



RESEARCH ARTICLE

How do eubacterial organisms manage aggregation-prone proteome? [v1; ref status: indexed, <http://f1000r.es/3k3>]

Rishi Das Roy^{1,4}, Manju Bhardwaj², Vasudha Bhatnagar³, Kausik Chakraborty¹, Debasis Dash^{1,4}

¹GNR Knowledge Centre for Genome Informatics, Institute of Genomics and Integrative Biology, Council of Scientific and Industrial Research, Delhi, 110007, India

²Department of Computer Science, Maitreyi College, Chanakyapuri, Delhi, 110021, India

³Department of Computer Science, Faculty of Mathematical Sciences, University of Delhi, Delhi, 110007, India

⁴Department of Biotechnology, University of Pune, Pune, 411007, India

v1 First published: 27 Jun 2014, 3:137 (doi: [10.12688/f1000research.4307.1](https://doi.org/10.12688/f1000research.4307.1))
Latest published: 27 Jun 2014, 3:137 (doi: [10.12688/f1000research.4307.1](https://doi.org/10.12688/f1000research.4307.1))

Abstract

Eubacterial genomes vary considerably in their nucleotide composition. The percentage of genetic material constituted by guanosine and cytosine (GC) nucleotides ranges from 20% to 70%. It has been posited that GC-poor organisms are more dependent on protein folding machinery. Previous studies have ascribed this to the accumulation of mildly deleterious mutations in these organisms due to population bottlenecks. This phenomenon has been supported by protein folding simulations, which showed that proteins encoded by GC-poor organisms are more prone to aggregation than proteins encoded by GC-rich organisms. To test this proposition using a genome-wide approach, we classified different eubacterial proteomes in terms of their aggregation propensity and chaperone-dependence using multiple machine learning models. In contrast to the expected decrease in protein aggregation with an increase in GC richness, we found that the aggregation propensity of proteomes increases with GC content. A similar and even more significant correlation was obtained with the GroEL-dependence of proteomes: GC-poor proteomes have evolved to be less dependent on GroEL than GC-rich proteomes. We thus propose that a decrease in eubacterial GC content may have been selected in organisms facing proteostasis problems.

Open Peer Review

Referee Status:

	Invited Referees	
	1	2
version 1 published 27 Jun 2014	 report	 report
1	Amnon Horovitz , Weizmann Institute of Science Israel	
2	Annalisa Pastore , MRC National Institute for Medical Research UK	

Discuss this article

Comments (0)

Corresponding authors: Kausik Chakraborty (kausik@igib.in), Debasis Dash (ddash@igib.in)

How to cite this article: Das Roy R, Bhardwaj M, Bhatnagar V *et al.* **How do eubacterial organisms manage aggregation-prone proteome? [v1; ref status: indexed, <http://f1000r.es/3k3>]** *F1000Research* 2014, 3:137 (doi: [10.12688/f1000research.4307.1](https://doi.org/10.12688/f1000research.4307.1))

Copyright: © 2014 Das Roy R *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](https://creativecommons.org/licenses/by/4.0/) (CC0 1.0 Public domain dedication).

Grant information: RDR acknowledges HP and CSIR for funding the study. DD and KC acknowledge CSIR for funding through IGIB. KC acknowledges Wellcome Trust – DBT India alliance for funding and fellowship.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 27 Jun 2014, 3:137 (doi: [10.12688/f1000research.4307.1](https://doi.org/10.12688/f1000research.4307.1))

First indexed: 01 Oct 2014, 3:137 (doi: [10.12688/f1000research.4307.1](https://doi.org/10.12688/f1000research.4307.1))

Introduction

Eubacterial organisms have genomes that vary largely in their nucleotide compositions. In this kingdom, the GC content varies from 20% to 70% of the genome and this large variation has been documented in a number of reports that have aimed to explain it¹⁻³. The amino acid compositions are also different in eubacterial proteomes due to the variation of GC content⁴. It has been reported that these difference of amino acid compositions alter the characteristics of proteomes and as a consequence, proteins of GC-poor genomes are more prone to misfolding and aggregation compare to GC-rich genomes^{5,6}. It has been hypothesized that GroEL plays a major role, if not an essential role, in the evolution of GC-poor organisms by buffering deleterious mutations that are fixed due to population bottlenecks⁷⁻⁹. This has been supported by the observation that many of the small GC-poor endosymbionts tend to overexpress GroEL¹⁰⁻¹².

However, the proposed chaperone dependence of GC-poor organisms does not explain why some of the GC-poor endosymbionts of the mycoplasma group have lost the *groEL* copy from their genome¹³. It is notable that these are the only known eubacterial organisms to have lost this gene. This observation led us to test the proposed relationship of GC poorness of genome with the aggregation propensity of the encoded proteome.

Obtaining information on the aggregation propensity of proteins from different organisms is a challenging task. However, there has already been a careful characterization of the aggregation propensity of different *Escherichia coli* proteins that was conducted in a high-throughput manner¹⁴⁻¹⁶. Kerner *et al.* classified the GroEL substrates into Class I, II or III based on the interaction strength and on the stringency of their requirement for GroEL. Class III (C3) substrates were completely dependent on GroEL for folding, whereas Class II (C2) substrates were partially dependent. Class I (C1) proteins interacted weakly with GroEL and were able to fold spontaneously. In a trivial approach, homologs of GroEL-dependent proteins may be identified in other organisms^{13,17}. This approach however fails to predict the evolution of protein dependence on GroEL correctly, as the sequence differences between species have the potential to introduce or remove kinetic traps from folding pathways, thereby altering their dependence on GroEL. In addition to the solubility of the *E. coli* proteome in a chaperone-free system, substrates of another chaperone DnaK were also identified by two independent research groups^{18,19}. Applications developed primarily on machine learning algorithms to classify soluble or GroEL substrates^{16,18,20-24} are already available. However, these classifiers have not been trained with curated data prepared from multiple experimental results^{14,15,18,19}. In this study, we have constructed a more reliable training dataset to build classifiers to determine the aggregation propensity and GroEL dependency in 1132 eubacterial proteomes, based solely on the amino acid sequences. We show a distinct trend in the aggregation propensity of proteins of an organism in relation to the GC content. Surprisingly, aggregation propensity decreased with lower GC content independent of symbiotic characteristics, suggesting that GC-poor organisms have indeed evolved a proteome that is devoid of aggregation-prone proteins.

Materials and methods

Data source

The aggregation-prone proteins of the eSOL database^{18,25} are dependent on the chaperone network of *E. coli* to get their three dimensional native structure. GroEL and DnaK are two important components of this network and their substrates have been extensively studied via different experimental methods^{14,15,19,26}. The integration of all the available information reveals that about half (457) of the soluble or chaperone-independent proteins identified by Niwa *et al.* were found to be GroEL- or DnaK-dependent¹⁸ (Figure 1). To construct a more reliable training set, we removed these proteins from the soluble set. Thus, proteins identified as chaperone-dependent by more than one study, were only considered as aggregation-prone proteins. Furthermore, the proteins which were more than 30% (amino acid) sequence similarity among the remaining proteins were removed using CD-HIT²⁷ clustering program. Therefore the final training set comprised of 502 aggregation prone and 475 soluble proteins.

Classifier building

The classifiers in this study were built with Pro-Gyan²⁸ software. Pro-Gyan builds classifiers directly from training data set given in FASTA format by selecting relevant features from a large number of unbiased features. Following metrics which are useful to evaluate performance of machine learning classifiers were reported by Pro-Gyan.

$$\text{Accuracy(Acc)} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Sensitivity or Recall (Sn)} = \frac{TP}{TP + FN}$$

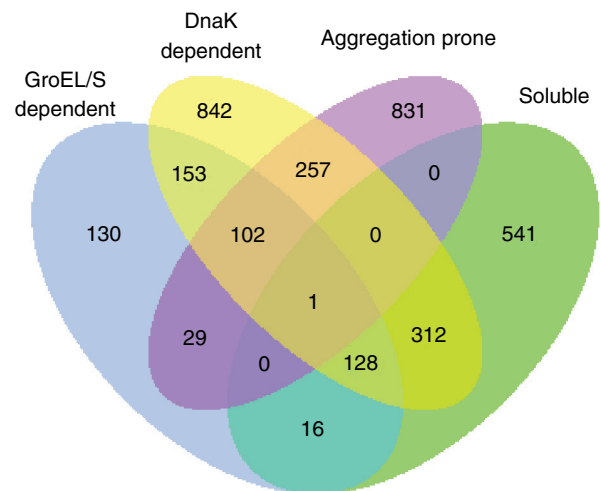


Figure 1. Integration of independent studies. A Venn diagram of proteins of *E. coli* identified by different experimental studies shows that ~45% of soluble proteins reported by Niwa *et al.* overlap with GroEL/S or DnaK substrates (soluble proteins are defined as having solubility >70% and aggregation-prone proteins have solubility <30%).

Sensitivity (Sp) = $TN/(FP + TN)$

Matthews correlation coefficient (MCC) = $(TP*TN-FP*FN)/\sqrt{(TP+FP)*(TN+FN)*(TP+FN)*(TN+FP)}$

where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative predicted by the classifier.

Additionally, receiver operating characteristic (ROC) curves and area under this curve (AUC)²⁹ were also generated.

Analysis on microbial genomes

The protein sequences of microbial genomes were downloaded from the **Microbial Genome Database**³⁰ (archive no. mbgd_2011-01). To identify the chaperonins in the microbial organisms, chaperonin homologs were searched for using BLAST (e-value $1*10^{-4}$) against a chaperonin database **cpnDB**³¹ downloaded on June 2011. The 16S rRNA nucleotide sequence of *E. coli* was acquired from **SILVA**³² and homologous were searched for in other microbial organisms using BLAST (e-value $1*10^{-4}$). GC contents for microbial genomes were calculated using following equation

$$GC \text{ content} = (G + C)/(\text{total bases}),$$

where G = number of guanosine and C = number of cytosine.

Statistical analysis

The Kendall correlation and analysis of covariance were performed in R³³ statistical computing environment using the package ‘stats’ version 2.15.3. To account the effect of evolution on different traits of bacterial genomes, we performed phylogenetic independent contrast through the PDAP³⁴ module on Mesquite³⁵ application.

Results and discussion

Development of machine learning tool to identify aggregation-prone proteins

Recently protein solubility has been carefully measured in a chaperone-free system and the information has been made available through the eSol database¹⁸. Few classification models developed on this database can segregate soluble proteins from chaperone-dependent proteins²²⁻²⁴. However, these web-based classifiers are not suitable to classify large numbers of proteomes, and their soluble or negative training dataset (proteins not aggregation-prone or soluble) are not carefully curated, as most of the soluble proteins from eSol database are substrates of DnaK¹⁹ or GroEL^{14,15} (Figure 1). Therefore we built a classifier containing a curated list of aggregation-prone

proteins and soluble proteins. The classifier was built using **Pro-Gyan**²⁸ which generates 5038 different features from a set of class labelled protein sequences and selects the “maximum relevant minimum redundant” feature subset. Finally, the tool built a support vector machine (SVM)³⁶ classifier by five-fold cross validation. The classifier attained an accuracy of 83.21% with 0.66 MCC (Table 1). Although Pro-Gyan generated classifier was trained with a rigorously curated training data set, it performs equivalent to Fang *et al.*'s classifier and better than others²²⁻²⁴. The receiver operating characteristic (ROC) curves of the classifier are shown in Figure 2. For interested users, the classifier is available in ZENODO (<https://zenodo.org/record/10442/>).

Discriminating features of aggregation prone proteins

To build the classifier, Pro-Gyan²⁸ selected 24 relevant features through an automated process. The top ten significant (by Mann-Whitney test) features were the sequence patterns, the pseudo amino acid composition³⁷ of phenylalanine (F), aspartic (D) and glutamic (E) acid, the distribution of positively charged amino acids, the features calculated from FoldIndex³⁸ and the auto-correlation of hydrophobicity and relative mutability (Table 2). The remaining

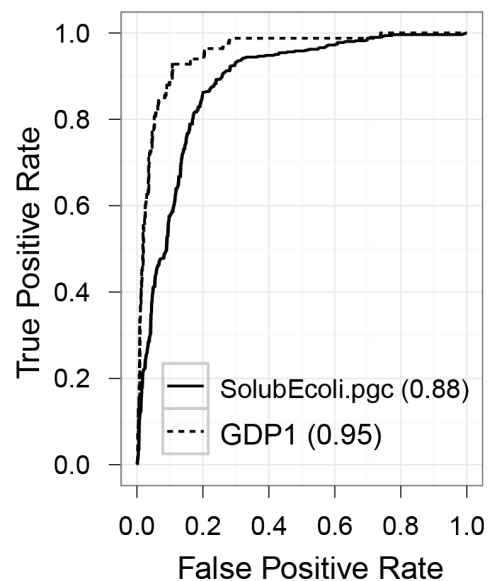


Figure 2. Receiver operating characteristic (ROC) curves. ROC curves of the soluble protein classifier (SolubEcoli.pgc) and the GroEL obligate protein classifier (GDP1.pgc). The areas under the curves (AUC) are given in the legend.

Table 1. Comparison of previous classifiers with our classifier.

Method	Sensitivity	Specificity	Accuracy	AUC	MCC
SVM ²⁵			80		
J48 (decision tree algorithm) ²³			72	0.72	
VTJ48 (visually tuned J48) ²³			76	0.81	
Fang <i>et al.</i> ²²	82.00	85.00	84	0.91	0.67
SolubEcoli.pgc*	86.25	80.00	83.21	0.88	0.66

* Built on a curated training data set.

Table 2. Selected features of proteins used to build the “SolubEcoli.pgc” classifier.

Serial no.	Feature id†	Description	p-value*
1	SW_SOC2	Quasi-sequence-order calculated from physicochemical distance matrix ⁵⁰ .	2.20E-16
2	PPR	Distribution of positively charged amino acids in sequence pattern ⁵¹ .	2.20E-16
3	H(8)M	Amino acid pair composition of histidine to methionine with 8 gaps ⁵² .	2.33E-15
4	M-B(Hydr)1	Moreau-Broto auto correlation (lag 1) of amino acid index; hydrophobicity ⁵³ .	2.24E-08
5	PseAAC_T1_3	Pseudo amino acid composition of aspartic acid (D) ³⁷ .	9.45E-06
6	PseAAC_T1_5	Pseudo amino acid composition of phenylalanine acid (F) ³⁷ .	6.87E-05
7	FI_16_psavgl	Average length of folded segments of proteins according to FoldIndex ³⁸ .	8.14E-05
8	PseAAC_T1_4	Pseudo amino acid composition of glutamic acid (E) ³⁷ .	0.000542
9	Dstrbu_Pol_2:3	Distribution of amino acids according to polarizability ⁵⁴ .	0.001289
10	M-B(mutblty)6	Moreau-Broto auto correlation (lag 6) of amino acid index; relative mutability ⁵³ .	3.65E-03
11	T	Composition of amino acid Threonine ⁵³ .	5.00E-03
12	Mrn(vlum)27	Moran auto correlation (lag 27) of amino acid index; residue volume ⁵³ .	0.00926
13	Mrn(Polar)22	Moran auto correlation (lag 22) of amino acid index; polarizability ⁵³ .	0.013
14	M-B(mutblty)9	Moreau-Broto auto correlation (lag 9) of amino acid index; relative mutability ⁵³ .	0.01988
15	Geary(sterc)4	Geary auto correlation (lag 4) of amino acid index; steric parameter ⁵³ .	0.03536
16	M-B(mutblty)24	Moreau-Broto auto correlation (lag 24) of amino acid index; relative mutability ⁵³ .	0.05416
17	M-B(Hydr)12	Moreau-Broto auto correlation (lag 12) of amino acid index; hydrophobicity ⁵³ .	5.92E-02
18	Mrn(RsdAcc)24	Moran auto correlation (lag 24) of amino acid index; residue accessible surface area in tripeptide ⁵³ .	0.1077
19	Mrn(Hydr)23	Moran auto correlation (lag 23) of amino acid index; hydrophobicity ⁵³ .	0.2106
20	Geary(Free)13	Geary auto correlation (lag 13) of amino acid index; free energy ⁵³ .	0.3271
21	Comp_Vol_2	Composition of normalized van der Waals volume of amino acids of range 2.95–4.0 ⁵³ .	0.4631
22	Geary(vlum)20	Geary auto correlation (lag 20) of amino acid index; residue volume ⁵³ .	4.95E-01
23	Geary(Free)14	Geary auto correlation (lag 14) of amino acid index; free energy ⁵³ .	0.499
24	M-B(vlum)30	Moreau-Broto auto correlation (lag 30) of amino acid index; residue volume ⁵³ .	0.9559

†Internal feature id of the Pro-Gyan application.

selected features (Table 2) were enriched with auto-correlation measurement of amino acid indices such as steric parameter, free energy, accessible surface area, polarizability, residue volume etc. The features which represent patterns of physico-chemical properties encrypted in protein sequences were unique to SolubEcoli.pgc when compared to earlier methods.

Genome wide prediction of aggregation prone proteins

From the analysis of features, it was noticed that the compositions of amino acids are significantly different within aggregation prone and soluble proteins. Sequence features of amino acids have been used to understand protein overexpression related to toxicity³⁹. Additionally, it has been also shown that the amino acid composition is drastically altered in organisms with GC-poor genomes^{4,40}. There

are multiple amino acids that change in frequency as a function of GC content (Figure 3) and this change that has been attributed to the difference in the GC content in the codons of these amino acids. On the basis of these differences, it has been reported that proteins encoded by GC-poor organisms should be more prone to aggregation than proteins encoded by GC-rich organisms^{5,6}. However, the GC composition of the training data showed that aggregation-prone proteins were significantly more GC-rich than the soluble proteins (Figure 4, Mann-Whitney test p-value = $1.3e-15$). Subsequently, we sought to verify the fraction of aggregation-prone proteins across different bacterial proteomes. We used the SolubEcoli.pgc classifier to predict aggregation-prone proteins in 1132 eubacterial species. Our prediction on bacterial genomes showed that the fAg (aggregation prone proteins as fraction of proteome) of a genome

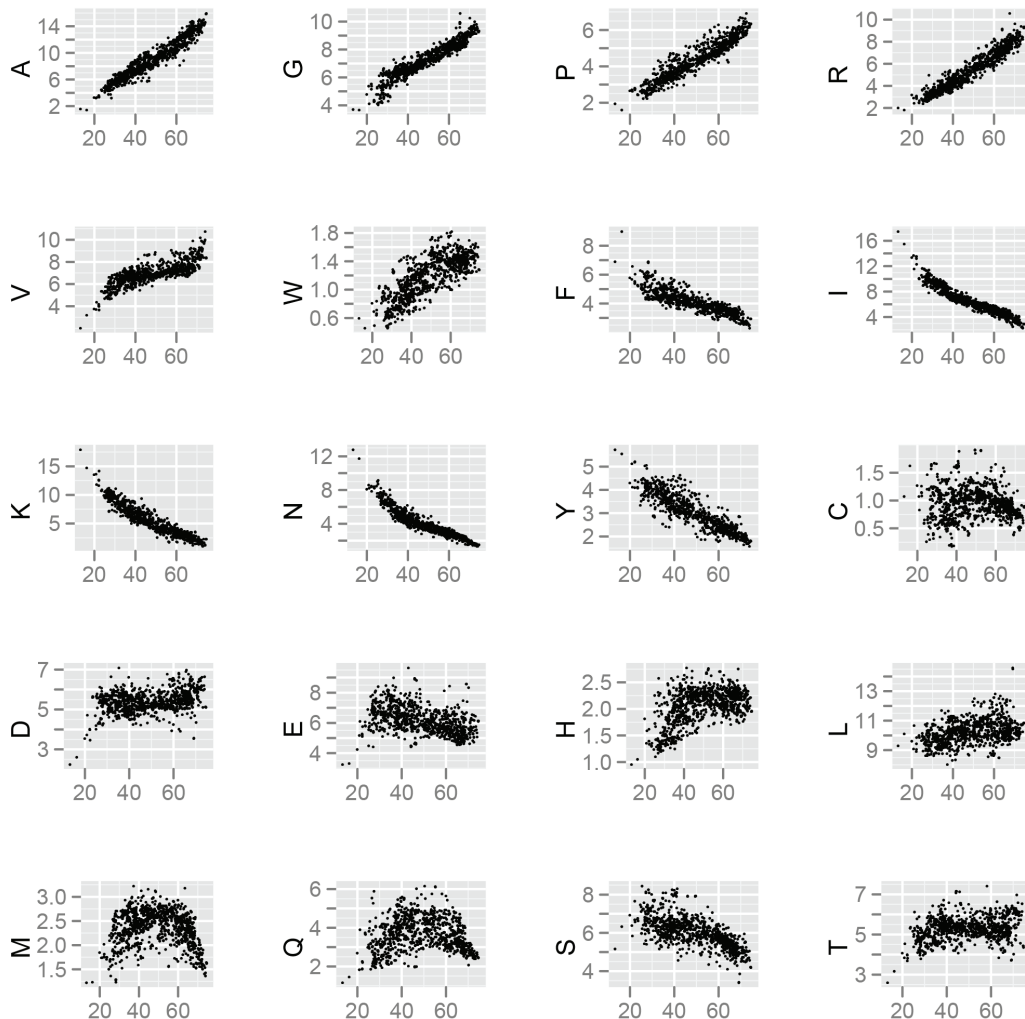


Figure 3. Composition of basic amino acids over ~1100 eubacterial genomes. The x-axis of each subplot shows for GC composition of each genome whereas y-axis shows corresponding amino acid composition.

correlates positively with the GC composition (Kendall tau=0.38 p-value < 2.2e-16) (Figure 5A). We further examined the correlation, with respect to phylogenetic ancestry, using the Mesquite system³⁵, because the Kendall correlation assumes that observations are independent even if organisms are linked through common ancestors⁴¹. The required phylogenetic tree was constructed from the 16S rRNA gene sequences of 570 bacteria⁴². We found a significant correlation (0.4, p-value < 2.2e-16) between independent contrasts of

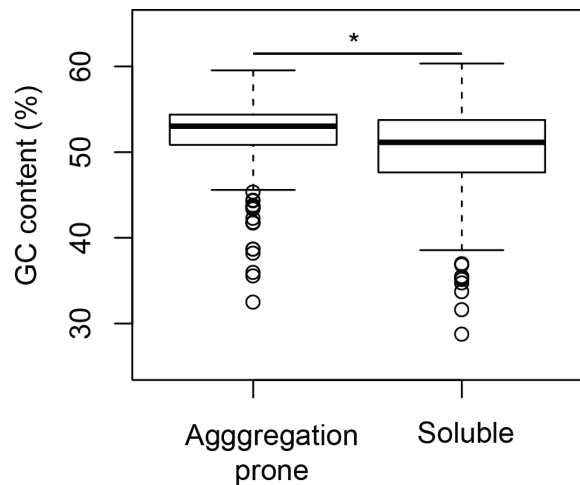


Figure 4. Aggregation-prone proteins are richer in GC-content than soluble proteins. In *E. coli*, aggregation-prone proteins contain higher GC-content than soluble proteins. Mann-Whitney test p-value (*) is 1.3e-15.

GC content and fAg (Figure 5B). This corroborated well with the difference seen between soluble and aggregation-prone proteins in *E. coli* (Figure 4). Thus the increase in the GC composition of a genome may encode proteome that harbours a higher fraction of aggregation-prone proteins.

This is in contrast to previous reports hypothesizing that GC-poor organisms have unstable and aggregation prone proteomes. Notably, the earlier hypothesis that GC poorness is associated with GroEL-dependent aggregation-prone proteomes was based on the observation that GroEL is overexpressed in GC-poor organisms. Therefore, to segregate GroEL-dependent proteins from aggregation-prone proteomes, we developed another classifier (ZENODO, <https://zenodo.org/record/10442/>) trained with 475 curated soluble and 83 GroEL obligate (Class 3 or C3) proteins¹⁴. The classifier achieved an accuracy of 92.29% with MCC of 0.69. We used GDP1.pgc to identify the C3 proteins within aggregation-prone proteins (predicted by SolubEcoli.pgc) to examine the evolution of the GroEL-dependent proteome with GC composition. Indeed we found that the fC3 (fraction of C3 proteins) of bacterial proteome are more correlated with GC content than the fAg fraction (Figure 6A). The phylogenetically independent contrasts of fC3 and GC also correlated strongly (0.7, p-value < 2.2e-16, Figure 6B). The phylum Tenericutes, members of which have GC-poor genomes, was predicted to encode less GroEL-dependent proteins. Mycoplasma and Ureaplasma are the main genera of the phylum Tenericutes and many species of these groups lack GroEL⁴³. In our analysis, we also observed that the Tenericutes without GroEL (red dots in Figure 6A) had very few fC3 proteins. This motivated us to investigate the effect of *groEL* copy number on misfolded proteins. Interestingly, there was a strong correlation between the *groEL* copy number and the

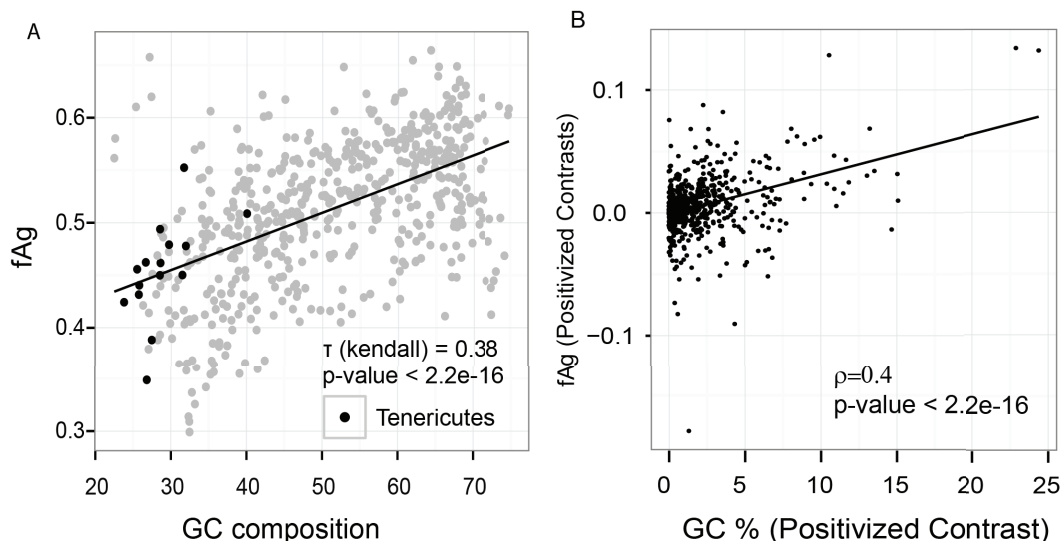


Figure 5. GC content is associated with fAg. (A) GC content of the genome correlates with the fraction of proteome that is aggregation-prone (fAg) (analysis of 570 bacterial genomes using the classifier). Rank-based correlation is provided along with the p-value. The black line shows a linear regression model. (B) The relationship between GC content and fAg was obtained through a phylogenetically independent contrast method (570 bacteria). A positive correlation (0.4) was identified between GC content and fAg (p-value < 2.1e-16).

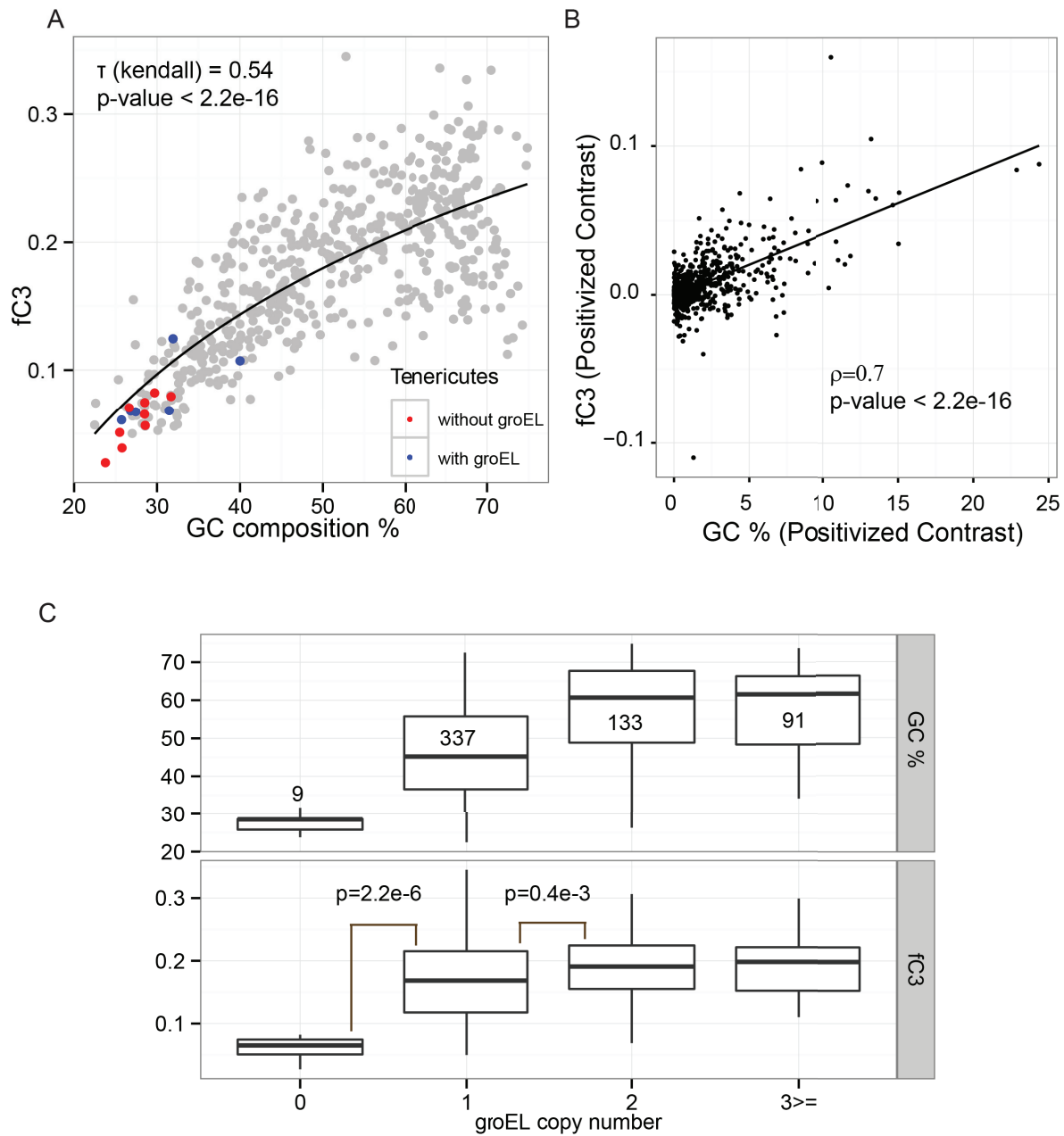


Figure 6. Decrease in GC content is associated with decrease in fC3. (A) Correlation of GC content with the fraction of the proteome that is GroEL obligate (fC3) over 570 bacterial genomes. Members of the phylum Tenericutes with and without the *groEL* gene are coloured in blue and red, respectively. Rank-based correlation is provided along with p-value. The black line shows a logarithmic regression model. (B) A positive correlation (0.7) was identified between independent contrast of GC content and fC3 with respect to phylogenetic information of bacterial genomes (570 bacteria, p-value < 2.2e-16). (C) The organisms were classified based on the number of *groEL* genes present in the genome. fC3 exhibited a significant increase with an increase in the number of genome-encoded *groEL* copies. The p-values were calculated by Mann-Whitney test using two-sided hypothesis.

fraction of genome coding for C3 proteins (Figure 6C). Due to the presence of noise in the experimental data, we tried to benchmark the classifiers. Fujiwara *et al.* reported that five C3 homologs of *groEL*-lacking *Ureaplasma urealyticum* are soluble in GroEL depleted cells²⁶. Hence, we also examined the tolerance of our classifiers by predicting the GroEL dependency of these homologs. Four of these homologs were predicted to be GroEL independent with a high confidence score (Table 3). Overall, the results indicated that C3 proteins and in general aggregation-prone proteins do decrease with the GC content of genomes.

Correlation of GC content with protein solubility is independent of the population bottleneck

Endosymbionts are crucial to this study as the literature suggests that these organisms have undergone bottlenecks during evolution⁴⁴.

It is hypothesized that these organisms have accumulated more deleterious mutations compared to non-endosymbionts⁸. If this were true then endosymbionts should show a greater aggregation propensity or dependence on GroEL than that predicted by the GC content of free-living eubacterial species. To measure the impact of a symbiotic relationship on C3 proteins, we performed an analysis of covariance ANCOVA on 570 eubacterial species⁴². There was no significant effect of a symbiotic relationship on fAG/fC3 (p-value 0.24/0.65, Data set) or significant interaction (p-value=0.36/0.38) with GC composition (Figure 7). Thus we were unable to obtain proof for any association of a bottleneck in evolutionary history with protein aggregation propensity. Therefore we rule out the possibility of bottleneck evolution as the reason for the evolution of GroEL-independent proteomes like *Ureaplasma* and GroEL-independent mycoplasma species.

Table 3. Evaluation of classifiers on five C3 homologous proteins of *groEL*-lacking *Ureaplasma urealyticum*.

The homologous were found in *U. urealyticum* by NCBI BLAST at a threshold of E value of 1e45. Then the aggregation propensity and GroEL dependency of these proteins were classified by SolubEcoli.pgc and GDP1.pgc.

C3 homologous proteins in <i>Ureaplasma urealyticum</i>	E value	Accession	Is aggregation prone? (classifier: SolubEcoli.pgc)	Is GroEL dependent? (classifier: GDP1.pgc)
UuMetK	2e-99	YP_002284849.1	Yes (0.739)	No (0.804)
UuDeoA	2e-80	WP_004026878.1	Yes (0.586)	No (0.938)
UuCsdB	4e-62	D82890	Yes (0.884)	Yes (0.665)
UuGatY	8e-46	H82870	Yes (0.672)	No (0.973)
UuYcfH	7e-41	E82944	Yes (0.518)	No (0.881)

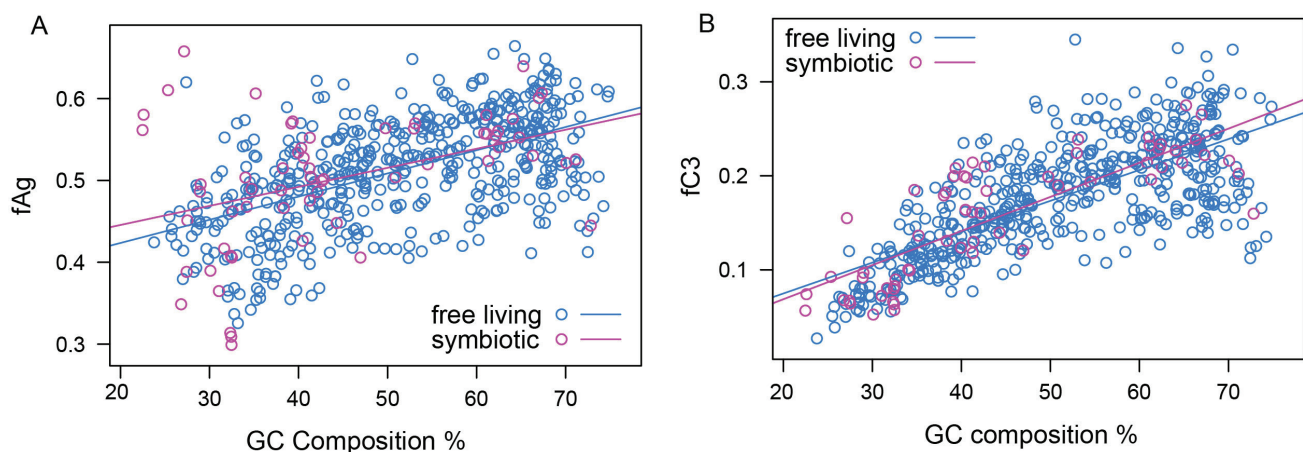


Figure 7. fAG and fC3 are correlated to the GC content independent from the species habitat. The ANCOVA test on 570 organisms showed that a symbiotic relationship has no significant effect or interaction with GC content on the aggregation propensity or GroEL-dependency of the proteins of an organism.

Data set 1. Application of SolubEcoli.pgc and GDP1.pgc classifiers.

1 Data File

<http://dx.doi.org/10.5256/f1000research.4307.d29624>

Proteome wide prediction of GroEL obligatory protein fraction (fC3) and aggregation prone protein fraction (fAg) in 1132 eubacterial genomes with their genome size, GC content and GroEL copy number.

Conclusions

Several machine learning (ML) classifiers have been developed to predict aggregation-prone or GroEL-dependent proteins, but very few of them used data sets generated and curated from multiple experimental studies. Our classifiers were based on curated data from multiple studies and performed well also against the false positive C3 homologs of *Ureaplasma*, showing accuracy and noise tolerance. According to previous theories, GC-poor organisms might have evolved through population bottlenecks. This allows mildly deleterious mutations to be fixed in the population with a high probability^{2,44}. It has been hypothesized that the GC-poor genomes that accumulated a large number of deleterious mutations in the course of evolution, through population bottlenecks and hence harbour proteins that are aggregation-prone. Although overexpressions of chaperones are observed in GC-poor organisms that have reduced genomes, there are also other GC-poor organisms that lack GroEL. Our work provides strong evidence that the general stability of the proteome increases with the decrease in GC content of eubacterial genomes. Decrease in GC content restricts the amino acid space that the organism can sample, thereby compromising protein evolution. We hypothesise that, even with this limited amino acid space, GC-poor organisms are still abundant as growth is facilitated under conditions that compromise protein folding capacity. This antagonism

between ability to evolve and folding advantage could be crucial in facilitating protein evolution in the presence of chaperones and other folding machineries^{45–48}.

Our work suggests that organisms facing continuous proteostasis problems would tend to shift towards a more GC-poor genome. This is supported by findings of Xia *et al.*⁴⁹ who have reported that the preponderance of GC to AT conversions during high temperature laboratory adaptation of *Pasteurella multocida*. Further *in vitro* evolution experiments will be required to demonstrate whether laboratory adaptation to low GC content may provide folding advantage.

Data availability

F1000Research: Dataset 1. Application of SolubEcoli.pgc and GDP1.pgc classifiers, [10.5256/f1000research.4307.d29624](https://doi.org/10.5256/f1000research.4307.d29624)⁵⁵.

ZENODO: Training data of protein classifier SolubEcoli.pgc and GDP1.pgc, doi: [10.5281/zenodo.10442](https://doi.org/10.5281/zenodo.10442)⁵⁶.

Author contributions

KC and DD designed the study. RDR carried out all the analysis. MB and VB participated in the design and discussion of the study. RDR, KC and DD prepared the manuscript and all authors agreed to the final content.

Competing interests

No competing interests were disclosed.

Grant information

RDR acknowledges HP and CSIR for funding the study, DD and KC acknowledge CSIR for funding through IGIB. KC acknowledges Wellcome Trust – DBT India alliance for funding and fellowship.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

- Nishida H: **Evolution of genome base composition and genome size in bacteria.** *Front Microbiol.* 2012; **3**: 420.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- McCutcheon JP, Moran NA: **Extreme genome reduction in symbiotic bacteria.** *Nat Rev Microbiol.* 2012; **10**(1): 13–26.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Guo FB, Lin H, Huang J: **A plot of G + C content against sequence length of 640 bacterial chromosomes shows the points are widely scattered in the upper triangular area.** *Chromosome Res.* 2009; **17**(3): 359–364.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Lightfield J, Fram NR, Ely B: **Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage.** *PLoS One.* 2011; **6**(3): e17677.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- van Ham RC, Kamerbeek J, Palacios C, *et al.*: **Reductive genome evolution in *Buchnera aphidicola*.** *Proc Natl Acad Sci U S A.* 2003; **100**(2): 581–586.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bastolla U, Moya A, Viguera E, *et al.*: **Genomic determinants of protein folding thermodynamics in prokaryotic organisms.** *J Mol Biol.* 2004; **343**(5): 1451–1466.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Fares MA, Moya A, Barrio E: **GroEL and the maintenance of bacterial endosymbiosis.** *Trends Genet.* 2004; **20**(9): 413–416.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Fares MA, Ruiz-Gonzalez MX, Moya A, *et al.*: **Endosymbiotic bacteria: groEL buffers against deleterious mutations.** *Nature.* 2002; **417**(6887): 398.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Moran NA: **Accelerated evolution and Muller's ratchet in endosymbiotic bacteria.** *Proc Natl Acad Sci U S A.* 1996; **93**(7): 2873–2878.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Aksoy S: **Molecular analysis of the endosymbionts of tsetse flies: 16S rDNA locus and over-expression of a chaperonin.** *Insect Mol Biol.* 1995; **4**(1): 23–29.
[PubMed Abstract](#)
- Clark MA, Baumann L, Baumann P: **Sequence analysis of a 34.7-kb DNA segment from the genome of *Buchnera aphidicola* (endosymbiont of aphids) containing groEL, dnaA, the atp operon, gidA, and rho.** *Curr Microbiol.* 1998; **36**(3): 158–163.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Wilcox JL, Dunbar HE, Wolfinger RD, *et al.*: **Consequences of reductive evolution for gene expression in an obligate endosymbiont.** *Mol Microbiol.* 2003; **48**(6): 1491–1500.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Williams TA, Fares MA: **The effect of chaperonin buffering on protein evolution.**

- Genome Biol Evol.* 2010; 2: 609–619.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Kerner MJ, Naylor DJ, Ishihama Y, *et al.*: **Proteome-wide analysis of chaperonin-dependent protein folding in *Escherichia coli*.** *Cell.* 2005; 122(2): 209–220.
[PubMed Abstract](#) | [Publisher Full Text](#)
 15. Chapman E, Farr GW, Usaite R, *et al.*: **Global aggregation of newly translated proteins in an *Escherichia coli* strain deficient of the chaperonin GroEL.** *Proc Natl Acad Sci U S A.* 2006; 103(43): 15800–15805.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 16. Raineri E, Ribeca P, Serrano L, *et al.*: **A more precise characterization of chaperonin substrates.** *Bioinformatics.* 2010; 26(14): 1685–1689.
[PubMed Abstract](#) | [Publisher Full Text](#)
 17. Bogumil D, Dagan T: **Chaperonin-dependent accelerated substitution rates in prokaryotes.** *Genome Biol Evol.* 2010; 2: 602–608.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 18. Niwa T, Kanamori T, Ueda T, *et al.*: **Global analysis of chaperone effects using a reconstituted cell-free translation system.** *Proc Natl Acad Sci U S A.* 2012; 109(23): 8937–8942.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 19. Calloni G, Chen T, Schermann SM, *et al.*: **DnaK Functions as a Central Hub in the *E. coli* Chaperone Network.** *Cell Rep.* 2012; 1(3): 251–264.
[PubMed Abstract](#) | [Publisher Full Text](#)
 20. Tartaglia GG, Dobson CM, Hartl FU, *et al.*: **Physicochemical determinants of chaperone requirements.** *J Mol Biol.* 2010; 400(3): 579–588.
[PubMed Abstract](#) | [Publisher Full Text](#)
 21. Noivirt-Brik O, Unger R, Horovitz A: **Low folding propensity and high translation efficiency distinguish *in vivo* substrates of GroEL from other *Escherichia coli* proteins.** *Bioinformatics.* 2007; 23(24): 3276–3279.
[PubMed Abstract](#) | [Publisher Full Text](#)
 22. Fang Y, Fang J: **Discrimination of soluble and aggregation-prone proteins based on sequence information.** *Mol BioSyst.* 2013; 9(4): 806–811.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 23. Stiglic G, Kocbek S, Pernek I, *et al.*: **Comprehensive decision tree models in bioinformatics.** *PLoS One.* 2012; 7(3): e33812.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 24. Klus P, Bolognesi B, Agostini F, *et al.*: **The cleverSuite Approach for Protein Characterization: Predictions of Structural Properties, Solubility, Chaperone Requirements and RNA-Binding Abilities.** *Bioinformatics.* 2014; 30(11): 1601–1608.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 25. Niwa T, Ying BW, Saito K, *et al.*: **Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins.** *Proc Natl Acad Sci U S A.* 2009; 106(11): 4201–4206.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 26. Fujiwara K, Ishihama Y, Nakahigashi K, *et al.*: **A systematic survey of *in vivo* obligate chaperonin-dependent substrates.** *EMBO J.* 2010; 29(9): 1552–1564.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 27. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics.* 2006; 22(13): 1658–1659.
[PubMed Abstract](#) | [Publisher Full Text](#)
 28. Das Roy R, Dash D: **Selection of relevant features from amino acids enables development of robust classifiers.** *Amino Acids.* 2014; 46(5): 1343–1351.
[PubMed Abstract](#) | [Publisher Full Text](#)
 29. Zweig MH, Campbell G: **Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine.** *Clin Chem.* 1993; 39(4): 561–577.
[PubMed Abstract](#)
 30. Uchiyama I, Higuchi T, Kawai M: **MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity.** *Nucleic Acids Res.* 2010; 38(Database issue): D361–D365.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 31. Hill JE, Penny SL, Crowell KG, *et al.*: **cpnDB: a chaperonin sequence database.** *Genome Res.* 2004; 14(8): 1669–1675.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 32. Pruesse E, Quast C, Knittel K, *et al.*: **SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.** *Nucleic Acids Res.* 2007; 35(21): 7188–7196.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 33. Team RC: **R: A language and environment for statistical computing.** *R foundation for Statistical Computing.* 2005.
[Reference Source](#)
 34. Midford P, Garland T Jr, Maddison W: **PDAP Package of Mesquite. Version 1.14.** 2008.
[Reference Source](#)
 35. **Mesquite: a modular system for evolutionary analysis. Version 2.75.**
[Reference Source](#)
 36. Vapnik V: **The nature of statistical learning theory.** 2nd edition. Springer; 1999.
[Reference Source](#)
 37. Chou KC: **Prediction of protein cellular attributes using pseudo-amino acid composition.** *Proteins.* 2001; 43(3): 246–255.
[PubMed Abstract](#) | [Publisher Full Text](#)
 38. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, *et al.*: **FoldIndex@: a simple tool to predict whether a given protein sequence is intrinsically unfolded.** *Bioinformatics.* 2005; 21(16): 3435–3438.
[PubMed Abstract](#) | [Publisher Full Text](#)
 39. Singh GP, Dash D: **Electrostatic mis-interactions cause overexpression toxicity of proteins in *E. coli*.** *PLoS One.* 2013; 8(5): e64893.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 40. Bohlin J, Brynildsrud O, Vesth T, *et al.*: **Amino acid usage is asymmetrically biased in AT- and GC-rich microbial genomes.** *PLoS One.* 2013; 8(7): e69878.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 41. Garland T, Bennett AF, Rezende EL: **Phylogenetic approaches in comparative physiology.** *J Exp Biol.* 2005; 208(Pt 16): 3015–3035.
[PubMed Abstract](#) | [Publisher Full Text](#)
 42. Mazurie A, Bonchev D, Schwikowski B, *et al.*: **Evolution of metabolic network organization.** *BMC Syst Biol.* 2010; 4: 59.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 43. Clark GW, Tillier ER: **Loss and gain of GroEL in the Mollicutes.** *Biochem Cell Biol.* 2010; 88(2): 185–194.
[PubMed Abstract](#) | [Publisher Full Text](#)
 44. Mira A, Moran NA: **Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria.** *Microb Ecol.* 2002; 44(2): 137–143.
[PubMed Abstract](#) | [Publisher Full Text](#)
 45. Bandyopadhyay A, Saxena K, Kasturia N, *et al.*: **Chemical chaperones assist intracellular folding to buffer mutational variations.** *Nat Chem Biol.* 2012; 8(3): 238–245.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 46. Rutherford SL, Lindquist S: **Hsp90 as a capacitor for morphological evolution.** *Nature.* 1998; 396(6709): 336–342.
[PubMed Abstract](#) | [Publisher Full Text](#)
 47. Queitsch C, Sangster TA, Lindquist S: **Hsp90 as a capacitor of phenotypic variation.** *Nature.* 2002; 417(6889): 618–624.
[PubMed Abstract](#) | [Publisher Full Text](#)
 48. Rohner N, Jarosz DF, Kowalko JE, *et al.*: **Cryptic variation in morphological evolution: HSP90 as a capacitor for loss of eyes in cavefish.** *Science.* 2013; 342(6164): 1372–1375.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 49. Xia X, Wei T, Xie Z, *et al.*: **Genomic changes in nucleotide and dinucleotide frequencies in *Pasteurella multocida* cultured under high temperature.** *Genetics.* 2002; 161(4): 1385–1394.
[PubMed Abstract](#)
 50. Chou KC: **Prediction of protein subcellular locations by incorporating quasi-sequence-order effect.** *Biochem Biophys Res Commun.* 2000; 278(2): 477–483.
[PubMed Abstract](#) | [Publisher Full Text](#)
 51. Bum Ju L, Keun Ho R: **Feature Extraction from Protein Sequences and Classification of Enzyme Function.** In *BioMedical Engineering and Informatics, 2008 BMEI 2008 International Conference on*; 27–30 May 2008. 2008; 138–142.
[Publisher Full Text](#)
 52. Shamim MTA, Anwaruddin M, Nagarajaram HA: **Support Vector Machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs.** *Bioinformatics.* 2007; 23(24): 3320–3327.
[PubMed Abstract](#) | [Publisher Full Text](#)
 53. Li ZR, Lin HH, Han LY, *et al.*: **PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence.** *Nucleic Acids Res.* 2006; 34(Web Server issue): W32–W37.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 54. Dubchak I, Muchnik I, Holbrook SR, *et al.*: **Prediction of protein folding class using global description of amino acid sequence.** *Proc Natl Acad Sci U S A.* 1995; 92(19): 8700–8704.
[PubMed Abstract](#) | [Free Full Text](#)
 55. Das Roy R, Bhardwaj M, Bhatnagar V, *et al.*: **Application of SolubEcoli.pgc and GDP1.pgc classifiers.** *F1000Research.*
[Data Source](#)
 56. Das Roy R, Chakraborty K, Das D: **Training data of protein classifier SolubEcoli.pgc and GDP1.pgc.** 2014.
[Data Source](#)

Open Peer Review

Current Referee Status:



Version 1

Referee Report 01 October 2014

doi:10.5256/f1000research.4611.r6283



Annalisa Pastore

MRC National Institute for Medical Research, London, UK

The genesis of this paper is the proposal that genomes containing a poor percentage of guanosine and cytosine (GC) nucleotide pairs lead to proteomes more prone to aggregation than those encoded by GC-rich genomes. As a consequence these organisms are also more dependent on the protein folding machinery. If true, this interesting hypothesis could establish a direct link between the tendency to aggregate and the genomic code.

In their paper, the authors have tested the hypothesis on the genomes of eubacteria using a genome-wide approach based on multiple machine learning models. Eubacteria are an interesting set of organisms which have an appreciably high variation in their nucleotide composition with the percentage of CG genetic material ranging from 20% to 70%. The authors classified different eubacterial proteomes in terms of their aggregation propensity and chaperone-dependence. For this purpose, new classifiers had to be developed which were based on carefully curated data. They took account for twenty-four different features among which are sequence patterns, the pseudo amino acid composition of phenylalanine, aspartic and glutamic acid, the distribution of positively charged amino acids, the FoldIndex score and the hydrophobicity. These classifiers seem to be altogether more accurate and robust than previous such parameters.

The authors found that, contrary to what expected from the working hypothesis, which would predict a decrease in protein aggregation with an increase in GC richness, the aggregation propensity of proteomes increases with the GC content and thus the stability of the proteome against aggregation increases with the decrease in GC content. The work also established a direct correlation between GC-poor proteomes and a lower dependence on GroEL. The authors conclude by proposing that a decrease in eubacterial GC content may have been selected in organisms facing proteostasis problems. A way to test the overall results would be through *in vitro* evolution experiments aimed at testing whether adaptation to low GC content provide folding advantage.

The main strengths of this paper is that it addresses an interesting and timely question, finds a novel solution based on a carefully selected set of rules, and provides a clear answer. As such this article represents an excellent and elegant bioinformatics genome-wide study which will almost certainly influence our thinking about protein aggregation and evolution. Some of the weaknesses are the not always easy readability of the text which establishes unclear logical links between concepts.

Another possible criticism could be that, as any *in silico* study, it makes strong assumptions on the sequence features that lead to aggregation and strongly relies on the quality of the classifiers used. Even

though the developed classifiers seem to be more robust than previous such parameters, they remain only overall indications which can only allow statistical considerations. It could of course be argued that this is good enough to reach meaningful conclusions in this specific case.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 15 July 2014

doi:[10.5256/f1000research.4611.r5273](https://doi.org/10.5256/f1000research.4611.r5273)



Amnon Horovitz

Department of Structural Biology, Weizmann Institute of Science, Rehovot, Israel

In this study, the authors describe machine-learning classifiers for predicting aggregation propensities of proteins. A novel aspect of this work is that the classifiers are based on experimental data obtained from different sources regarding chaperone dependence (GroEL or DnaK) and solubility in chaperone-free systems. The authors then use their machine-learning approach to test for ~1100 eubacterial proteomes whether a low GC content correlates with a greater tendency to aggregate or mis-fold as suggested by earlier studies. One possibility is that the GC content affects the amino acid composition of proteins that, in turn, affects their folding and aggregation propensities. The authors show in Figure 3 of the paper that amino acid compositions do indeed correlate strongly with GC content (the Figure shows data for all 20 naturally occurring amino acids, although its title suggests otherwise). They then show that aggregation-prone proteins have a higher GC content than soluble proteins. This finding is not really new since it has already been reported (see ref. 21 in the paper) that GroEL substrates tend to have a relatively high GC content. However, the observation for a large number of genomes that the fraction of aggregation-prone proteins increases with an increasing GC content is novel. A weakness of the paper is that the authors do not discuss the problems of distinguishing between (i) causation and correlation; and (ii) cause and effect. In other words, is the correlation between high GC content and an increased tendency to mis-fold due to some unidentified factor(s) that correlates with both GC content and mis-folding or to a direct effect? In addition, it is also possible (at least in principle) that a high GC content reflects selection against mis-folding rather than being one of its 'causes'.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.
