Synthetic and Systems Biotechnology 9 (2024) 304–311

Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: www.keaipublishing.com/en/journals/synthetic-and-systems-biotechnology

Original Research Article

Omics data analysis reveals the system-level constraint on cellular amino acid composition

Yuanyuan Huang [a,c,d,1], Zhitao Mao [c,d,*,1], Yue Zhang [a,c,d,1], Jianxiao Zhao [c,d,e], Xiaodi Luan [a,c,d], Ke Wu [c,d], Lili Yun [f], Jing Yu [c,d], Zhenkun Shi [c,d], Xiaoping Liao [b,c,d,**], Hongwu Ma [c,d,***]

[a] College of Biotechnology, Tianjin University of Science and Technology, Tianjin, 300457, China
[b] Haihe Laboratory of Synthetic Biology, Tianjin, 300308, China
[c] Biodesign Center, Key Laboratory of Engineering Biology for Low-carbon Manufacturing, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, 300308, China
[d] National Center of Technology Innovation for Synthetic Biology, Tianjin, 300308, China
[e] Frontier Science Center for Synthetic Biology and Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin, 300072, China
[f] Tianjin Medical Laboratory, BGI-Tianjin, BGI-Shenzhen, Tianjin, 300308, China

ARTICLE INFO

Keywords:
Omics data
System-level constraint
Amino acid composition

ABSTRACT

Proteins play a pivotal role in coordinating the functions of organisms, essentially governing their traits, as the dynamic arrangement of diverse amino acids leads to a multitude of folded configurations within peptide chains. Despite dynamic changes in amino acid composition of an individual protein (referred to as AAP) and great variance in protein expression levels under different conditions, our study, utilizing transcriptomics data from four model organisms uncovers surprising stability in the overall amino acid composition of the total cellular proteins (referred to as AACell). Although this value may vary between different species, we observed no significant differences among distinct strains of the same species. This indicates that organisms enforce system-level constraints to maintain a consistent AACell, even amid fluctuations in AAP and protein expression. Further exploration of this phenomenon promises insights into the intricate mechanisms orchestrating cellular protein expression and adaptation to varying environmental challenges.

1. Introduction

Proteins, comprised of 20 amino acids (AAs), are vital components of the biological system, serving diverse functions in metabolic reactions, cell growth regulation, signal transmission, structural support, and immune defense within living organisms [1]. When the environment changes, cells can sense these changes and adapt to them by adjusting gene expression [2]. AA composition of an individual protein (referred to as AAP) can be easily calculated from the protein sequence, and it is well known that different proteins exhibit distinct AA compositions [3–5]. Additionally, the expression levels of numerous proteins can vary significantly under different cultivation conditions due to gene regulation [6–8]. In the context of general reasoning, one would expect that the AA composition for all cellular proteins (referred to as AACell) may have great variance under different cultivation conditions.

On the other hand, AACell can also be measured experimentally, but the experimental methods come with a certain degree of error and difficulty in distinguishing between intracellular aspartate and asparagine, as well as glutamate and glutamine. Based on real experimental data, Choi et al. discovered that the AACell in *Escherichia coli* remains

Peer review under responsibility of KeAi Communications Co., Ltd.
* Corresponding author. Biodesign Center, Key Laboratory of Engineering Biology for Low-carbon Manufacturing, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, 300308, China.
** Corresponding author. Haihe Laboratory of Synthetic Biology, Tianjin, 300308, China.
*** Corresponding author. Biodesign Center, Key Laboratory of Engineering Biology for Low-carbon Manufacturing, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, 300308, China.
E-mail addresses: mao_zt@tib.cas.cn (Z. Mao), liao_xp@tib.cas.cn (X. Liao), ma_hw@tib.cas.cn (H. Ma).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.synbio.2024.03.001
Received 12 October 2023; Received in revised form 1 March 2024; Accepted 1 March 2024
Available online 5 March 2024

relatively stable among 11 studies [9]. Among 20 amino acids, only lysine (Lys) and glycine (Gly) exhibited higher coefficient of variation (CV) values. Similar trends were observed in *Saccharomyces cerevisiae*, yet this conclusion was based on data from only three studies, necessitating further validation through statistical analysis. Due to a lack of experimental data, Chen et al. utilized proteome data to calculate AACell [10], revealing its stability through the analysis of 30 datasets of *S. cerevisiae*. The great variance on AAP and the extensive regulation of protein expression levels across conditions seems to be in contrast with the relatively stable AACell measured experimentally. It is desirable to investigate whether AACell remains stable across a wide variety of cultivation conditions and test the preliminary finding for more species.

The wide availability of transcriptomics data offers a new option to study AACell under a much wide range of conditions by calculating AACell from AAP and the protein (gene) expression levels. In the present study, we collected transcriptome data for model organisms, such as *E. coli, S. cerevisiae*, *Bacillus subtilis*, and *Corynebacterium glutamicum* from literature and public databases. Surprisingly, we found that although the AAP of individual proteins and expression levels of proteins varied significantly among different samples, the AACell was almost constant. This suggests that organisms impose system-level constraints on protein expression to maintain the AACell of the entire cell.

2. Materials and methods

2.1. Calculation of AAPs of a particular protein

The mass ratio of amino acid i in a particular protein j (g/g) was calculated using the following formula (Eq. 1), based on the protein sequence:

$$AAP_{ij} = \frac{N_{ij} * MW_i}{PMW_j} \quad (1)$$

Here, MW_i represents the molecular weight of AA i, N_{ij} represents the number of AA i in protein j, and PMW_j is the molecular weight of the protein j. Protein sequences and molecular weight data for the organisms used in the study were obtained through the UniProt API [11].

2.2. Calculation of the mass ratio of individual proteins in all cellular proteins at a particular condition

To calculate the mass percentage (g/g total protein) of a protein in the whole transcriptome of a sample, we used Eq. (2):

$$MP_{jk} = \frac{E_{jk} * PMW_j}{\sum_{l}^{n} E_{lk} * PMW_l} \quad (2)$$

Here, E_{jk} represents the expression value of protein j at a particular condition k, PMW_j represents the molecular weight of protein j, n is the total number of proteins, and l takes values between 1 and n.

2.3. Calculation of AACells at a particular condition

To determine the AACells of different samples, we calculated the mass distribution of AAs, such as alanine, arginine, and valine, per unit mass of total protein. This was achieved by multiplying AAP_{ij} by MP_{jk} (Eq. (3)).

$$AACell_{ik} = \sum_{j}^{n} AAP_{ij} * MP_{jk} \quad (3)$$

Here, $AACell_{ik}$ represents the mass ratio of amino acid i in all cellular proteins at a particular condition k (g/g), AAP_{ij} represents the mass ratio of AA i in a particular protein j (g/g), MP_{jk} represents the mass ratio of protein j in all cellular protein at a particular condition k (g/g).

2.4. Acquisition of omics data for E. coli

In order to accurately calculate the AACell of different samples, protein sequence, molecular weight, and absolute level protein abundance are needed. However, current technology limits the available proteome, making it challenging to cover all proteins in the cell. Furthermore, recent studies have shown that in *E. coli*, changes in gene expression (and hence final protein concentrations) are mostly determined at the transcriptional stage, and researchers have provided simple quantitative formulas to link regulation to mRNA and protein levels [12]. So, in this study, we used gene expression levels from the transcriptome as a proxy for protein expression levels. The transcriptome of *E. coli* was obtained from Ecomics [13], which converted relative expression measurements to absolute RNA copies per cell. To ensure data quality, we only included four strains (DH1, W3110, BW25113, and MG1655) with more than 100 samples.

2.5. Acquisition of omics data for other species

To study the differences in AACell among different species, we obtained transcriptome data for *S. cerevisiae*, *B. subtilis*, and *C. glutamicum*. Same as the case of *E. coli*, we acquired relative quantitative transcriptome data from literature and databases. For *B. subtilis* (microarray) and *S. cerevisiae* (single-cell RNA-Seq), we directly obtained the transcriptome data from literature-processed expression data [14,15]. For *C. glutamicum*, we retrieved samples from the NCBI Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra, published before Jan 12, 2023) and processed the raw sequencing files using the prokaryotic RNA-seq processing pipeline (https://github.com/avsastry/modulome-workflow). Briefly, the pipeline utilizes fastq-dump (https://github.com/ncbi/sra-tools/wiki/HowTo: fasterq-dump), Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore), FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for raw data processing. Sequencing reads were aligned to the reference genome (BA000036.3) using Bowtie2 [16]. Read counts were generated by using featureCounts [17] and converted into log2 transcripts per million (TPM).

2.6. Statistical analysis

We calculated the standard deviation (SD) and coefficient of variation (CV) using Equations (4) and (5) to assess the dispersion of the dataset, including the number of AAs, MP, AAP and AACell. Furthermore, we calculated the Pearson correlation coefficient (PCC) [18] to investigate the characteristics of the AACell (mean value in different experiments) of AAs in different species, as described in Equation (6). Additionally, we conducted an independent-samples *t*-test to compare the MP under different conditions. When the proteins had equal population variances, we conducted a standard independent 2-sample test; otherwise, we utilized Welch's *t*-test. All of the statistical analyses were performed using Python.

$$SD = \sqrt{\frac{\sum_{i}^{N}(x_i - \mu)^2}{N}} \quad (4)$$

$$CV = \frac{SD}{\mu} \quad (5)$$

$$PCC_{data_1, data_2} = \frac{cov(data_1, data_2)}{SD_{data_1} * SD_{data_2}} \quad (6)$$

Here, N represents the number of proteins or cultivation conditions, μ stands for the mean value, *data* represents the dataset for calculation, and *cov* stands for the covariance between two datasets.

3. Results

3.1. The proportion of the same AA varies greatly among different proteins

The protein sequences and molecular weight information for the organisms included in this study were retrieved using the UniProt API, which provided us with 4082 *E. coli* protein sequences, 4160 *B. subtilis* protein sequences, 5619 *S. cerevisiae* protein sequences, and 3089 *C. glutamicum* protein sequences, as well as their corresponding molecular weight data. The ranges of protein molecular weight across different species were quite similar, with a distribution between 1.7 kDa and 609.6 kDa (Fig. S1). While the average protein molecular weights of *B. subtilis* (32.98 ± 29.7 kDa), *E. coli* (35.54 ± 22.81 kDa), and *C. glutamicum* (33.72 ± 23.85 kDa) were comparable, *S. cerevisiae* (54.55 ± 41.84 kDa) exhibited the highest average molecular weight. We also observed significant variations in the AAP among the proteins in different organisms, with a CV exceeding 0.27 for all AAs (Fig. 1 and Table S1). Moreover, we observed that not all proteins encompass the entirety of the twenty AAs, with tryptophan, methionine, cysteine, and histidine being notably scarce in the majority. Current research indicates that, spanning various taxonomic divisions, including bacteria, archaea, and eukaryotes, the occurrence frequency of cysteine, histidine, methionine, and tryptophan in proteins is consistently below 3% [19]. It is worth noting that tryptophan and methionine, in particular, are each encoded by only one codon, while cysteine and histidine have two codons (Fig. S2). For AAs with exceptionally high AAP, such as leucine, arginine, and valine, the number of codons is indeed relatively high (Fig. S2). However, the correlation between the remaining AAs and the number of codons is relatively low (Fig. S2). Therefore, the number of codons encoding AAs may be a crucial factor influencing the AAP in proteins.

3.2. The MP varies greatly under different conditions

Protein expression levels can vary under different cultivation conditions, but it is unclear if the MP of proteins (as defined in Eq. (2)) would also differ. To investigate this, we obtained transcriptomic data for *E. coli*, *B. subtilis*, *S. cerevisiae*, and *C. glutamicum* from literature and databases (Table 1). For *E. coli*, we obtained transcriptomic data for 4080 proteins under 3034 cultivation conditions from the Ecomics database [13], including 2307 MG1655 strains, 344 BW25113 strains, 280 W3110 strains, and 103 DH1 strains. Transcriptomic data for *S. cerevisiae* and *B. subtilis* were obtained from the literature [14,15], including single-cell RNA-Seq data for 5612 proteins under 175 cultivation conditions and microarray data for 4160 proteins under 269 cultivation conditions, respectively. In addition, we analyzed RNA-Seq data for 3081 proteins under 292 cultivation conditions for *C. glutamicum* from the SRA database. Upon utilizing Eq. (2), we subsequently computed the MP of proteins across various experimental trials. Significant variations in the MP of these proteins were revealed across different cultivation conditions. To illustrate, within *E. coli* MG1655, an impressive majority of over 97.2% of the protein pairs manifest discernible variances in inter-MPs (*t*-test, $P < 0.05$) (Fig. S3). As for the *C. glutamicum*, *S. cerevisiae*, and *B. subtilis*, the corresponding ratios approximate 96.3%, 87.8%, and 98.4% respectively (Fig. S3). By sorting protein-coding genes according to their median MP levels across the experiments, we observed that MP levels varied significantly among proteins (Fig. 2). Besides, we found that even for the same protein, there is a significant difference in MP values under different cultivation conditions (Fig. 2).

Table 1
Sources and statistics of omics data.

Species	Experiment number	Protein number	Types	Sources
E. coli BW25113	344	4080	Microarray and RNA_seq	Ecomics [13]
E. coli MG1655	2307	4080	Microarray and RNA_seq	Ecomics
E. coli W3110	280	4080	Microarray and RNA_seq	Ecomics
E. coli DH1	103	4080	Microarray and RNA_seq	Ecomics
S. cerevisiae	175	5612	Single-cell RNA-Seq	Literature [14]
B. subtilis	269	4160	Microarray	Literature [15]
C. glutamicum	292	3081	RNA-Seq	NCBI Sequence Read Archive

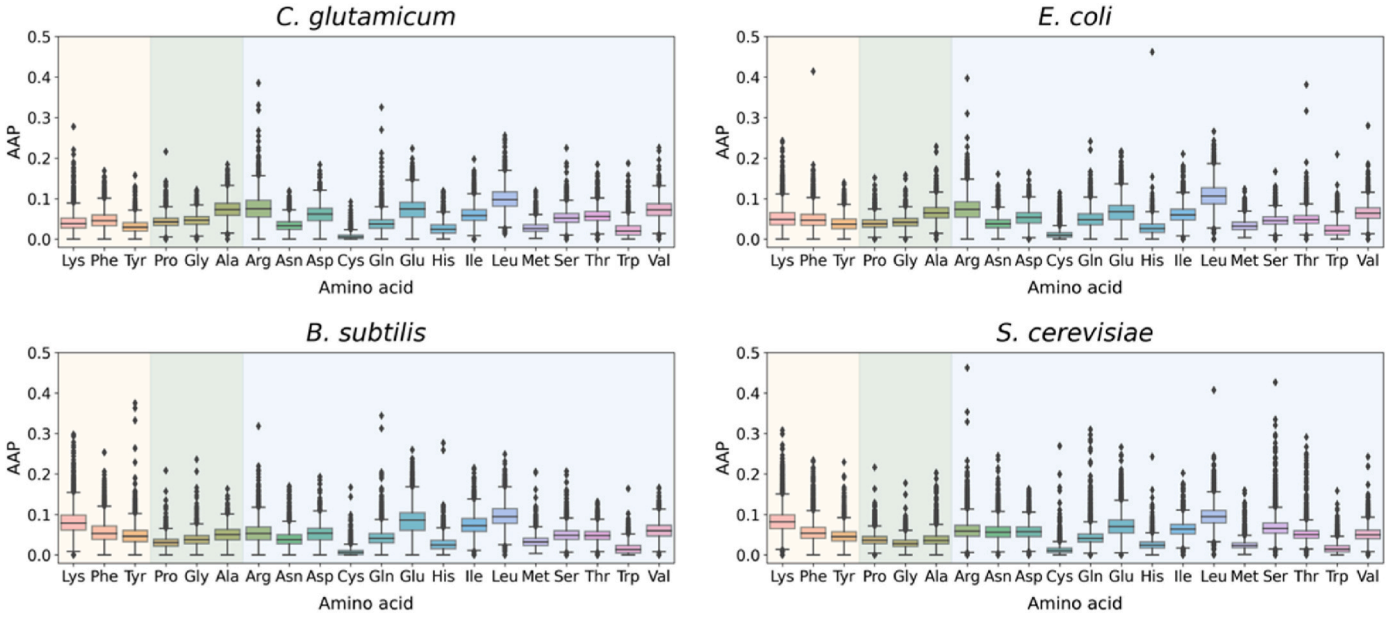

Fig. 1. The distribution of AAP in various species. Twenty AAs are classified into three categories: AAs coded by AT-rich codons (off yellow), AAs coded by GC-rich codons (sea mist), and other AAs (azureish white). The meanings of the abbreviations can be found in the abbreviation table.

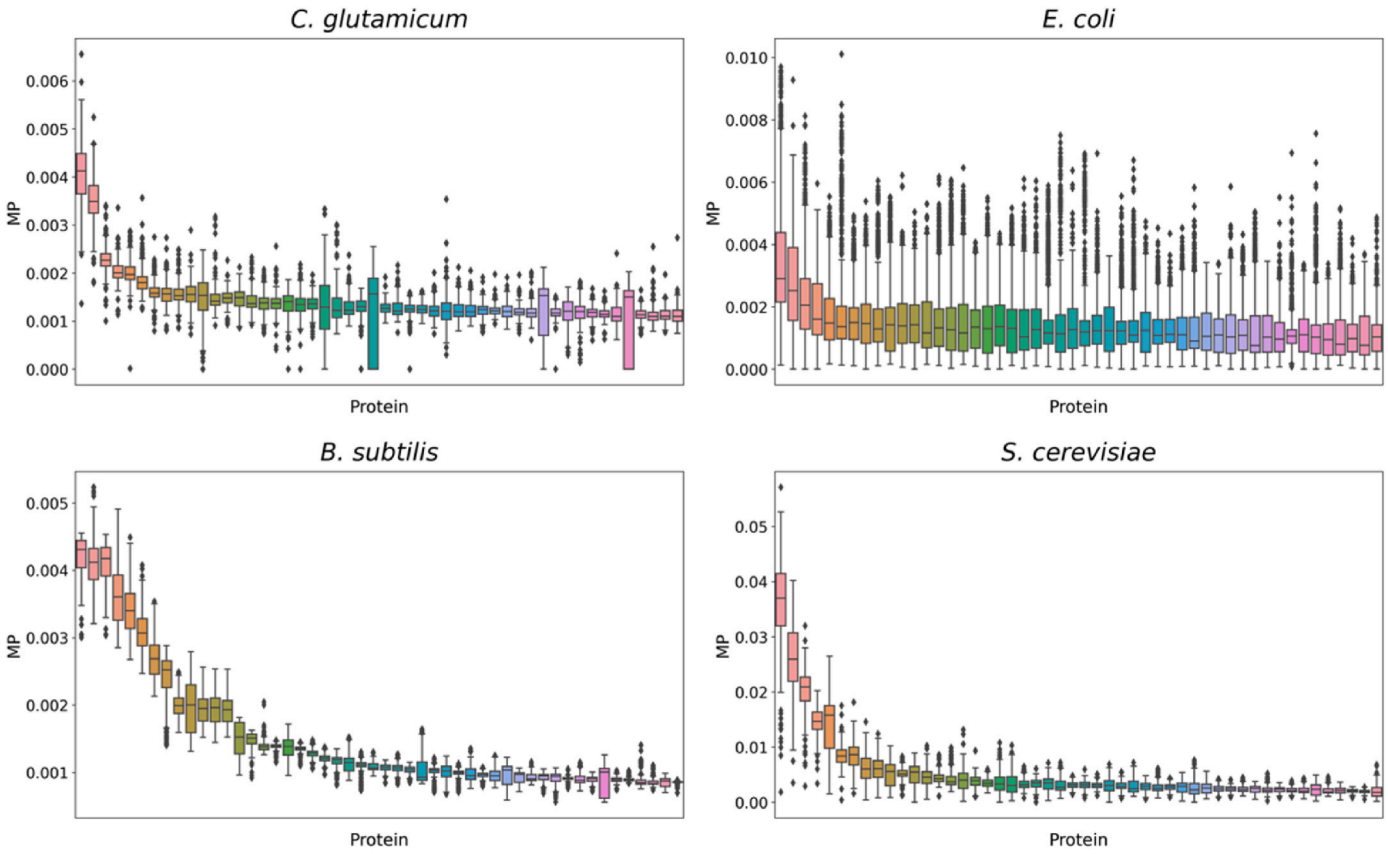

Fig. 2. The distribution of MP for different proteins (only the proteins with a median MP among the top 50 are displayed). The plotting procedure involves calculating the median MP for each protein, sorting them in descending order, and then generating MP distribution plots for the top 50 proteins. The meanings of the abbreviations can be found in the abbreviation table.

3.3. The AACell in four model organisms remains stable across different conditions

It's crucial to emphasize that existing literature predominantly concentrates on analyzing protein sequences when considering the aspect of AAP. The exploration of AACell, particularly concerning protein abundance, has received comparatively limited attention. Our study seeks to bridge this gap by precisely examining how the abundance of proteins correlates with their AACell. In doing so, we aim to provide novel insights into the relationship between protein abundance and

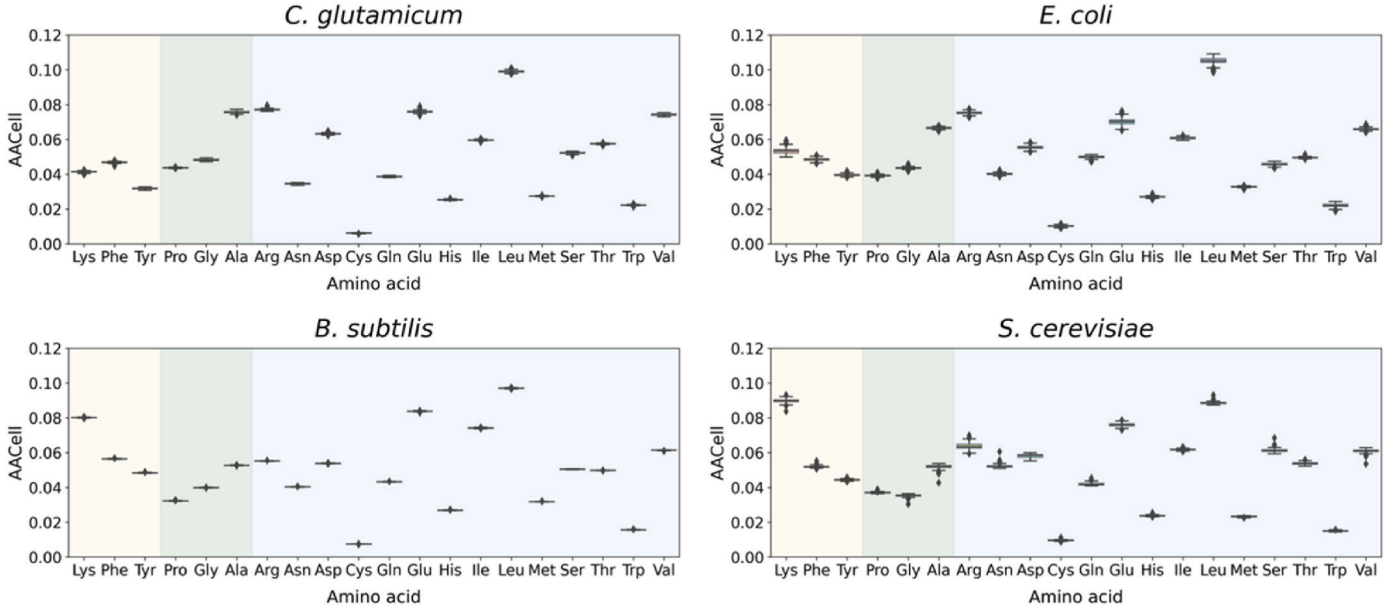

Fig. 3. The distribution of AACell in different species. Twenty AAs are classified into three categories: AAs coded by AT-rich codons (off yellow), AAs coded by GC-rich codons (sea mist), and other AAs (azureish white). The meanings of the abbreviations can be found in the abbreviation table.

AACell, thereby enhancing our understanding of cellular processes at the molecular level.

We analyzed data from *E. coli* MG1655, *B. subtilis*, *S. cerevisiae*, and *C. glutamicum*, and revealed that when we compared the AACell, we found that there was a systemic stability of AA levels in these four organisms, with a CV being less than 0.06 for all AAs (Fig. 3 and Table S2). Further analysis revealed that in *E. coli*, AACell calculated through transcriptome and the AACell computed by Choi et al. [9] using experimental data exhibited a high PCC of 0.94 (Table S3). Similarly, in *S. cerevisiae*, the AACell calculation results showed a PCC of 0.987 concerning Chen's AACell computation [10] based on the proteome (Table S4). Additionally, we conducted a comparison between the AACell obtained through transcriptome analysis and the AACell derived from four models (iML1515^{R} [20], iCW773^{R} [21], iBsu1147^{R} [22], and Yeast8 [23]). Remarkably, we observed a high level of consistency (PCC approximately 0.86) between the calculated AACell and the model-derived AACell for three species: *E. coli*, *S. cerevisiae*, and *B. subtilis* (Table S5). Furthermore, our analysis revealed significant discrepancies in the model for *C. glutamicum* (iCW773^{R}), where the levels of alanine, glutamate, and glutamine were notably higher than the computationally derived AACell and exceeded the corresponding values in the other three species (Table S5). This anomaly resulted in a lower PCC of 0.717 between the computed AACell and the model-derived AACell for *C. glutamicum*. This highlights a potential requirement for adjustments to the AACell of the iCW773^{R} model and emphasizes the value of computationally derived AACell in refining models.

We also observed that AACell of leucine was significantly higher than others in these three species, while cysteine was relatively less abundant (Fig. 3). Additionally, we observed that this systemic stability was species-specific (Fig. 3). In comparison to *E. coli* MG1655, the PCC for the AACell was 0.955 for *C. glutamicum*, 0.876 for *B. subtilis*, and 0.832 for *S. cerevisiae*. This indicates a generally consistent trend in the AACell across the 20 different AAs in these diverse species. Interestingly, we observed a pattern that the smaller the difference in GC content in the genomes of the species (*E. coli* is 50.79% [24], C. *glutamicum* is 53.8% [25], *B. subtilis* is 43% [26], and *S. cerevisiae* is 38.3% [24]), the closer the distribution of the AACell (Table S6). Extensive research has demonstrated a correlation between genomic GC content and interspecies variations in codon frequencies [27,28] as well as AAs [29,30]. Previous study indicated that the investigation of GC-rich coding sequences would yield proteins with elevated levels of glycine, alanine, arginine, and proline, while AT-rich coding sequences would encode proteins rich in phenylalanine, tyrosine, methionine, isoleucine, asparagine, and lysine [4]. Recent research also indicates that the ratios of AAs in the last universal common ancestor (LUCA) proteins, coded by GC-rich codons, positively correlate with the GC content of various bacterial genomes, while the ratios of AAs coded by AT-rich codons exhibit a negative correlation with the increasing GC content of genomes [31]. In order to investigate the influence of GC content on the AACell, we classified twenty AAs into categories: AAs coded by AT-rich codons (lysine, phenylalanine, and tyrosine), AAs coded by GC-rich codons (proline, glycine, and alanine), and other AAs. Our study, conducted on four species, including three bacteria and one eukaryote, suggests that those with higher GC content tend to exhibit elevated levels of proline, glycine, and alanine per unit mass of total protein, but lower levels of lysine, phenylalanine, and tyrosine (Figs. S4 and S5). This finding confirms the association between genomic GC content and the AACell.

3.4. The AACell in distinct strains of the same species remains stable across different conditions

To determine whether these AA distribution characteristics exist in distinct strains of the same species, we gathered transcriptomes of other *E. coli* strains with transcriptome profiles exceeding 100 (e.g., BW25113, DH1, and W3110), and computed the AACell. Surprisingly, we discovered that these *E. coli* K-12 strains consistently maintained a stable AACell, with a CV of less than 0.06 for all AAs (Table S7). There were no significant differences in the distribution pattern of AAs (Fig. 4), indicating that this may be a common characteristic among *E. coli* strains.

In exploring whether this systemic stability is shaped by the cell or the pathway, a systematic analysis of *E. coli* MG1655 was conducted, making use of its comprehensive transcriptome data. We firstly obtained 55 pathways belonging to the second level (level B) of BRITE from the KEGG [32] database, such as 'Carbohydrate metabolism', 'Energy metabolism', and 'Translation'. After mapping the transcriptomic data, we finally identified 25 pathways that contained genes expressed in the transcriptome. Then, we compared the distribution of AACell in these KEGG pathways and found that there were significant differences in the distribution of AACell in the various pathways (Fig. S6). Thus, we concluded that this stable value was reached only when considering the whole cell level.

Furthermore, we compared the calculated values with two latest sets of experimental AACell [33,34], and we found a very high degree of similarity between the calculated and experimental values (PCC all exceeding 0.91). The standard procedure for experimental determination of AACell involves measuring the total cellular protein content via acid hydrolysis. Subsequently, AA derivatization and quantification are conducted using HPLC, following the protocol outlined by Noble et al. [35] However, a limitation arises during this process as glutamine and asparagine undergo deamination, leading to the formation of glutamate and aspartate [36]. Consequently, the experimental methods face challenges in distinguishing between aspartate and asparagine, as well as glutamate and glutamine. However, AACell calculated through transcriptome allowed for quantitative determination, and the total values closely matched the experimental data (Table S3). For instance, the sum of the AACell of aspartate and asparagine was calculated to be 0.095 g/g total protein, while the experimental value was 0.097, and 0.099, respectively (Table S3).

4. Discussion

Proteins play a crucial role in the functioning of living organisms, and their AA composition varies greatly among different species. This variation is closely related to the evolutionary process of the species and the environment in which the organisms grow [37–40]. For instance, the frequency of oxygen, sulfur, carbon, and hydrogen in proteins of eukaryotes is higher than that of prokaryotes [37]. Moreover, the AAP of different proteins also varies considerably, depending on various factors such as subcellular localization (e.g., membrane proteins are rich in hydrophobic or non-polar AAs) [41,42], protein function (e.g., oxygenated photosynthetic proteins contain more oxygen atoms) [43,44], and GC content (e.g., GC-rich coding sequences produce proteins rich in glycine, alanine, arginine, and proline) [3–5].

In this study, we observed that despite the significant variability of AAP across proteins and the upregulation or downregulation of numerous proteins under different cultivation conditions, AACell remains largely unchanged. This consistency aligns well with experimental measurements, exhibiting a high PCC of 0.94 for *E. coli* and an impressive 0.987 for *S. cerevisiae*. The robust correlation underscores the reliability of inferred AACell and its reflection of evolutionary conservation in AA levels across species. Despite challenges in experimental methods distinguishing certain AAs [36], our transcriptome-based AACell calculations demonstrated quantitative accuracy, closely matching experimental data. This suggests that the genuine AACell within cells can be directly deduced from genomic and transcriptomic information. Additionally, when comparing AACell derived from transcriptomics with high-quality Genome-Scale Metabolic Models (GEMs) of four model organisms, including *E. coli*, *S. cerevisiae*, and *B. subtilis*, a high consistency (PCC approximately 0.86) was observed. However, anomalies in the model for *C. glutamicum* (iCW773^{R}) resulted in a reduced PCC of 0.717, emphasizing the need for adjustments to enhance model accuracy.

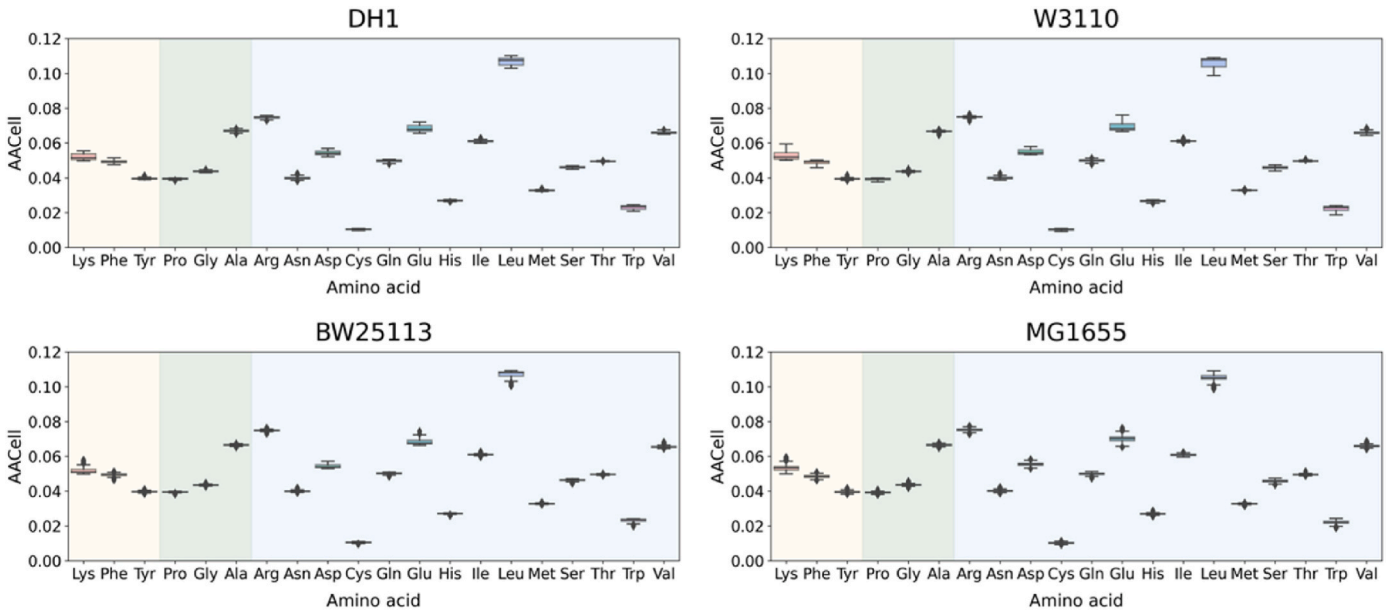

Fig. 4. The distribution of AACell in different *E. coli* K-12 strains. The strains are arranged based on different cultivation conditions from left to right, namely DH1, W310, BW25113, and MG1655. The data in the figure is sourced from the Ecomics database, with DH1 having 103 cultivation conditions, W310 having 280 cultivation conditions, BW25113 having 344 cultivation conditions, and MG1655 having 2307 cultivation conditions. Twenty AAs are classified into three categories: AAs coded by AT-rich codons (off yellow), AAs coded by GC-rich codons (sea mist), and other AAs (azureish white). The meanings of the abbreviations can be found in the abbreviation table.

This study hints at a systemic constraint on the regulation of individual proteins. When a group of proteins rich in a particular AA is upregulated, another set of proteins with a high content of the same AA needs to be downregulated. This systemic constraint allows cells to maintain a stable distribution of intracellular fluxes toward different AAs, even when protein content changes in varying conditions, to meet the growth requirements of different AAs. This may confer an evolutionary advantage, as the nearly constant AACell appears to be a result of the lengthy evolutionary process involving the replacement of AA residues in cellular proteins. There appears to be a trade-off between functional optimization at the individual protein level and optimization at the whole-cell level. Certain enzymes, with low impact on controlling pathway fluxes, experience lower selective evolutionary pressure [45, 46] and thus exhibit flexibility in protein structure and AA usage. Moreover, many AA residues that are distant from the protein's active center can be substituted without affecting protein function. This flexibility in AA composition at the protein level is likely shaped by evolutionary pressure at the whole-cell level. This finding holds significance for metabolic engineering research, where foreign proteins are frequently introduced into a host organism to gain new metabolic capacity, or the foreign protein itself is the objective to produce. The constraint on AA composition at the whole-cell level implies that optimizing a foreign protein involves considerations not only at the codon level but also at the AA composition level. A highly expressed protein with a markedly different AA composition than that of the host cell will necessitate widespread changes in intracellular fluxes and trigger a global cellular response.

5. Conclusions

This study utilized omics data to analyze cellular AACell at a wide range of conditions and discovered that the distribution of AACell remains constant across different cultivation conditions. This systemic stability of AAs is exclusive to each species and shows no difference between distinct strains of the same species. The findings imply that there are whole cell level constraints on the up/down regulation of proteins so that the intracellular fluxes toward different AAs do not need to change significantly despite the great expression variance of individual proteins. These system level constraints are the results of the long evolution process and should be considered in engineering organism for the expression of foreign proteins.

Availability of data and materials

All data used and generated in this study are available either in the Genome or Sequence Read Archive (SRA) databases of the NCBI or in the Supplementary Data associated with this manuscript. In-house codes used in this study are available through a GitHub repository (https://github.com/tibbdc/aaomics).

CRediT authorship contribution statement

Yuanyuan Huang: comprehensively analyzed data from all results, Writing – original draft. **Zhitao Mao:** designed and supervised the project, comprehensively analyzed data from all results, Writing – original draft, Writing – review & editing. **Yue Zhang:** comprehensively analyzed data from all results, Writing – original draft. **Jianxiao Zhao:** contributed to the data collection. **Xiaodi Luan:** assisted in analyzing relevant data. **Ke Wu:** assisted in analyzing relevant data. **Lili Yun:** assisted in analyzing relevant data. **Jing Yu:** assisted in analyzing relevant data. **Zhenkun Shi:** assisted in analyzing relevant data. **Xiaoping Liao:** Writing – review & editing. **Hongwu Ma:** designed and supervised the project, All authors have read and approved the final manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the National Key R&D Program of China

(2022YFC2106000), National Natural Science Foundation of China (32300529, 32201242, 12326611), Tianjin Synthetic Biotechnology Innovation Capacity Improvement Projects (TSBICIP-PTJS-001, TSBICIP-PTJJ-007), Major Program of Haihe Laboratory of Synthetic Biology (22HHSWSS00021), and Strategic Priority Research Program of the Chinese Academy of Sciences (XDC0120201).

Abbreviation

Abbreviation	*Full Name/Explanation*
AA	Amino Acid
AAP_{ij}	The mass ratio of amino acid i in a particular protein j (g/g)
$AACell_{ik}$	The mass ratio of amino acid i in all cellular proteins at a particular condition k (g/g)
MP_{jk}	The mass ratio of protein j in all cellular protein at a particular condition k (g/g)
SD	Standard Deviation
PCC	Pearson Correlation Coefficient
Ala	Alanine
Arg	Arginine
Asn	Asparagine
Asp	Aspartate
Cys	Cysteine
Gln	Glutamine
Glu	Glutamate
Gly	Glycine
His	Histidine
Ile	Isoleucine
Leu	Leucine
Lys	Lysine
Met	Methionine
Phe	Phenylalanine
Pro	Proline
Ser	Serine
Thr	Threonine
Trp	Tryptophan
Tyr	Tyrosine
Val	Valine

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.synbio.2024.03.001.

References

[1] Murray JE, Laurieri N, Delgoda R. Chapter 24 - proteins. In: Badal S, Delgoda R, editors. Pharmacognosy. Boston: Academic Press; 2017. p. 477–94.
[2] Marashi S-A, Behrouzi R, Pezeshk H. Adaptation of proteins to different environments: a comparison of proteome structural properties in Bacillus subtilis and Escherichia coli. J Theor Biol 2007;244(1):127–32.
[3] Moura A, Savageau MA, Alves R. Relative amino acid composition signatures of organisms and environments. PLoS One 2013;8(10):e77319.
[4] Singer GA, Hickey DA. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol 2000;17(11):1581–8.
[5] Brbić M, Warnecke T, Kriško A, Supek F. Global Shifts in genome and proteome composition are very tightly coupled. Genome Biology and Evolution 2015;7(6):1519–32.
[6] Zhou S, Campbell TG, Stone EA, Mackay TFC, Anholt RRH. Phenotypic plasticity of the Drosophila transcriptome. PLoS Genet 2012;8(3):e1002593.
[7] Vinuela A, Snoek LB, Riksen JA, Kammenga JE. Genome-wide gene expression regulation as a function of genotype and age in C. elegans. Genome Res 2010;20(7):929–37.
[8] Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee TI, True HL, Lander ES, Young RA. Remodeling of yeast genome expression in response to environmental changes. Mol Biol Cell 2001;12(2):323–37.
[9] Choi Y-M, Choi D-H, Lee YQ, Koduru L, Lewis NE, Lakshmanan M, Lee D-Y. Mitigating biomass composition uncertainties in flux balance analysis using ensemble representations. Comput Struct Biotechnol J 2023;21:3736–45.
[10] Chen Y, Nielsen J. Yeast has evolved to minimize protein resource cost for synthesizing amino acids. Proc Natl Acad Sci USA 2022;119(4):e2114622119.
[11] UniProt C. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47(D1):D506–15.
[12] Balakrishnan R, Mori M, Segota I, Zhang Z, Aebersold R, Ludwig C, Hwa T. Principles of gene regulation quantitatively connect DNA to RNA and proteins in bacteria. Science 2022;378(6624):eabk2066.
[13] Kim M, Rai N, Zorraquino V, Tagkopoulos I. Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. Nat Commun 2016;7:13090.
[14] Nadal-Ribelles M, Islam S, Wei W, Latorre P, Nguyen M, de Nadal E, Posas F, Steinmetz LM. Sensitive high-throughput single-cell RNA-seq reveals within-clonal transcript correlations in yeast populations. Nature Microbiology 2019;4(4):683–92.
[15] Nicolas P, Mader U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, Bidnenko E, Marchadier E, Hoebeke M, Aymerich S, et al. Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. Science 2012;335(6072):1103–6.
[16] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009;10(3):R25.
[17] Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2013;30(7):923–30.
[18] Pearson K, Galton F, VII. Note on regression and inheritance in the case of two parents. Proc Roy Soc Lond 1895;58(347–352):240–2.
[19] Bogatyreva NS, Finkelstein AV, Galzitskaya OV. Trend of amino acid composition of proteins of different taxa. J Bioinf Comput Biol 2006;4(2):597–608.
[20] Mao Z, Zhao X, Yang X, Zhang P, Du J, Yuan Q, Ma H. ECMpy, a simplified workflow for constructing enzymatic constrained metabolic network model. Biomolecules 2022;12(1):65.
[21] Niu J, Mao Z, Mao Y, Wu K, Shi Z, Yuan Q, Cai J, Ma H. Construction and analysis of an enzyme-constrained metabolic model of Corynebacterium glutamicum. Biomolecules 2022;12(10):1499.
[22] Wu K, Mao Z, Mao Y, Niu J, Cai J, Yuan Q, Yun L, Liao X, Wang Z, Ma H. ecBSU1: a genome-scale enzyme-constrained model of Bacillus subtilis based on the ECMpy workflow. Microorganisms 2023;11(1):178.
[23] Lu H, Li F, Sánchez BJ, Zhu Z, Li G, Domenzain I, Marcišauskas S, Anton PM, Lappa D, Lieven C, et al. A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nat Commun 2019;10(1):3586.
[24] Piovesan A, Pelleri MC, Antonaros F, Strippoli P, Caracausi M, Vitale L. On the length, weight and GC content of the human genome. BMC Res Notes 2019;12(1):106.
[25] Ikeda M, Nakagawa S. The Corynebacterium glutamicum genome: features and impacts on biotechnological processes. Appl Microbiol Biotechnol 2003;62(2):99–109.
[26] Kovács ÁT. Bacillus subtilis. Trends Microbiol 2019;27(8):724–5.
[27] Muto A, Osawa S. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA 1987;84(1):166–9.
[28] Osawa S, Ohama T, Yamao F, Muto A, Jukes TH, Ozeki H, Umesono K. Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets. Proc Natl Acad Sci USA 1988;85(4):1124–8.
[29] Collins DW, Jukes TH. Relationship between G + C in silent sites of codons and amino acid composition of human proteins. J Mol Evol 1993;36(3):201–13.
[30] Foster PG, Jermiin LS, Hickey DA. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J Mol Evol 1997;44(3):282–8.
[31] Du MZ, Zhang C, Wang H, Liu S, Wei W, Guo FB. The GC content as a main factor shaping the amino acid usage during bacterial evolution process. Front Microbiol 2018;9:2948.
[32] Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res 2022;51(D1):D587–92.
[33] Beck AE, Hunt KA, Carlson RP. Measuring cellular biomass composition for computational Biology applications. Processes 2018;6(5):38.
[34] Simensen V, Schulz C, Karlsen E, Bråtelund S, Burgos I, Thorfinnsdottir LB, García-Calvo L, Bruheim P, Almaas E. Experimental determination of Escherichia coli biomass composition for constraint-based metabolic modeling. PLoS One 2022;17(1):e0262450.
[35] Noble JE, Knight AE, Reason AJ, Di Matola A, Bailey MJA. A comparison of protein quantitation assays for biopharmaceutical applications. Mol Biotechnol 2007;37(2):99–111.
[36] Holt L, Milligan B, Roxburgh C. Aspartic acid, asparagine, glutamic acid, and glutamine contents of wool and two derived protein fractions. Aust J Biol Sci 1971;24(3):509–14.
[37] Zhang YJ, Yang CL, Hao YJ, Li Y, Chen B, Wen JF. Macroevolutionary trends of atomic composition and related functional group proportion in eukaryotic and prokaryotic proteins. Gene 2014;534(2):163–8.
[38] Kreil DP, Ouzounis CA. Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res 2001;29(7):1608–15.
[39] Fontanillas E, Galzitskaya OV, Lecompte O, Lobanov MY, Tanguy A, Mary J, Girguis PR, Hourdez S, Jollivet D. Proteome evolution of deep-sea hydrothermal vent alvinellid polychaetes supports the ancestry of thermophily and subsequent adaptation to cold in some lineages. Genome Biology and Evolution 2017;9(2):279–96.
[40] Venev SV, Zeldovich KB. Thermophilic adaptation in prokaryotes is constrained by metabolic costs of proteostasis. Mol Biol Evol 2017;35(1):211–24.

[41] Schwartz R, Ting CS, King J. Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. Genome Res 2001;11(5):703–9.
[42] Zhang YJ, Zhu C, Ding Y, Yan ZW, Li GH, Lan Y, Wen JF, Chen B. Subcellular stoichiogenomics reveal cell evolution and electrostatic interaction mechanisms in cytoskeleton. BMC Genom 2018;19(1):469.
[43] Chen M, Zhang Y. Tracking the molecular evolution of photosynthesis through characterization of atomic contents of the photosynthetic units. Photosynth Res 2008;97(3):255–61.
[44] Baudouin-Cornu P, Surdin-Kerjan Y, Marliere P, Thomas D. Molecular evolution of protein atomic composition. Science 2001;293(5528):297–300.
[45] Sahin A, Weilandt DR, Hatzimanikatis V. Optimal enzyme utilization suggests that concentrations and thermodynamics determine binding mechanisms and enzyme saturations. Nat Commun 2023;14(1):2618.
[46] Coton C, Dillmann C, de Vienne D. Evolution of enzyme levels in metabolic pathways: a theoretical approach. Part 2. J Theor Biol 2023;558:111354.