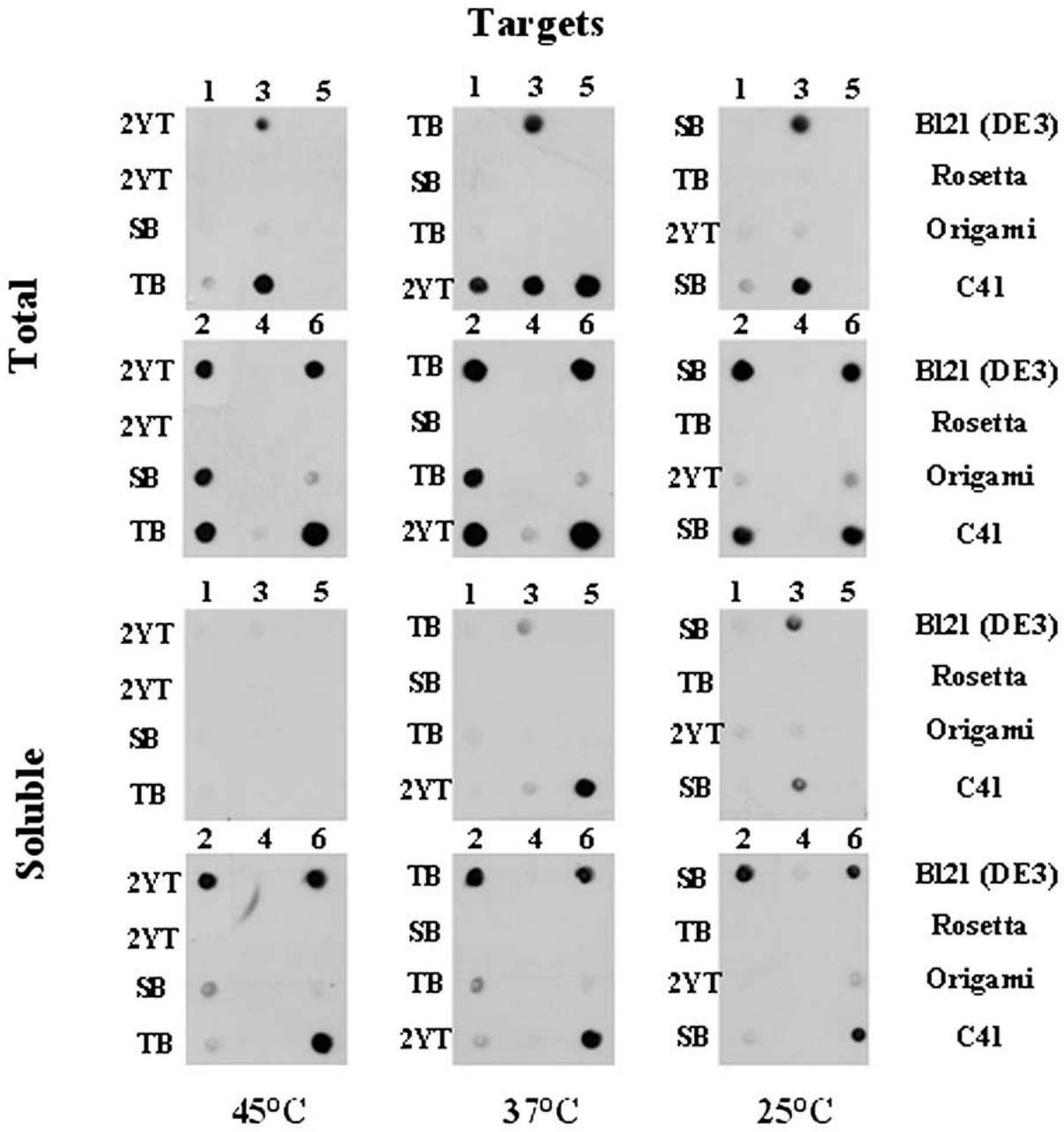

Figure 4. Dot Blot results of the expression screening on 6 different targets. The top 6 Dot blots correspond to the *total* protein expression results. The 6 lower Dot blots correspond to the *soluble* protein expression measurements. The numbers above each Dot Blot refer to the targets. Dot Blot results are ranked by decreasing temperature (left to right) and the strains indicated on the right. The growth medium is indicated on the left of each Dot Blot.

two yhbO monomers as seen in the crystal structure is entirely different from the one found in the 1G2I structure. The functional assignment of the *E. coli* yhbO protein is thus not yet certain. The putative protease activity of yhbO is currently studied.

YqhE: significant difference in the active sites.
The yqhE structure was solved by molecular replacement using AmoRe with the PDB entry 1A80 (a 2, 5-diketo-D-gluconic acid reductase A structure from *Corynebacterium*). The yqhE structure was then

Table 3: Project status. List and status of our targets as deposited in targetdb (http://targetdb.pdb.org) under the project name: SGI-SG (BIGS). The first column indicates the target name (according to EcoGene [10]), the second column indicates the relevant genome. The status (from 'Selected' to 'Crystal Structure″) of each target protein is shown in pink. For each target, *arrows* are used to indicates the next step to be completed, *stars* correspond to targets currently stopped, and *squares* correspond to 'on hold' targets (failure during the scale-up or crystallization processes) pending further testing.

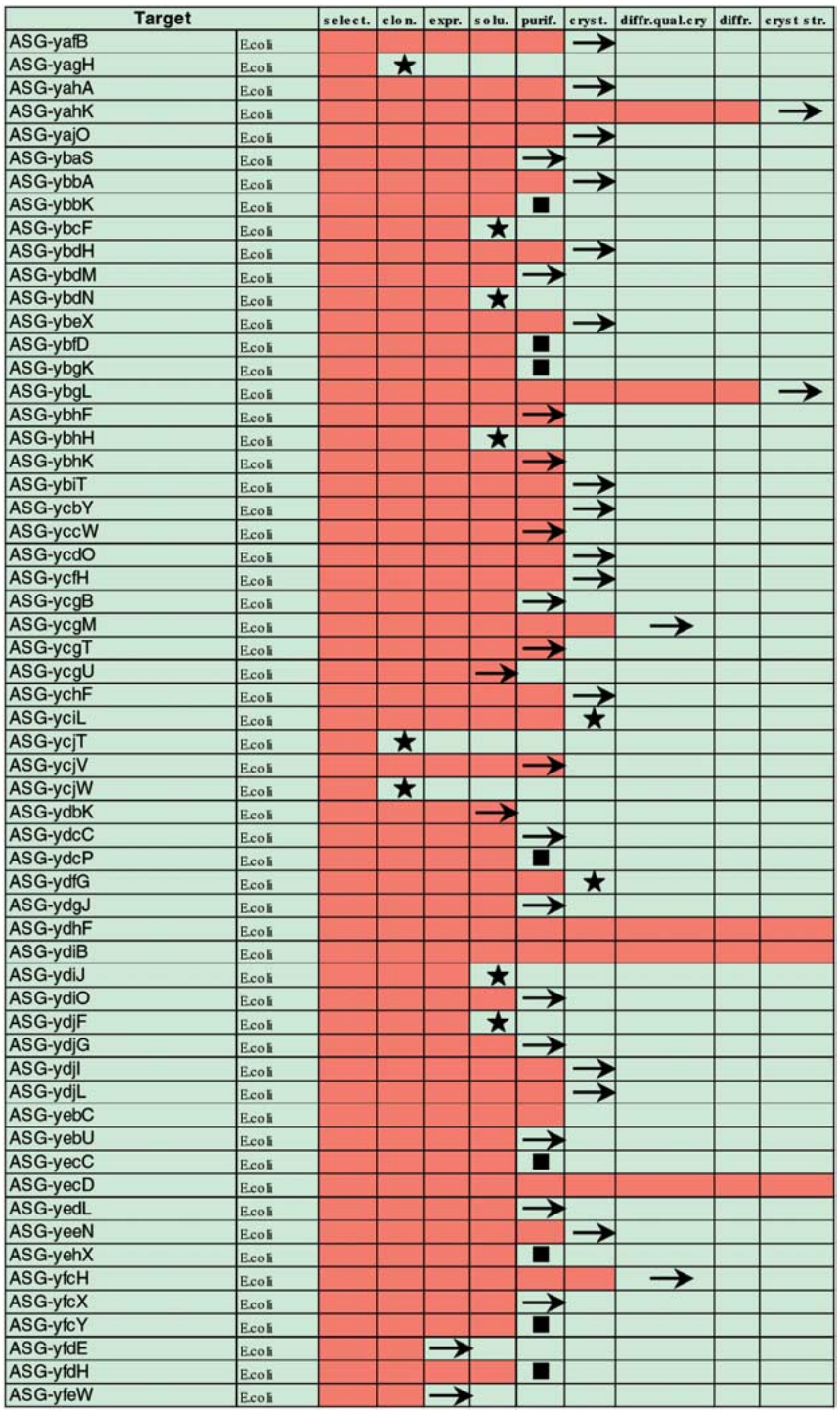

Table 3: (Continued).

Target		select.	clon.	expr.	solu.	purif.	cryst.	diffr.qual.cry	diffr.	cryst str.
ASG-yfgK	E.coli				★					
ASG-ygbJ	E.coli					■				
ASG-ygcA	E.coli				★					
ASG-ygcE	E.coli					→				
ASG-ygcT	E.coli					■				
ASG-ygcU	E.coli					→				
ASG-ygfF	E.coli				→					
ASG-ygfN	E.coli					→				
ASG-yggF	E.coli					→				
ASG-yggV	E.coli									
ASG-yggW	E.coli						→			
ASG-yghZ	E.coli						→			
ASG-ygiS	E.coli			→						
ASG-ygiX	E.coli						→			
ASG-ygjD	E.coli						→			
ASG-yhbG	E.coli					→				
ASG-yhbO	E.coli									
ASG-yhbZ	E.coli						→			
ASG-yhdG	E.coli				★					
ASG-yhdZ	E.coli				→					
ASG-yheS	E.coli					→				
ASG-yhfR	E.coli					→				
ASG-yhiN	E.coli					■				
ASG-yiaY	E.coli				→					
ASG-yibK	E.coli					→				
ASG-yicC	E.coli					→				
ASG-yicP	E.coli					→				
ASG-yieK	E.coli					→				
ASG-yifB	E.coli				★					
ASG-yjcS	E.coli			→						
ASG-yjeQ	E.coli						→			
ASG-yjeS	E.coli					■				
ASG-yjfH	E.coli					→				
ASG-yjjK	E.coli						→			
ASG-yliA	E.coli				★					
ASG-yliB	E.coli									
ASG-yliK	E.coli					→				
ASG-yneJ	E.coli					■				
ASG-ynfF	E.coli				★					
ASG-ynfL	E.coli				★					
ASG-yodA	E.coli			★						
ASG-yohF	E.coli							→		
ASG-yojH	E.coli				★					
ASG-ypdE	E.coli					→				
ASG-ypdF	E.coli					→				
ASG-ypfJ	E.coli				★					
ASG-yphE	E.coli				→					
ASG-yqeF	E.coli						→			
ASG-yqhE	E.coli									
ASG-yraL	E.coli						→			
ASG-yrbF	E.coli				★					
ASG-ytfR	E.coli				★					
ASG	111	111	108	103	84	46	12	9	9	7

refined to 1.93 Å resolution (PDB 1MZR). YqhE belongs to the vast NADPH-dependent aldo-keto reductase family. Multiple alignment was used to compare the various sequences of the family. This revealed two sub-families, including all eukaryotic sequences on one hand, and all the prokaryotic sequences on the other hand. Structurally, the eukaryotic proteins differ from the prokaryotic ones by the presence of much longer loops in the co-factor and ligand binding regions. These marked structural differences should make possible the design of small inhibitory molecule with a good specificity for the prokaryotic targets.

YdhF: an NADPH-dependent oxidoreductase of unknown specificity

The ydhF structure was solved using the MAD technique with 8 seleno-methionines. The structure was then refined to 2.5 Å resolution using data collected on the BM30 beam line at ESRF. YdhF belongs to the NADPH-dependent oxidoreductase family. Its structure was compared to the other members of the family using multiple alignment. YqhE shares 26.7% identical residues over a 217-aa segment with its closest structural homologue ydhF . YdhF exhibits a sequence insertion in a flexible loop not well resolved in the structure. This difference is not located near the co-factors and ligand binding sites, but may still affects the enzyme specificity. The structure of the NADP containing ydhF is currently under refinement using data to 2.8 Å resolution.

YliB: a putative n-peptide binding protein?

The yliB structure was solved using the MAD technique with 20 seleno-methionines. The structure was then refined to 2.3 Å resolution using data collected on the ID14-EH2 beam line at ESRF. This membrane-anchored periplasmic protein shares 29% identical residues with 1DPE (a dipeptide binding protein from *E. coli*) and 1DPP (the same protein in complex with glycyl-L-leucine). The loops surrounding the ligand-binding pocket are shorter in the yliB structure compared to the homologous structures, suggesting that larger peptides might be accommodated by yliB. We also noticed the presence of zinc ion, and of some residual electron density in the region homologous to the 1DPP dipeptide binding site. We are currently producing crystals of yliB in the presence of various dipeptide in order to verify its affinity for such substrates.

YdiB: a new structural fold?

The ydiB structure was solved using the MAD technique with 20 seleno-methionines. The structure building is still in progress using data collected at 2.66 Å resolution on the BM30 beam line at ESRF. The preliminary tracing suggests that the ydiB structure might be an original fold which is confirmed by the deposition in the PDB of the ydiB structure by the Northeast Structural Genomics Research Consortium (PDB 1NPD) and the Montreal-Kingston Bacterial Structural Genomics Initiative (PDB 1O9B).

Conclusion

This structural genomics project is an opportunity to develop new approaches for the various stages of gene expression. The use of the 'one tube reaction' GATEWAY technology allowed the successful cloning of 97.2% of the selected ORFs. The development of both *in vitro* and *in vivo* expression screening with the efficient sampling of the various parameters influencing soluble protein expression, allowed 80% of the tested ORFs to be expressed as soluble proteins (an additional 18% being expressed as inclusion bodies). The automation of the screening using dot-blot revelation now allows us to process 12 targets weekly.

In our hands, the purification step is not a limiting step since 100% of the 46 proteins that have reached that stage were successfully purified. However, it was more challenging to keep them intact and soluble during the following steps of concentration and buffer exchange required for crystallization. The introduction of buffer screening through dialysis controlled by UV scanning has proven very effective, allowing the identification of a suitable buffer for 95% of the purified samples.

Finally, the quality control we imposed through the entire project, albeit time consuming, has proven useful to trace mistakes such as target mixing and/or cross-contamination, as well as protein degradation. On a more positive side, it also provided information on unexpected protein/protein interactions and/or ligand binding, eventually leading to improved crystallization protocols. To this days, 40 of the 46 purified recombinant proteins have been through biophysical characterization (circular dichroïsm spectroscopy and monodispersity assays using dynamic light scattering), 32 have entered crystallization trials and 31 have been sequenced. Twelve proteins have produced crystals, out of which 9 have resulted in us-

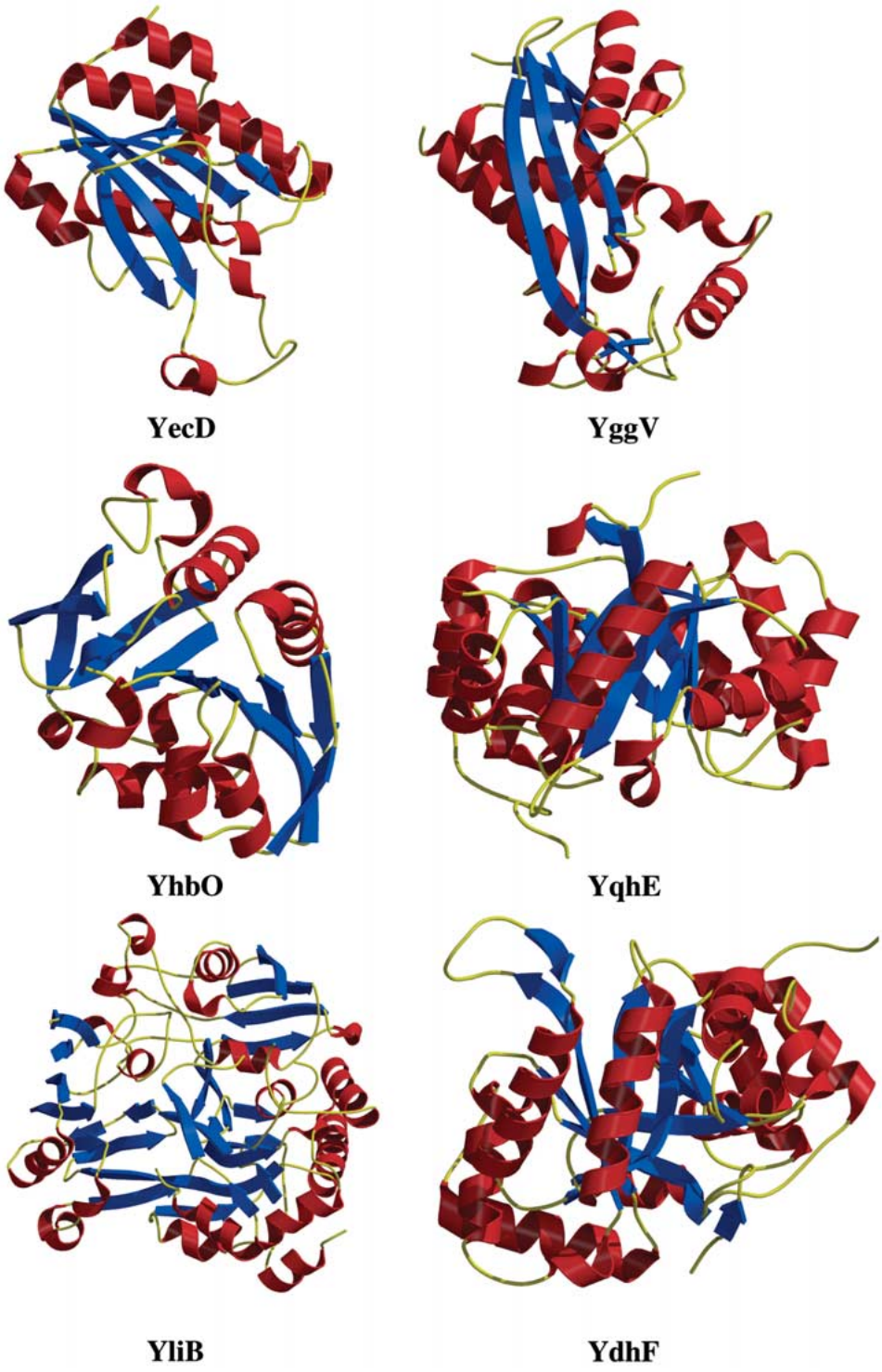

Figure 5. Protein structures solved in this study. The structures are drawn using MOLSCRIPT [31]. The YdiB structure was also solved by the consortium.

able diffraction data, allowing 7 structures to be solved (Figure 5). This number is now expected to grow rapidly, as more expressed gene products are reaching the purification/crystallization stages.

Acknowledgements

We thank our collaborators from Aventis Pharma: Drs V. Blanc, E. Remy, M. Black and J. Hogdson for their scientific input. We also thank Drs. J-L. Ferrer, F. Borel, P. Carpentier, A. Thompson, G. Leonard and J. McCarthy for their help on the FIP, BM14, ID29, ID14 ESRF beamlines. We thank P. Pham-Trong (Greiner) for his help and gift of materials. This project is funded by a grant (#014906055) from the French Ministry of Industry 'post genomics research program'.

References

1. Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., *et al.* (2000) *Nature*, **406**, 959–964.
2. Cassell, G.H., and Mekalanos, J. JAMA 285, (2001) 601–605.
3. Projan, S. (2002) *Curr. Opin. Pharmacol.* **2**, 513–522.
4. Clements, J.M., Beckett, R.P., Brown, A., Catlin, G., Lobell, M., Palan, S., Thomas, W., Whittaker, M., Wood, S., Salama, S., Baker, P.J., Rodgers, H.F., Barynin, V., Rice, D.W., and Hunter, M.G. (2001) *Antimicrob. Agents Chemother.* **45**, 563–570.
5. Barbachyn, M.R., Brickner, S.J., Gadwood, R.C., Garmon, S.A., Grega, K.C., Hutchinson, D.K., Munesada, K., Reischer, R.J., Taniguchi, M., Thomasco, L.M., Toops, D.S., Yamada, H., Ford, C.W., and Zurenko, G.E. (1998) *Adv. Exp. Med. Biol.*, **456**, 219–238.
6. Johnson A.P. (2001) *Curr. Opin. Investig. Drugs*, **12**,1691–1701
7. Auckland, C., Teare, L., Cooke, F., Kaufmann, M.E., Warner, M., Jones, G., Bamford, K., Ayles, H., Johnson, A.P. (2002) *J. Antimicrob. Chemother.* **50**, 743–746.
8. Claverie, J.M. *Nature*, **2000**, 403, 12–12.
9. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) *Nucleic Acids Res.* **25**, 3389–3402.
10. Rudd, K.E. (2000) *Nucleic Acids Res.* **28**, 60–64.
11. Persson, B., and Argos, P. (1997) *J. Protein Chem.* **16**, 453–457.
12. Hartley, J.L., Temple, G.F., and Brasch, M.A. (2000) *Genome Res.* **10**, 1788–1795.
13. Monchois V., Vincentelli, R., Deregnaucourt, C., Abergel, C., and Claverie J.-M. (2002) In *Cell-Free Translation Systems* (Ed., Spirin, A.S.), Springer Verlag, New York, pp. 197–202.
14. Audic, S., Lopez, F., Claverie, J.-M., Poirot, O., and Abergel, C. (1997) *Proteins* **29**, 252–257.
15. Monchois, V., Abergel, C., Sturgis, J., Jeudy, S., and Claverie, J.-M. (2001) *J. Biol. Chem.* **276**, 18437–18441.
16. Bhat, T.N., Bourne, P., Feng, Z., Gilliland, G., Jain, S., Ravichandran, V., Schneider, B.,Schneider, K., Thanki, N., Weissig, H., Westbrook, J., and Berman, H.M. (2001) *Nucleic Acids Res.* **29**, 214–218.
17. Navaza, J. (2001) *Acta Crystallogr.* **D57**, 1367–1372.
18. Liu, Y., Ogata, C.M., and Hendrickson,W.A. (2001) *Proc Natl Acad Sci U S A* **98**, 10648–10653.
19. Hendrickson, W.A., Horton, J.R., and LeMaster, D.M. (1990) *EMBO J.* **9**, 1665–1672.
20. Abergel. C., Bouveret, E., Claverie, J.M., Brown, K., Riga, A., Lazdunski, C., and Benedetti, H. (1999) *Structure Fold. Des*. **7**,1291–1300.
21. Notredame, C., Higgins, D.G., Heringa, J. (2000) *J. Mol. Biol.* **302**, 205–217.
22. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofmann, K., and Bairoch, A. (2002) *Nucleic Acids Res.* **30**, 235–238.
23. Notredame, C., and Abergel, C. (2003) In *Bioinformatics and Genomes* (Andrade, M.A., ed.) Horizon Scientific Press, Wymondham,UK, pp 27–49.
24. Abergel, C., Robertson, D.L., Claverie, J.-M. (1999) *J. Virol.* **73**, 751–753.
25. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D. (1999) *Nature* **402**, 83–86.
26. Enault, F., Suhre, K., Poirot, O., Abergel, C. and Claverie, J-M. (2003) *Nucleic Acids Res.*, In press.
27. Madabushi, S., Yao, H., Marsh, M., Kristensen, D.M., Philippi, A., Sowa, M.E., and Lichtarge, O. (2002) *J. Mol. Biol.* **316**, 139–154.
28. Klebe, G. (2000) *J. Mol. Med.* **78**, 269–281.
29. Apostolakis, J., and Caflisch, A. (1999) *Comb Chem High Throughput Screen.* **2**, 91–104.
30. Sali, A., and Blundell, T. L. (1993) *J. Mol. Biol.* **234**, 779–815.
31. Kraulis, P.J. (1991) *J. Appl. Crystallogr.* **24**, 946–950.