Low folding propensity and high translation efficiency distinguish in vivo substrates of GroEL from other Escherichia coli proteins

Orly Noivirt-Brik¹, Ron Unger² and Amnon Horovitz¹,.*

¹Department of Structural Biology, Weizmann Institute, Rehovot 76100 and ²The Mina & Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan 52900, Israel

Received on July 10, 2007; revised on August 28, 2007; accepted on October 8, 2007

ABSTRACT

Motivation: Theoretical considerations have indicated that the amount of chaperonin GroEL in Escherichia coli cells is sufficient to fold only ~2–5% of newly synthesized proteins under normal physiological conditions, thereby suggesting that only a subset of E.coli proteins fold in vivo in a GroEL-dependent manner. Recently, members of this subset were identified in two independent studies that resulted in two partially overlapping lists of GroEL-interacting proteins. The objective of the work described here was to identify sequence-based features of GroEL-interacting proteins that distinguish them from other E.coli proteins and that may account for their dependence on the chaperonin system.

Results: Our analysis shows that GroEL-interacting proteins have, on average, low folding propensities and high translation efficiencies. These two properties in combination can increase the risk of aggregation of these proteins and, thus, cause their folding to be chaperonin-dependent. Strikingly, we find that these properties are absent in proteins homologous to the E.coli GroEL-interacting proteins in Ureaplasma urealyticum, an organism that lacks a chaperonin system, thereby confirming our conclusions.

Contact: amnon.horovitz@weizmann.ac.il

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The Escherichia coli chaperonin system facilitates protein folding in vivo and in vitro in an ATP-dependent manner (Horovitz and Willison, 2005; Horwich et al., 2007). It comprises GroEL and its helper-protein GroES, which are both essential proteins (Fayet et al., 1989). GroEL is able to bind and assist the folding of a wide range of proteins in vitro. For example, it was shown that ~40% of the soluble proteins in E. coli can interact with GroEL in vitro (Viitanen et al., 1992) and that even randomly generated artificial polypeptides are GroEL binders (Aoki et al., 2000). Theoretical considerations have indicated, however, that the amount of GroEL in the cell is sufficient to fold only ~2–5% of newly synthesized proteins under normal physiological conditions (Lorimer, 1996), thereby suggesting that only a subset of E.coli proteins fold in vivo in a GroE-dependent manner. An initial attempt to identify this set of proteins resulted in a list of about 300 proteins that are comprised preferentially of αβ domains (Houy et al., 1999). More recently, two additional sets of GroEL-interacting proteins (GIP) were identified for which there is less experimental uncertainty regarding whether they represent the true in vivo substrates (Chapman et al., 2006; Kerner et al., 2005). Overall, it appears that ~10% of the cytosolic proteins interact with GroEL under normal growth conditions (Ewalt et al., 1997; Houy et al., 1999).

The existence of a subset(s) of E.coli proteins that require the GroE system for folding raises the following two questions. The first concerns the mechanism by which GroEL recognizes its substrates. In other words, what are the features that confer binding to GroEL that distinguish members of the set of GIP from other E.coli proteins that are non-binders? Given that the mobile loop of GroES and protein substrates share the same binding site in GroEL, it was suggested that one possible answer to this question is that members of the set of GIP contain sequence motifs similar to that of the GroES mobile loop (Chaudhuri and Gupta, 2005; Stan et al., 2005, 2006). The second question concerns identifying the properties of members of the set of GIP that render them GroE-dependent. In this article, we focus on this second question. It should be mentioned, however, that the two questions are not entirely unrelated. For example, slow folding may cause a protein to be prone to aggregation and, thus, GroE-dependent and, at the same time, facilitate binding to GroEL.

The work described here involves DNA and protein sequence analysis of the set of GIP identified by Kerner et al. (2005). This set was partitioned into three classes: (i) class I substrates that are assisted by GroE but do not require it; (ii) class II substrates that require both GroEL and GroES at 37°C but do not require GroES at 25°C and (iii) class III substrates that require the GroE-system for folding in a stringent manner (Kerner et al., 2005). This partitioning enabled us to compare GroEL binders and non-binders at a higher resolution by analyzing differences also between the three classes. Our results show that GroEL binders have, on average, relatively low folding propensities but higher translation efficiencies as compared with all other E.coli proteins (i.e. non-binders). These results were confirmed by showing that these features are also present in the set of GIP
identified by Chapman et al. (2006) but absent in the proteins homologous to the set of GIP in *Ureaplasma urealyticum*, an organism that lacks a chaperonin system.

2 METHODS

Differences between sequence-based features of the set of GIP and all other *E.coli* proteins were evaluated by comparing the mean value for the set of GIP to the distribution of the mean values of 1000 sets of randomly selected *E.coli* proteins (from a set of 3444 proteins that exclude GIP and their homologues) each with the same number of proteins as in the set of GIP. Other comparisons, such as between the separate classes of GIP or the GIP homologues in *U.urealyticum* and all other *E.coli* proteins, were carried out in a similar way. A P-value of < 0.001 was assigned in cases where the mean value of the examined set (e.g. that of GIP) is more extreme than (or equal to) the mean values of all the 1000 sets of randomly selected proteins. It is important to note that in the case of all the features examined, the values for individual members of the set of GIP were not found to be significantly separable in space from the values for members of the random sets and, therefore, only the mean values for each feature are compared. In addition, the Wilcoxon rank sum test was implemented in order to evaluate whether the sample (GIP, *U.urealyticum* homologues or the different classes) and the rest of the *E.coli* proteome have equal medians.

The full genome and proteome sequences of the K12 strain of *E.coli* (corresponding to accession number NC_000913.1) were downloaded from the Refseq database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_K12/) (Riley et al., 2006). The full genome and proteome sequences of *U.urealyticum* (Glass et al., 2000) were downloaded from the PEDANT database (http://pedant.gsf.de/) (Frishman and Mewes, 1997). The sequences of GIP reported by Kerner et al. (2005) were also downloaded from the PEDANT database (http://pedant.gsf.de/links.jsp). This data set contains 252 proteins that are divided into classes I–III comprising 38, 126 and 84 proteins, respectively, and 4 proteins whose class was not determined. The protein P00810 (that belongs to class I) was excluded from our analysis as it is a β-lactamase that is not encoded by the *E.coli* K12 genome. The GIP set of Chapman et al. (2006) was built using the accession-number list provided in their supporting information. This data set contains 317 proteins of which 136 appear also in the set of Kerner et al. (2005). The other 181 proteins found only by Chapman et al. (2006) constitute a set that is used to confirm our findings for the set of Kerner et al. (2005).

In the analysis, we use only 180 sequences (instead of 181) since protein YP_026165.1 was not found in the proteome of the K12 strain of *E.coli* (accession number NC_000913.1). Identification of proteins in the *E.coli* and *U.urealyticum* proteomes that are homologous to proteins in the GIP set of Kerner et al. (2005) was carried out using BLAST (Altschul et al., 1990). All proteins found to have an E-value smaller than 0.1 were assigned as GIP homologues.

The tRNA adaptation index (tAI) was calculated using the codonM script and tAI.R program provided by dos Reis et al. (2004). The tAI is a measure of tRNA usage by the coding sequences. Each codon is assigned a weight based on the gene copy numbers of all the tRNA that recognize it. This measure relies on the fact that the tRNA gene copy number has a high correlation with the tRNA abundance within the cell and with the codon preferences in the genomes (Ikemura, 1981a, b). The tAI values are in the range of 0–1 where high values correspond to high translation efficiency. The tRNA copy numbers were taken from the genomic tRNA database (Lowe and Eddy, 1997) (http://lowelab.ucsc.edu/GrtRNAAdh/).

The FoldIndex (Prilusky et al., 2005) measure was calculated for each protein using the server (http://bip.weizmann.ac.il/Fldbin/index). This measure is based on a linear equation proposed by Uversky et al. (2000) that separates ‘folded’ proteins from ‘disordered’ ones according to the absolute value of their net charge and average hydrophobicity. The values of this measure are in the range of −1 to 1 where larger values correspond to a higher probability that the protein is able to fold. Here, we use FoldIndex as a measure of folding propensity.

3 RESULTS AND DISCUSSION

The means of the folding propensity (Prilusky et al., 2005), translation efficiency (dos Reis et al., 2004) and GC content of the set of GIP (Kerner et al., 2005) were found to be significantly different from the respective means of sets of randomly selected *E.coli* proteins (Fig. 1 and Table 1). Similar results were obtained when classes II and III of the GIP were compared individually with the random sets. In the case of class I, however, only the mean value of the translation efficiency was found to be significantly different from the respective means of the random sets (Table 1). Importantly, similarly significant results were obtained for all three features when the full set of GIP of Chapman et al. (2006), or the subset of GIP found by Chapman et al. (2006) that is not included in the set of Kerner et al. (2005), were compared to the random sets (Table 1). All the differences were found to be significant also according to the Wilcoxon rank sum test except for the GC content of the subset of GIP found by Chapman et al. (2006) but not by Kerner et al. (2005) in which case we were not able to reject the hypothesis that its median value is equal to that of the rest of the *E.coli* proteome. It should be noted that other properties we examined, such as contact order (Plaxco et al., 1998) and α-helix and β-strand secondary structure content, were found to not distinguish between GIP and other *E.coli* proteins (data not shown).

The mean folding propensity (as measured by the FoldIndex) of the different sets of GIP was found to be lower than those of the random sets (Table 1). This difference is, as expected, less significant in the case of class I proteins that are assisted by

![Fig. 1. Comparison of sequence features of GroEL-dependent and independent proteins. The mean GC content (a), FoldIndex (b) and tAI (c) values of the set of GroEL-interacting proteins (GIP) identified by Kerner et al. (2005) (in dashed black line) is shown versus the distribution of the respective mean values for 1000 sets of similar size of randomly selected proteins that are not included in the set of GIP or are homologous to any of its members.](image-url)
GroEL in folding but do not require it. In contrast, the mean tAI values of all the GIP sets was found to be significantly higher than the mean values of the random sets. Taken together, our results indicate that in vivo substrates of GroEL are characterized by low folding propensities and high translation efficiencies. These observations can be rationalized by assuming that proteins that fold slowly but are synthesized rapidly are more aggregation prone and, as a result, are chaperonin-dependent.

A negative control for the above findings is provided by analysis of the proteins that are homologous to GIP in *U.urealyticum*, an organism that lacks the GroE system. The full *U.urealyticum* genome contains 613 proteins of which 21, 37 and 28 are homologous to members of classes I, II and III, respectively, in Kerner et al. (2005) (here only the best hits were considered to be homologous). The non-GIP set that was used for generating the random sets contains all *U.urealyticum* proteins except for those that are homologous (best hit or otherwise) to the set of GIP in *E.coli* (i.e., 502 proteins). It may be seen in Table 1, Figure 2 and the Supplementary Material that the set of *U.urealyticum* proteins homologous to GIP in *E.coli* has a mean FoldIndex value that is similar to the means of the random sets. Moreover, the mean tAI value of this set is significantly smaller than the mean values of the random sets instead of being larger as found in the case of the GIP set of *E.coli*. These findings support our conclusion that GIP differ, on average, from other *E.coli* proteins in their folding potential and translation efficiency as these differences are absent when a similar comparison is made for an organism that lacks the GroE system. The set of GIP also differs from that of other *E.coli* proteins in its mean GC content but a similar difference is found in *U.urealyticum* and this feature is, therefore, not considered further.

**Table 1.** Comparison of the mean values of sequence features of different sets of GroEL-interacting proteins (\(N_{\text{GIP}}\)) with the respective mean values of 1000 sets of randomly selected proteins that do not interact with GroEL (\(N_{\text{rand}}\))

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Set size</th>
<th>GC content (P)-value for (N_{\text{GIP}} &gt; N_{\text{rand}})</th>
<th>FoldIndex (P)-value for (N_{\text{GIP}} &gt; N_{\text{rand}})</th>
<th>tAI (P)-value for (N_{\text{GIP}} &gt; N_{\text{rand}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIP(^a)—all classes</td>
<td>251</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>GIP(^a)—class I</td>
<td>37</td>
<td>0.072</td>
<td>0.056</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>GIP(^a)—class II</td>
<td>126</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>GIP(^a)—class III</td>
<td>84</td>
<td>&lt;0.001</td>
<td>0.003</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>GIP(^b)</td>
<td>316</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>GIP(^c)</td>
<td>180</td>
<td>0.004</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>GIP homologues in <em>U.urealyticum</em></td>
<td>86</td>
<td>&lt;0.001</td>
<td>0.167</td>
<td>0.978(^d)</td>
</tr>
</tbody>
</table>

\(^a\)The set of GroEL-interacting proteins of Kerner et al. (2005).

\(^b\)The set of GroEL-interacting proteins of Chapman et al. (2006).

\(^c\)The set of GroEL-interacting proteins of Chapman et al. (2006) that are not found in the set of Kerner et al. (2005).

\(^d\)Note that the mean tAI value of the GIP homologues from *U.urealyticum* is significantly smaller than the mean values of the tAI values of the random sets (\(P\)-value = 0.022).

It may be seen in Figure 2 that the GIP set of *E.coli* has very unique properties as a group and is clearly separable from the random sets (Fig. 2a) whereas the GIP homologues in *U.urealyticum* are indistinguishable from the other proteins in this organism (Fig. 2b). A possible explanation for this observation is that two independent evolutionary processes that compensate for the absence of a chaperonin system occurred in *U.urealyticum*. These processes led to (i) increased folding propensities as reflected in higher FoldIndex values and (ii) reduced translation rates (as reflected in lower tAI values) that may cause folding to occur in a more efficient co-translational manner. The weak (\(r \approx 0.3\)) negative correlation in Figure 2b between the translation efficiency and FoldIndex value is consistent with the above explanation as it may reflect the subset of *U.urealyticum* proteins with a potential problem in folding that is circumvented by having both a high FoldIndex value and a low translation efficiency. In addition, it has been shown that there is a significant and positive correlation between the tAI values and the protein
abundance levels in yeast (Man and Pilpel, 2007) and between the tAI values and the mRNA expression levels in E.coli (dos Reis et al., 2003). These correlations, if they also exist in U.urealyticum, may indicate that the significant reduction in the tAI values of the GIP homologues in this organism reflects a tendency to reduce the production of proteins that require a chaperonin system. This line of reasoning is also supported by the observation that a very high anti-correlation was found for humans between mRNA expression levels and aggregation rates (Tartaglia et al., 2007).

In conclusion, we have identified sequence-based characteristics related to folding propensity and translation efficiency that significantly differentiate GIP in E.coli from the rest of the E.coli proteome. These characteristics, although reflected only in the mean values of the sets and not in each protein separately, highlight the relationship between the folding properties of a protein and its chaperonin dependence. Further analysis is required in order to reveal the mechanism by which GroEL recognize these substrates.

ACKNOWLEDGEMENTS

This work was supported by grants to A.H. from the Israel Science Foundation (67/05) and the Kimmelman Center for Macromolecular Assembly.

Conflict of Interest: none declared.

REFERENCES

Stun,G. et al. (2006) Residues in substrate proteins that interact with GroEL in the capture process are buried in the native state. Proc Natl Acad Sci. USA, 103, 4433–4438.