Nucleic Acids Research, 2000, Vol. 28, No. 16 **3075–3082**

Proteomics of *Mycoplasma genitalium*: identification and characterization of unannotated and atypical proteins in a small model genome

Suganthi Balasubramanian[1], Tamara Schneider[1], Mark Gerstein[1,2] and Lynne Regan[1,3,*]

[1]Department of Molecular Biophysics and Biochemistry, [2]Department of Computer Science and [3]Department of Chemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06520-8114, USA

Received May 16, 2000; Revised and Accepted June 28, 2000

ABSTRACT

We present the results of a comprehensive analysis of the proteome of *Mycoplasma genitalium* (MG), the smallest autonomously replicating organism that has been completely sequenced. Our aim was to identify and characterize all soluble proteins in MG that are structurally and functionally uncharacterized. We were particularly interested in identifying proteins that differed significantly from typical globular proteins, for example, proteins which are unstructured in the absence of a 'partner' molecule or those that exhibit unusual thermodynamic properties. This work is complementary to other structural genomics projects whose primary aim is to determine the three-dimensional structures of proteins with unknown folds. We have identified all the full-length open reading frames (ORFs) in MG that have no homologs of known structure and are of unknown function. Twenty-five of the total 483 ORFs fall into this category and we have expressed, purified and characterized 11 of them. We have used circular dichroism (CD) to rapidly investigate their biophysical properties. Our studies reveal that these proteins have a wide range of structures varying from highly helical to partially structured to unfolded or random coil. They also display a variety of thermodynamic properties ranging from cooperative unfolding to no detectable unfolding upon thermal denaturation. Several of these proteins are highly conserved from mycoplasma to man. Further information about target selection and CD results is available at http://bioinfo.mbb.yale.edu/genome

INTRODUCTION

The avalanche of DNA sequence information being generated allows us to perform a comprehensive analysis of the entire proteome of an organism. Many open reading frames (ORFs) that are identified in a particular organism have structurally or functionally well-characterized homologs in other organisms and therefore it is possible to assign their fold or function on the basis of sequence comparison alone. Nevertheless, analysis of several genomes indicates that for a large fraction of ORFs, neither structure nor function can be assigned by sequence comparisons. For these ORFs there are either no known homologs or if the homologs are known, they do not as yet have structural or functional annotations.

There is a tremendous effort underway to determine 3D structures of proteins that cannot be structurally or functionally characterized by sequence homology methods alone (1–9). The motivations behind large-scale structural genomics are manifold. All currently known protein structures have been categorized into a few hundred folds. It has been estimated that the complete protein fold space encompasses ~1000–10 000 different folds (10–14). Large-scale structural genomics will help determine all the folds in the basic parts list (15,16). The entire parts list will enable us to take a global, unbiased view on protein structures and their physical properties. Proteins in the current structure databank are not representative of those in a complete genome because the structures solved to date have been influenced by the individual researchers' specific needs or by the ease of obtaining protein samples amenable for crystallization or nuclear magnetic resonance (NMR) methods (17). An unbiased representative sampling of all proteins is essential to obtain a comprehensive database of all existent folds. Three-dimensional (3D) structure determination of proteins is a key element in the assignment of protein function because the spatial arrangement of amino acids gives a more accurate representation of the active site than comparisons based on one-dimensional primary sequence alignments.

With the above-mentioned goals in mind, several groups have adopted different approaches to study uncharacterized proteins. These are enumerated below.

(i) High throughput expression of proteins for mass production and 3D structure determination of the most easily obtainable proteins, the so-called 'low-hanging' fruits (18).

(ii) Classification of all sequences from different organisms into different categories based on sequence homology and then determining the 3D structure of one member from each class. This class-directed initiative aims to assign the 3D structure of every ORF sequenced by solving the structure of one representative example from each class (19,20).

(iii) 3D structure elucidation of proteins that have functional annotation but are of unknown structure in order to better understand the relationship between structure and function.

*To whom correspondence should be addressed. Tel: +1 203 432 9843; Fax: +1 203 432 5175; Email: lynne@csb.yale.edu

(iv) Complete structure determination of proteins which have no functional annotation and are of unknown structure (21). The idea behind this approach is the identification of function based on structural homology that was not discernable from sequence comparisons alone.

Our approach towards the study of uncharacterized proteins is complementary to other structural genomics projects. We have biophysically characterized a set of proteins in order to differentiate between folded stable proteins and atypical proteins rather than focusing only on structured proteins. We are particularly interested in proteins that are not structured on their own and proteins with unusual biophysical properties. There are several examples of proteins that become structured in the presence of their interacting partners (22–24). In addition, examples of several proteins that are unfolded or disordered in their native state are known where lack of regular structure is essential for their biological function (25,26). It has been shown computationally that ~15 000 proteins in the SwissProt database contain disordered regions based on neural network prediction methods (27). In addition, a theoretical study elegantly demonstrates the importance of disordered regions for fine tuning and modulating the affinity and specificity of molecular interactions (26). Structure determination of largely unstructured proteins tend to be difficult because they are hard to crystallize and generally not very amenable to NMR. This may be due to many factors such as aggregation and degradation by proteases. Consequently, such proteins will be underrepresented in the current structural genomics efforts. Our studies aim to include proteins with unstructured regions which are not very easily amenable to high resolution 3D structure determination methods.

We have used an objective criterion for target selection in a systematic study of the uncharacterized proteins of *Mycoplasma genitalium* (MG). Our aim is to identify and characterize all functionally and structurally uncharacterized (ORFs whose fold cannot be assigned by sequence homology methods) soluble proteins in the proteome of MG, the smallest autonomously self-replicating organism for which the entire genome sequence has been determined (28). The small size permits a comprehensive and experimentally tractable analysis of the entire proteome. For this reason, MG has also attracted considerable interest in studies which aim to define the minimal gene set that represents all genes necessary and sufficient for a free living organism. The proteomes of two different ancient lineage, MG (a gram-positive bacteria) and *Haemophilus influenzae* (a gram-negative bacterium) have been compared, and proteins present in both have been proposed to represent the minimal gene set (29) with the assumption that genes conserved over a long evolutionary distance are likely to be essential for cellular function. MG has also been investigated experimentally in studies which aim to delineate the set of genes that are essential for viability (30). In these studies, transposon mutagenesis was used to introduce random insertions into the entire MG genome. Location of transposon insertions were identified by sequencing viable cell cultures obtained on defined growth media. The presence of numerous insertions within a gene suggests that the protein is not essential under the defined laboratory growth condition. Conversely, genes which contain no disruptive transposon insertions are probably essential for cellular growth. From this study, it is estimated that of the 483 genes in MG, about 265 are essential. Of these 265 essential genes, 100 are of unknown function. Interestingly, an analysis of distribution of the predicted fold assignments in MG among the various functional categories show that ORFs with no functional annotation constitutes the largest category of soluble proteins which cannot be assigned to a known fold (31). These observations underscore the importance of studying uncharacterized ORFs to enhance our understanding of protein structure and function. In undertaking such a study, we have used circular dichroism (CD) as a rapid means by which to screen the secondary structure and thermodynamic properties of ORFs that are both structurally and functionally uncharacterized.

MATERIALS AND METHODS

Bioinformatics

For the PSI-blast analyses, we ran all PDB domains in SCOP 1.39 (excluding coiled-coils and small Cys-rich proteins) against MG embedded in NRDB (which had been masked in the default fashion by SEG; 32–35). These comparisons used 20 iterations, an inclusion threshold into the matrix of 0.0005 and an overall match cutoff of 0.0001. Further matches were identified by running PSI-blast in two-way fashion plus pre-clustering the ORFs in MG. By 'two-way', we mean that the PDB was first run against MG embedded in NRDB and then unmatched regions of MG were cut out and run against the PDB embedded in NRDB. The pre-clustering was done with GEANFAMMER (36). The putative membrane proteins were identified as ORFs which contain segments of at least 20 residues with an average GES hydrophobicity less than –1 kcal/mol (37) in a protein that had at least one TM-segment with an average hydrophobicity less than –2 kcal/mol. Low complexity regions were identified using the program SEG with a trigger complexity K(1) of 3.4, an extension complexity K(2) of 3.75, and a window of length 45 (32,34). Signal sequences were identified based on their having a pattern of a charged residue within the first seven residues at the N-terminus followed by a stretch of 14 hydrophobic residues.

The *Escherichia coli* (EC) and *Bacillus subtilis* (BS) homologs were identified from single-sequence BLAST (38) using an MG query against the EC and SWISSPROT database at NCBI and the BS database at http://genolist.pasteur.fr/SubtiList/ . Two sequences were considered to be homologous if the *E*-values were $<1 \times 10^{-4}$.

It is important to realize that the state of the databases is constantly changing. For instance, when we began the project the MG ORF file available from TIGR had 479 ORFs, whereas the current one has 483. To maintain consistency, all of our analysis is based on a particular snapshot of this process: as of 15 February 1999.

Experimental methods

The 14 target genes were cloned into a vector containing a His_6 tag and a TEV-protease cleavage site, pPROEXHT (Gibco BRL, Rockville, MD). This vector was chosen for ease of cloning using the multiple cloning sites and for one-step purification applicable to all the proteins using the affinity of the hexahistidine tail to Talon (Clontech, Palo Alto, CA), Co(II)-column. The targets were cloned using genomic DNA as the template for PCR reactions. In the case of the EC

homologs, genomic DNA was isolated from cell cultures of *Escherichia coli* strain MJ1655 provided by the *E.coli* Genetic Stock Center (Yale University, New Haven, CT). The BS homologs were cloned by PCR amplification of the genes using BS genomic DNA: strain 168 as template and MG homologs were cloned using genomic DNA from MG strain G37 as template. The genomic DNA from these specific strains were used because they have been completely sequenced (28,39,40). The EC and MG homologs were cloned between the *Nco*I and *Hin*dIII sites and the BS homologs were cloned between the *Bam*HI and *Xba*I sites in pPROEX-HT vector. Appropriate restriction sites were added to the primers for insertion of PCR product into the plasmid. The N-terminal methionine in the target proteins were not included in the primers because a methionine is already present in the plasmid as part of the His_6 tag. Thus, the EC and MG target proteins included a 27-residue overhang at the N-terminus, MSYYHHHHHHDYDIPTTENLYFQGAMG, and the BS proteins had a 28-residue overhang at the N-terminus, MSYYHHHHHHDYDIPTTENLYFQGAMGS. All the clones were sequenced at the Keck Foundation Biotechnology research Foundation at Yale University. 332-EC has an asparagine to threonine change at position 9 with respect to the published sequence (40) and 448-BS has a threonine to proline change at position 20 with respect to the published sequence (39).

EC cell cultures were grown either at 37, 30 or 26°C in LB media. The cells were induced around an OD_{600} of 0.6 with 1 mM IPTG unless explicitly stated otherwise in the text. Cells grown at 37°C were harvested 2.5 h after induction and cells grown at lower temperatures were harvested 4–5 h after induction.

Proteins were purified both in the presence and absence of phenylmethanesulfonyl fluoride (PMSF, a protease inhibitor). Addition of PMSF did not seem to have an impact on the yield of purified protein. In some cases, protein degradation was occasionally observed after prolonged storage periods (3–6 months) and hence all the experiments were performed with freshly purified protein. Proteins were purified using a Talon, Co(II)-affinity, column. Proteins were eluted from the Talon column with 100 mM imidazole. After elution from the Talon column, samples purified from the soluble fraction were extensively dialyzed into 20 mM phosphate at pH 7.0: a physiologically relevant pH or 20 mM phosphate at pH 8.0 in case of solubility problems at pH 7.0. Proteins from inclusion bodies were similarly purified under denaturing conditions and refolded with 10 mM DTT in the appropriate dialysis buffer.

CD experiments were performed on an AVIV 62DS spectropolarimeter (Aviv Instruments, Lakewood, NJ). The concentration of samples used for CD experiments, as measured by A_{280}, varied between 5 and 20 μM. Wavelength scans were obtained from 190 to 260 nm at 4°C with an averaging time of between 5 and 20 s. The unfolding of the proteins was monitored by thermal denaturation studies. Ellipticity at 222 nm was followed as temperature was increased. DTT (2 mM) was added to all samples containing cysteines at the beginning of a thermal melt. Unfolding was monitored from 4 to 96°C in steps of 1 or 2°C. The samples were equilibrated at each temperature for 1 min and the signal was averaged over 30 s.

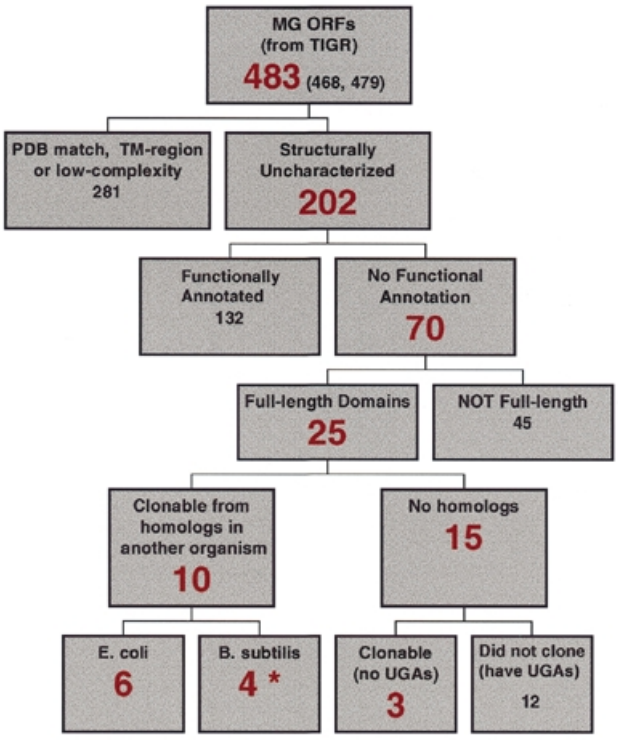

Figure 1. Schematic depiction of the target selection strategy. The numbers in parentheses in block 1 illustrate the dynamic nature of databases which get updated as more information is gathered. The numbers indicate the changing nature of the ORF file with time. *, Number of BS homologs cloned not including the homolog of MG448 which was cloned both from EC and BS.

RESULTS

Target selection

Our target selection strategy is schematically illustrated in Figure 1 and consists of the following steps.

(i) Sequence masking to identify structurally uncharacterized domains. To generate the initial target list, we used a sequence 'masking procedure' which has been described in detail elsewhere (17,41). Briefly, we identified and removed regions of sequence with homology to known structures using standard approaches such as FASTA (42) and PSI-blast (43). We then generated a consensus list from all the structural matches reported by various groups and our set of matches (41,44–48). We removed all putative transmembrane (TM) regions and regions containing long stretches of low sequence complexity. Finally, we removed hydrophobic signal sequences and short peptides (<80 residues) which link segments of sequence already accounted for by the PDB homology matches, TM-helices and low complexity regions.

(ii) Identification of 25 structurally uncharacterized ORFs with no known function. The regions remaining after the masking procedure outlined above are annotated as structurally uncharacterized. In total, we found 202 ORFs containing at least one uncharacterized region. We used the functional annotations from TIGR MG database (28) to identify ORFs with no functional annotation from this set of 202 ORFs. Of the

Table 1. List of the target proteins and their characteristics

ORFID	Length	Source	Cloned	Expression	Fraction at 37°C	Purification	2° Structure	Thermal melts	Minimal	Essential
9	262	EC	Yes	++	IB	Sol (0.1mM IPTG)	H	No	Yes	No
17	176	MG	Yes	++	IB	IB	H	No	No	No
56	277	EC	Yes	-	-	-	-	-	Yes	Yes
134	100	EC	Yes	+	Sol	Sol	H	Yes	No	Yes
208	196	MG	Yes	+	IB+Sol	Sol	H	Yes	No	No
221	154	EC	Yes	++	IB+Sol	Sol [1]	H+C	No [2]	Yes	Yes
240	292	BS	Yes	++	IB+Sol	Sol	H	Yes	No	Yes
296	129	MG	Yes	++	IB	Sol (25°C)	H	Yes	No	No
332	239	EC	Yes	++	IB+Sol	Sol	H	Yes	Yes	Yes
352	166	BS	Yes	++	IB	Sol (30°C)	H	Yes	No	NO
369	557	BS	Yes	++	IB+Sol	Sol	H+C	Yes	No	Yes
448	150	BS	Yes	++	IB+Sol	Sol	C	-	Yes	Yes
448	150	EC	Yes	++	IB+Sol	Sol	C	-	Yes	Yes
461	425	BS	No	-	-	-	-	-	No	Yes

The first column corresponds to the MG ORF #. The second column denotes the number of residues in the MG gene, Source indicates the organism from which the corresponding homolog of MG was cloned. Symbols: ++ indicates good expression, + moderate expression and – for no expression. Column 6 indicates if the expressed protein is in the soluble phase (Sol) or inclusion body (IB) at 37°C. In column 7, the growth conditions under which the expressed proteins were obtained in the soluble phase is indicated in parentheses. H, helix; C, Coil. In column 9, 'Yes' indicates that a thermal transition was observed and 'No' indicates that the protein did not unfold on thermal denaturation, i.e. the secondary structure did not change with temperature except in the case of 221-EC as indicated below. The column with the heading 'Minimal' indicates if the protein belongs to the proposed minimal gene set (29) and the column with the heading 'Essential' tells if the protein was found to be essential from transposon mutagenesis experiments (30).

[1]221-EC was purified under denaturing conditions even though it was expressed in the soluble fraction because it did not stick to the Talon column under native conditions.

[2]An increase in ellipticity was observed with increase in temperature for 221-EC. This could be an artifact due to improper refolding of 221-EC purified under denaturing conditions, presumably due to aggregation.

202 structurally uncharacterized ORFS, 70 proteins have no functional annotation and 132 proteins have a functional annotation. The group of 132 proteins are more likely to have a globular fold with properties of a typical protein than the first group of proteins because they have a known function. In our search for atypical proteins, we therefore focused our attention on the 70 proteins without known function. We divided these into two categories: in the first (25 proteins), the entire ORF is structurally uncharacterized; and in the second (45 proteins), only part of an ORF is uncharacterized, but structural assignments exist for other regions. We targeted the first group of 25 ORFs, representing full-length proteins with no homologs of known structure and no known function, for expression and purification. A full list of MG proteins with complete annotation describing their target-selection status is available at http://bioinfo.mbb.yale.edu/genome

Problem with tryptophan codon usage in MG

At this stage, it was necessary to consider a practical problem associated with expression of proteins in EC directly from the genomic DNA of MG. In MG, one of the codons for tryptophan is UGA (49). In most other organisms, including EC, UGA is a translational stop codon. Therefore, MG proteins containing this codon would be prematurely terminated when expressed in EC or other hosts. To circumvent this problem, we identified homologs of the MG genes of interest in EC or BS and cloned and expressed them from the genomic DNA of these organisms.

Homologs in EC were the first choice because this is the most commonly used organism for cloning and expression of proteins. If homologs were not present in EC, we searched for homologs in BS, the paradigm gram-positive bacterium. Finally, we divided the target genes into four groups according to their phylogenetic occurrence (Fig. 1): group I consisting of six EC homologs; group II consisting of four (five) BS homologs; group III consisting of three MG proteins; and group IV consisting of the remaining 12 proteins which had no homologs in EC or BS. [The BS homolog of MG448 was cloned first and all the colonies that were sequenced had a consistent threonine to proline change at position 20 with respect to the published sequence (39). Proline is a helix breaker and could have a deleterious effect on the structure of proteins and therefore we cloned the EC variant also. Therefore, the numbers in parentheses indicate the one extra protein that was cloned in both EC and BS. From now on, the total number of targets includes this extra protein.] We were able to clone the three group III MG proteins directly from MG genomic DNA although they had no homologs in EC or BS because they either contained no tryptophan or only had tryptophans with the commonly used UGG codon. The twelve group IV MG proteins had the alternate UGA Trp codon and were left out of this analysis because 13 (14) of the 25 full-length target proteins should provide a good representative sample. The targets include proteins that are highly conserved amongst many organisms and proteins unique to mycoplasma. Interestingly, many of these proteins are included both in the proposed minimal gene set as well as the experimental set of essential genes as shown in Table 1. The selected target proteins are denoted by the MG ORF number followed by the organism from which the corresponding homolog of MG was cloned. The proteins cloned directly using MG genomic DNA are denoted as MG followed by the ORF number.

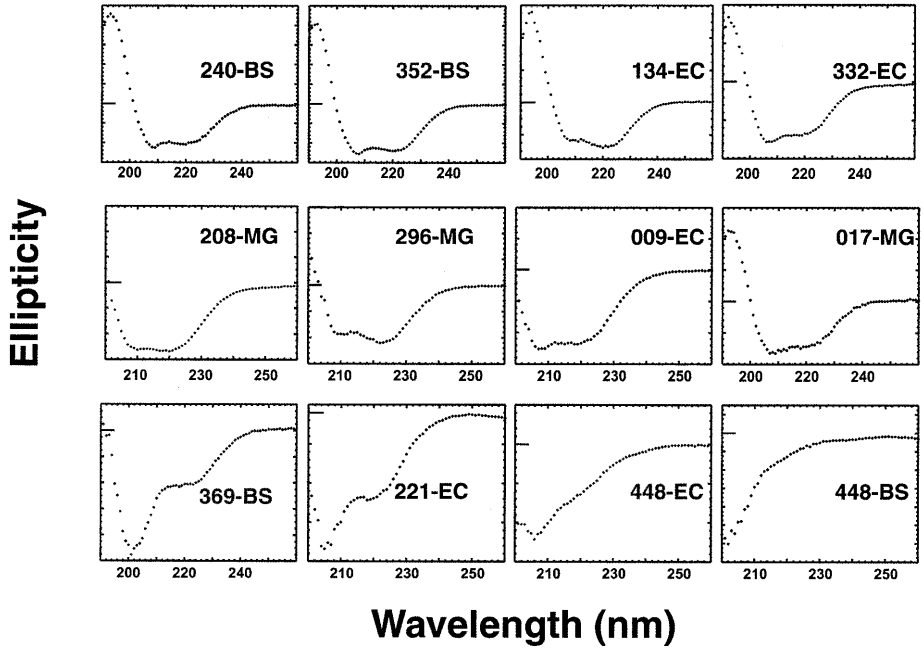

Figure 2. CD spectra of the 12 target proteins at 4°C. The proteins are named beginning with the ORF # associated with a MG gene followed by the source from which the homolog was cloned. The large horizontal line marker on the *y*-axes corresponds to an ellipticity value of zero millidegrees.

Cloning and expression

The 14 genes of interest were cloned from the appropriate genomic DNA using PCR methods. Repeated attempts to clone 461-BS failed for unknown reasons. The remaining 13 genes were successfully cloned and the gene sequences were confirmed by sequencing the entire gene. The 13 proteins were expressed in EC. No protein expression was detected for 056-EC under a number of growth conditions and in different strains. The rest of the 12 proteins expressed at moderate to high levels in EC and were found to partition between the soluble phase and the inclusion body to varying degrees depending on the exact growth conditions.

Although several methods for purification of proteins from inclusion bodies are known (50), we do not know if we have obtained the properly folded native state upon refolding the proteins from inclusion bodies. Therefore, we attempted to purify all the proteins from the soluble phase. This is especially pertinent to our studies because it is difficult to assess whether the protein is properly folded in the absence of a functional assay for these uncharacterized proteins. Cells were grown at lower temperatures (30 or 26°C) to facilitate protein expression in the soluble phase. In the case of 009-EC, protein was obtained in the soluble phase by addition of a lower concentration of IPTG (0.1 mM) for induction as opposed to 1 mM IPTG that we used routinely in cell cultures for protein expression. In spite of these methods, MG017 and 221-EC had to be purified under denaturing conditions and refolded from them. The results of cloning, expression and purification of these proteins are tabulated in Table 1. From Table 1, it can be seen that four of the proteins completely localize in inclusion bodies at 37°C, seven of them partition into both phases to varying degrees and only one of them is expressed completely in the soluble phase. Expression of proteins in the soluble phase, finding the optimal concentration for solubility and testing of different renaturation methods for purifying proteins from inclusion bodies were the rate-limiting steps in the characterization of the target proteins.

Circular dichroism (CD)

After purification, we characterized the proteins by CD. The advantages of CD for high-throughput analyses are that it is rapid, relatively low concentrations of protein are required, and a wide range of solution conditions and temperatures can be explored. The small amounts of material needed for CD methods makes it possible to verify if protein purified from the inclusion body is folded correctly by comparing its CD spectra and thermal melt characteristics with that of the protein purified from the soluble phase. This would allow the structure determination of proteins that localize in inclusion bodies on over-expression because large amounts of protein can be purified from the inclusion body once it is established that the protein can be refolded correctly.

We used CD as a rapid screen by which to assess the secondary structure of the selected target proteins and to identify interesting targets for further characterization. We performed thermal denaturation experiments to qualitatively assess the thermodynamic properties and stability of proteins that had significant secondary structural content.

Figure 2 shows the CD wavelength scans of 12 proteins. Our aim was to qualitatively evaluate the secondary structural content (helical, sheet or random coil) of these proteins rather than precise deconvolution of their CD spectra. Of the 12 proteins, eight have predominantly helical character, two are a mixture of helix and coil (369-BS and 221-EC) and both the EC and BS homologs of MG448 appear to be unstructured. We have classified the 12 target proteins into three different

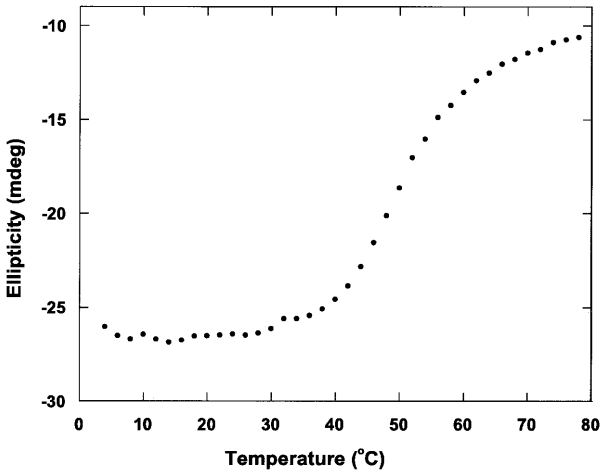

Figure 3. Thermal melt of 332-EC, a representative protein which has significant secondary structure content as discussed in the text. Ellipticity at 222 nm is plotted as a function of temperature.

classes on the basis of their secondary structure profile and thermal unfolding characteristics.

(i) Structured. The CD spectrum resembles that of a typical structured protein, containing varying percentages of α-helix and β-sheet as seen in Figure 2. Seven of the 12 proteins studied (134-EC, 332-EC 240-BS, 352-BS, 369-BS, 208-MG, 296-MG) undergo cooperative thermal denaturation transitions: a representative thermal melt is depicted in Figure 3. These proteins represent important targets for structure determination projects because the 3D structures of these proteins cannot be predicted by homology. These proteins could have novel folds or may belong to an existing fold that could not be predicted due to low sequence identity with the parent fold.

(ii) Unstructured. The CD spectra of the EC and BS homologs of MG448 resemble that of a random coil indicating that the protein is unstructured (Fig. 2). The lack of any significant secondary structure suggests that it is unfolded on its own and may require additional partners such as a protein, nucleic acid or some other cofactor for folding. Alternatively, the unfolded conformation may be essential for its biological role. MG448 is highly conserved in both prokaryotes and eukaryotes which includes yeast, worm and human homologs. In addition, MG448 is both a member of the proposed minimal gene set as well as an essential gene based on transposon mutagenesis experiments (Table 1). Thus, MG448 represents an interesting target for further studies.

(iii) Unusual. The CD spectra of 009-EC and MG017 indicate that they are helical. MG009 is found in many different organisms whereas MG017 is specific to mycoplasma. Interestingly, both proteins are extremely resistant to thermal denaturation and neither protein shows any loss of structure from 4 to 96°C (Fig. 4A). The CD spectrum of 009-EC shows significant secondary structure content over a wide range of temperatures (Fig. 4B). It is known that helices can associate to form stable coiled-coil motifs. However, analysis of MG009

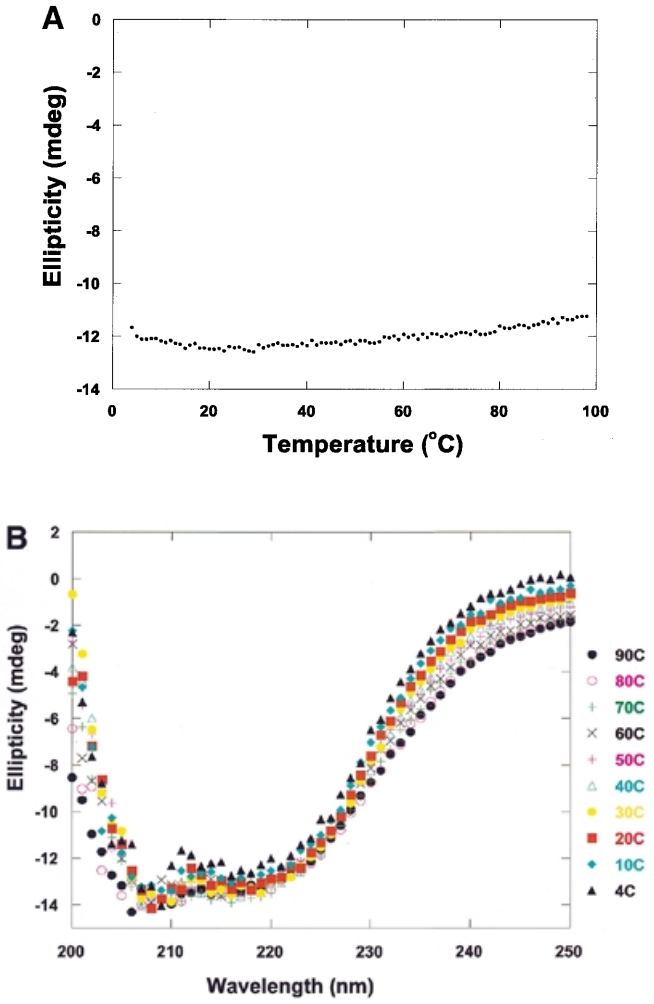

Figure 4. (**A**) Thermal melt of 009-EC. Ellipticity at 222 nm is plotted as a function of temperature. (**B**) CD spectra of 009-EC at various temperatures: black triangle, 4°C; cyan diamond, 10°C; red square, 20°C; yellow circle, 30°C; cyan triangle, 40°C; magenta plus, 50°C; black cross, 60°C; green plus, 70°C; magenta circle, 80°C; black circle, 90°C.

and 009-EC sequences by two different programs, COILS (51) and Multicoil (52), does not reveal any coiled-coil regions.

MG009 is highly conserved from mycoplasma to man (Fig. 5). A BLAST search of MG009 against all the proteins in the non-redundant database at NCBI reveals 46 homologs in various organisms whose sequences have significant *E*-values ranging from 1×10^{-5} to 1×10^{-6} (this includes only two sequences from other mycoplasma species). Although MG009 is classified as a member of the minimal gene set (29), transposon knockout experiments suggest that it is not an essential gene. Nine independent disruptions were observed in this gene. This seemingly paradoxical result can be rationalized in many ways. It is possible that while MG009 is individually dispensable, it may be essential in the absence of another gene. MG009 may be non-essential for MG under the defined laboratory growth conditions but essential in its physiological state. A third, but highly unlikely, possibility is that gene function was not

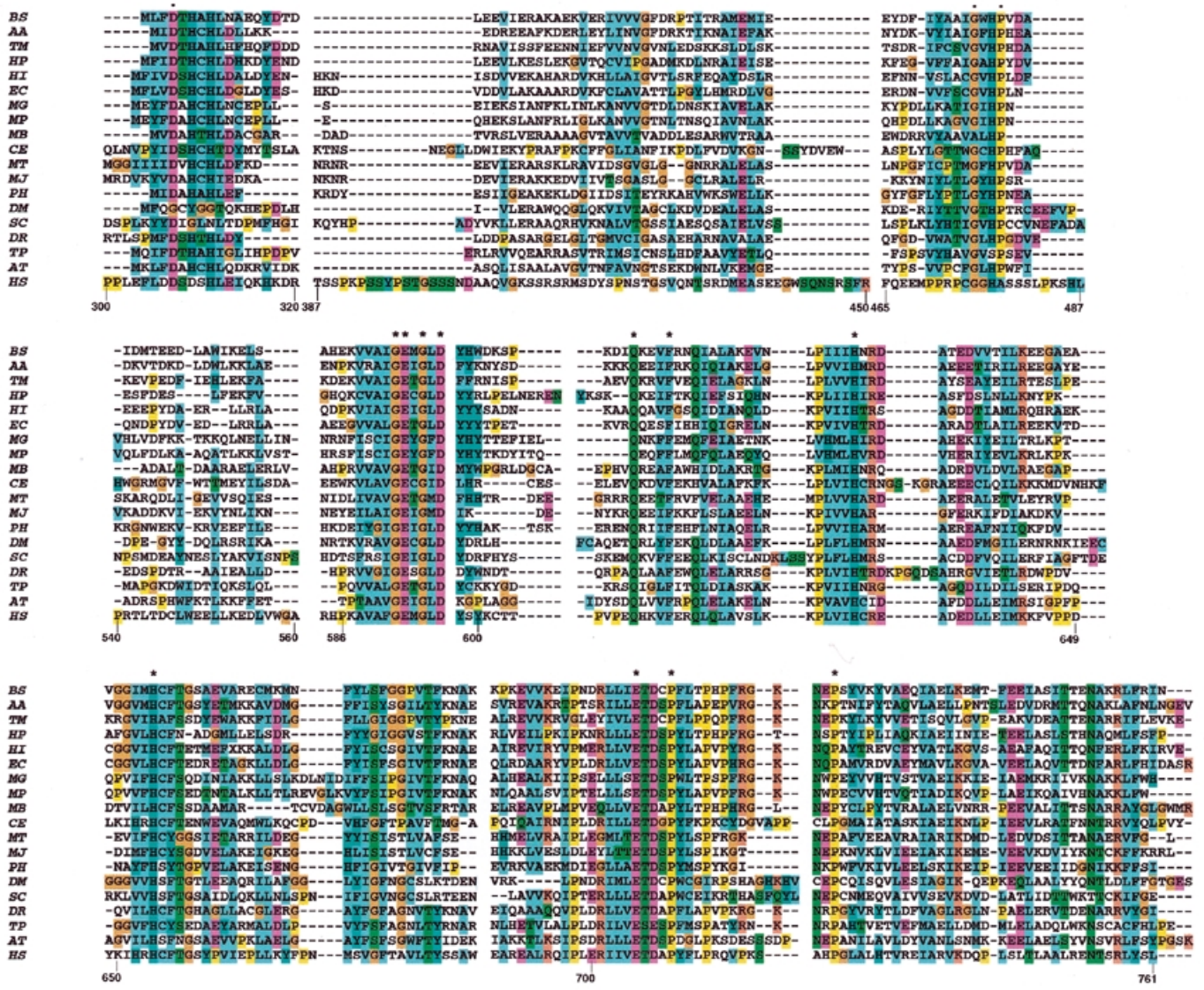

Figure 5. Alignment of selected sequences obtained from a BLAST search of the NR database at NCBI using MG009 as the query sequence. The complete alignment was generated using the ClustalX program (53,54). The name of the homologs from various organisms are abbreviated and depicted to the left of the alignment. *BS*, *B.subtilis*; *AA*, *Aquifex aeolicus*; *TM*, *Thermotoga maritama*; *HP*, *Helicobacter pylori*; *HI*, *H.influenzae*; *EC*, *E.coli*; *MG*, *M.genitalium*; *MP*, *Mycoplasma pneumoniae*; *MB*, *Mycobacterium tuberculosis*; *CE*, *Caenorhabditis elegans*; *MT*, *Methanobacterium thermoautotrophicum*; *MJ*, *Methanococcus jannaschii*; *PH*, *Pyrococcus horikoshii*; *DM*, *Drosophila melanogaster*; *SC*, *Saccharomyces cerevisiae*; *DR*, *Deinococcus radiodurans*; *TP*, *Treponema pallidum*; *AT*, *Arabidopsis thaliana*; *HS*, *Homo sapiens*. The color coding for the alignment uses ClustalX's default protein color file. *, Positions which have a single fully conserved residue. Regions in the human, yeast and *C.elegans* homologs that are not a part of the alignment have not been included The residue number corresponding to the amino acid position in the human sequence is included in the alignment. The entire MG gene is shown in the alignment.

compromised despite the presence of transposon insertions. The unusual thermal stability of this protein and the fact that it is universally conserved make it an important target for further detailed studies.

CONCLUSIONS

We present an analysis of soluble proteins of MG that have no homologs of known 3D structure and no functional annotation. We have demonstrated that the preliminary biophysical characterization of these proteins helps to rapidly identify proteins with unusual properties. Secondary structure profiling of the proteins by CD allows the screening of many proteins essential for structure-based high-throughput genome analyses. These results rapidly identify atypical proteins and distinguish them from folded and well-behaved proteins.

From our study, we have found a variety of proteins: proteins that appear fairly structured and undergo cooperative thermal denaturation, unstructured proteins and thermostable proteins. The structured proteins are ideal targets for 3D structure determination projects and will be pursued in that direction. Several of these proteins are highly conserved from mycoplasma to

man (MG009, MG221, MG448 and MG332). From our studies, we have selected MG448, unstructured by itself and MG009, a thermostable protein, as particularly interesting candidates for further investigation. MG448 may exist as a folded entity when it is bound to other partners such as proteins, nucleic acid or other cofactors. Potential partners for MG448 will be identified. It is also possible that MG448 is unfolded because it is essential for its function. These aspects will be probed in future experiments. MG009 is highly conserved, yet its cellular role is unidentified. In addition, the unusual stability of this protein makes it an interesting subject for further extensive biophysical and biochemical characterization.

ACKNOWLEDGEMENTS

We gratefully acknowledge all the members of the Regan laboratory for their intellectual contributions, support and encouragement throughout the project. We thank Prof. Setlow for a generous gift of BS genomic DNA and Prof. Baseman for MG genomic DNA. M.G. thanks the Keck and Donaghue foundations for financial support. We thank P.Bertone and N.Echols for development of the web database.

REFERENCES

1. Kim,S.H. (1998) *Nature Struct. Biol.*, **5**(Suppl.), 643–645.
2. Rost,B. (1998) *Structure*, **6**, 259–263.
3. Sali,A. (1998) *Nature Struct. Biol.*, **5**, 1029–1032.
4. Shapiro,L. and Lima,C.D. (1998) *Structure*, **6**, 265–267.
5. Burley,S.K., Almo,S.C., Bonanno,J.B., Capel,M., Chance,M.R., Gaasterland,T., Lin,D., Sali,A., Studier,F.W. and Swaminathan,S. (1999) *Nature Genet.*, **23**, 151–157.
6. Montelione,G.T. and Anderson,S. (1999) *Nature Struct. Biol.*, **6**, 11–12.
7. Teichmann,S.A., Chothia,C. and Gerstein,M. (1999) *Curr. Opin. Struct. Biol.*, **9**, 390–399.
8. Eisenstein,E., Gilliland,G.L., Herzberg,O., Moult,J., Orban,J., Poljak,R.J., Banerjei,L., Richardson,D. and Howard,A.J. (2000) *Curr. Opin. Biotechnol.*, **11**, 25–30.
9. Skolnick,J., Fetrow,J.S. and Kolinski,A. (2000) *Nat. Biotechnol.*, **18**, 283–287.
10. Chothia,C. (1992) *Nature*, **357**, 543–544.
11. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure*, **5**, 1093–1108.
12. Zhang,C. and DeLisi,C. (1998) *J. Mol. Biol.*, **284**, 1301–1305.
13. Govindarajan,S., Recabarren,R. and Goldstein,R.A. (1999) *Proteins*, **35**, 408–414.
14. Orengo,C.A., Pearl,F.M., Bray,J.E., Todd,A.E., Martin,A.C., Lo Conte,L. and Thornton,J.M. (1999) *Nucleic Acids Res.*, **27**, 275–279.
15. Gerstein,M. and Hegyi,H. (1998) *FEMS Microbiol. Rev.*, **22**, 277–304.
16. Skolnick,J. and Fetrow,J.S. (2000) *Trends Biotechnol.*, **18**, 34–39.
17. Gerstein,M. (1998) *Fold Des.*, **3**, 497–512.
18. Christendat,D., Yee,A., Dharamsi,A., Kluger,Y., Gerstein,M., Arrowsmith,C. and Edwards,A. (2000) *Prog. Biophys. Mol. Biol.*, in press.
19. Terwilliger,T.C., Waldo,G., Peat,T.S., Newman,J.M., Chu,K. and Berendzen,J. (1998) *Protein Sci.*, **7**, 1851–1856.
20. Cort,J.R., Koonin,E.V., Bash,P.A. and Kennedy,M.A. (1999) *Nucleic Acids Res.*, **27**, 4018–4027.
21. Zarembinski,T.I., Hung,L.W., Mueller-Dieckmann,H.J., Kim,K.K., Yokota,H., Kim,R. and Kim,S.H. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 15189–15193.
22. Gosser,Y.Q., Nomanbhoy,T.K., Aghazadeh,B., Manor,D., Combs,C., Cerione,R.A. and Rosen,M.K. (1997) *Nature*, **387**, 814–819.
23. Campbell,K.M., Terrell,A.R., Laybourn,P.J. and Lumb,K.J. (2000) *Biochemistry*, **39**, 2708–2713.
24. Mercier,P., Li,M.X. and Sykes,B.D. (2000) *Biochemistry*, **39**, 2902–2911.
25. Wright,P.E. and Dyson,H.J. (1999) *J. Mol. Biol.*, **293**, 321–331.
26. Dunker,A.K., Garner,E., Guilliot,S., Romero,P., Albrecht,K., Hart,J., Obradovic,Z., Kissinger,C. and Villafranca,J.E. (1998) *Pac. Symp. Biocomput.*, pp. 473–484.
27. Romero,P., Obradovic,Z., Kissinger,C.R., Villafranca,J.E., Garner,E., Guilliot,S. and Dunker,A.K. (1998) *Pac. Symp. Biocomput.*, pp. 437–448.
28. Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M. *et al*. (1995) *Science*, **270**, 397–403.
29. Mushegian,A.R. and Koonin,E.V. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 10268–10273.
30. Hutchison,C.A., Peterson,S.N., Gill,S.R., Cline,R.T., White,O., Fraser,C.M., Smith,H.O. and Venter,J.C. (1999) *Science*, **286**, 2165–2169.
31. Fischer,D. (1999) *Protein Eng.*, **12**, 1029–1030.
32. Wootton,J.C. (1994) *Comput. Chem.*, **18**, 269–285.
33. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
34. Wootton,J.C. and Federhen,S. (1996) *Methods Enzymol.*, **266**, 554–571.
35. Holm,L. and Sander,C. (1998) *Bioinformatics*, **14**, 423–429.
36. Park,J. and Teichmann,S.A. (1998) *Bioinformatics*, **14**, 144–150.
37. Engelman,D.M., Steitz,T.A. and Goldman,A. (1986) *Annu. Rev. Biophys. Biophys. Chem.*, **15**, 321–353.
38. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
39. Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessieres,P., Bolotin,A., Borchert,S., Borriss,R., Boursier,L., Brans,A., Braun,M., Brignell,S.C., Bron,S., Brouillet,S., Bruschi,C.V., Caldwell,B., Capuano,V., Carter,N.M., Choi,S.K., Codani,J.J., Connerton,I.F., Danchin,A. *et al*. (1997) *Nature*, **390**, 249–256.
40. Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) *Science*, **277**, 1453–1474.
41. Gerstein,M. (1998) *Proteins*, **33**, 518–534.
42. Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
43. 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.
44. Fischer,D. and Eisenberg,D. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 11929–11934.
45. Huynen,M., Doerks,T., Eisenhaber,F., Orengo,C., Sunyaev,S., Yuan,Y. and Bork,P. (1998) *J. Mol. Biol.*, **280**, 323–326.
46. Teichmann,S.A., Park,J. and Chothia,C. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 14658–14663.
47. Jones,D.T. (1999) *J. Mol. Biol.*, **287**, 797–815.
48. Wolf,Y.I., Brenner,S.E., Bash,P.A. and Koonin,E.V. (1999) *Genome Res.*, **9**, 17–26.
49. Minion,F.C. (1998) *Methods Mol. Biol.*, **104**, 259–265.
50. Lilie,H., Schwarz,E. and Rudolph,R. (1998) *Curr. Opin. Biotechnol.*, **9**, 497–501.
51. Lupas,A., Van Dyke,M. and Stock,J. (1991) *Science*, **252**, 1162–1164.
52. Wolf,E., Kim,P.S. and Berger,B. (1997) *Protein Sci.*, **6**, 1179–1189.
53. Jeanmougin,F., Thompson,J.D., Gouy,M., Higgins,D.G. and Gibson,T.J. (1998) *Trends Biochem. Sci.*, **23**, 403–405.
54. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) *Nucleic Acids Res.*, **25**, 4876–4882.