Sequence and structure features of GroE substrates orthologs in *Mycoplasma*

Marina Parr

Laboratory of Bioinformatics, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia, mar.ark.parr@gmail.com

Dmitrii Frishman

Laboratory of Bioinformatics, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia; Department of Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, Freising, Germany; d.frishman@wzw.tum.de

Molecular chaperons help *de novo* synthesized polypeptides to reach native conformation. One specific type of chaperons, the bacterial chaperonin GroEL, together with its co-factor GroES, prevents its substrates from aggregation by shielding them from other molecules in a large cage. The *E.coli* GroEL interactome contains 250 proteins, which are subdivided into three classes according to their GroEL dependency [1]: proteins that can fold correctly without GroEL (class I), proteins that need GroEL for correct folding under stress conditions (class II), and obligate substrates, *i.e.* proteins that can only reach the native conformation with the help of GroEL (class III).

In spite of the importance of the GroEL/GroES system, the majority of the genomes from the *Mycoplasma* genus lack the *groe* operon (GroE- *Mycoplasma*). Using the *E. coli* GroEL sequence as query for BLAST searches against the UniProt database we identified only four GroE+ *Mycoplasma* species: *Mycoplasma pneumoniae*, *Mycoplasma genitalium*, *Mycoplasma gallisepticum*, and *Mycoplasma penetrans*. Although the GroE+ and GroE- *Mycoplasmas* are similar in genome size and GC content, we found significant differences in the amino acid composition between the soluble proteins from these two groups of genomes, with Ala, Asn, Gln, His, Leu, Pro, Val being enriched and Asp, Glu, Ile, Lys, Met, Tyr being depleted in GroE+ proteomes.

Using the OMA database we found the total of 77 orthologs of the E.coli GroEL substrates which are present in both GroE+ and GroE- *Mycoplasma*, with 24, 32, and 21 of them being orthologs of class I, II, and III substrates, respectively. Irrespective of the class all GroEL substrate orthologs differ in amino acid composition from other cytoplasmic proteins and display a depletion in aromatic amino acid residues and enrichment in hydrophobic residues.

We sought to identify the features that enable orthologs of obligate GroEL substrates to fold autonomously in the absence of GroE. Georgescauld *et al.* [2] compared the amino acid usage of the NanA protein from E.coli (GroE+ bacteria) (EcNanA), which is an obligate GroEL substrate, with the amino acid usage of its ortholog, the NanA protein from GroE- *Mycoplasma synoviae* (MsNanA). This comparison revealed that MsNanA has the two times more Lys residues and two times fewer Arg residues than EcNanA, and that its amino acid sequence is also strongly enriched in Phe and Tyr. Here we attempted to reproduce the results presented in [2] at a proteome-wide scale by extending amino acid usage comparison to include all orthologs of obligate GroEL substrates in GroE+ and GroE- *Mycoplasma*. The majority of the amino acid residues overrepresented in MsNanA are helix-promoting residues. This observation prompted us to test the hypothesis, originally formulated by Georgescauld *et al*., that it is the higher content of helix-formers that drives the folding of proteins in GroE- organisms and eliminates their GroEL dependence. To test this hypothesis we obtained secondary structure assignments for proteins with known 3D structure using the DSSP algorithm [3]. Our analysis of amino acid usage in the orthologs of GroEL obligate substrates in GroE- *Mycoplasma* only partially confirms the findings reported in [2], as we only detected a significant enrichment of Lys and Phe in the helical regions. These results were normalized by taking into account the amino acid usage in the helical regions of all proteins from GroE+ and GroE- *Mycoplasma* proteomes (except transmembrane proteins). Lys is known to be a strong helix-former, while Phe is a hydrophobic amino acid with a large side-chain, known to be favorable for secondary structure formation. Work is underway to determine other features of chaperons and protease systems in GroE- *Mycoplasma* that might compensate for the absence of GroEL.

1. M.J.Kerner et al. (2005) Proteome-wide analysis of chaperonin-dependent protein folding in Escherichia coli, *Cell*, **122(2):**209-20.

2. F.Georgescauld et al. (2014) GroEL/ES Chaperonin Modulates the Mechanism and Accelerates the Rate of TIM-Barrel Domain Folding, *Cell*, **157(4):**922-34.

3. W.Kabsch and Ch.Sander (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, **12:**2577–2637.