Proteome and cellular amino acid compositions may be mutually constrained and in a state of narrow convergence

Genshiro Esumi
University of Occupational and Environmental Health, Kitakyushu, Japan

Proteins are composed of 20 amino acids, and their amino acid compositions are known to vary. Today, it is possible to examine all the amino acid compositions of all proteins in a proteome, and several reports have already reported on the differences in the compositions among the proteomes of different organisms. Meanwhile, the distribution of amino acid compositions within the proteome of a single organism has seldom been reported.

In this study, I examined the amino acid composition distribution of proteomic proteins based on publicly available information on human and *Escherichia coli* (*E. coli*) proteomes and found three things. First, each compositional distribution of all the examined amino acids showed a typical bell-shaped distribution. Second, the mean values of the distributions of each proteome were correlated with the measured amino acid compositions of human and E. coli cells, respectively. Third, in the distribution analysis of the distances between the protein compositions and their means, each proteome showed almost identical distributions in shape, and most of the proteins gathered toward the narrowed center near the mean.

The largest source of protein synthesis is intracellular proteins, and it is reasonable to assume that the source composition could constrain the proteome composition. On the other hand, proteome proteins consist of intracellular proteins, and therefore, the proteome composition likely constrains the intracellular composition. Hence, the bell-shaped narrow distributions of the amino acid compositions and the correlations with the actual cellular compositions seem to suggest that the amino acid compositions of proteome proteins are in a state of narrow convergence as a result of mutual constraint with the source amino acids of intracellular proteins.

Keywords: amino acid composition, proteome, convergence

E-mail: esumi@clnc.uoeh-u.ac.jp

※The author has no conflicts of interest relevant to the content of this article.

プロテオームタンパク質のアミノ酸組成分布は細胞のアミノ酸組成と相互に制約し狭い範囲に収束している可能性がある

江角 元史郎
産業医科大学病院 小児外科

タンパク質は20種類のアミノ酸により構成され、そのアミノ酸組成は多様であることが知られている。さらに今日ではプロテオーム全体のタンパク質のアミノ酸組成について検討することが可能となっており、複数の生物プロテオーム間での組成の差についてはすでにいくつもの報告がなされている。しかし、各生物のプロテオーム内でタンパク質のアミノ酸組成がどのように分布しているかということについては全く報告されていなかった。

今回、公開されているヒトと大腸菌のプロテオームの情報をもとに、プロテオームタンパク質のアミノ酸組成分布の検討を行った。その結果、3つのことが明らかとなった。第1に、検討した全てのアミノ酸についてその個々の組成分布は典型的な釣鐘型分布を呈していた。第2に、各プロテオームの分布の平均値はヒト、大腸菌それぞれの細胞のアミノ酸組成の実測値と相関していた。第3に、各生物のタンパク質組成と各平均との距離の分布を検討したところ、各生物においてこれらの分布はほとんど同一形状を呈し、その大半が平均に近い中央の狭い範囲に集中していた。

タンパク質合成の最大の原料は細胞内タンパク質であること、そして細胞内タンパク質はプロテオームのタンパク質によって構成されていることを考慮すると、今回認められたプロテオームアミノ酸組成の釣鐘型分布、組成平均値と細胞実測組成との相関、およびプロテオームアミノ酸組成の中央集中分布は、プロテオームタンパク質のアミノ酸組成が細胞内タンパク質のアミノ酸組成と相互に制約し、比較的狭い範囲に収束している状態にあることを示唆すると考えられた。

キーワード：　アミノ酸組成　プロテオーム　収束

E-mail: esumi@clnc.uoeh-u.ac.jp

背景

タンパク質は20種類のアミノ酸により構成され、そのアミノ酸組成は多様であることが知られている。さらに今日ではプロテオーム全体のタンパク質のアミノ酸組成について検討することが可能となっており、複数の生物プロテオーム間での組成の差についてはすでにいくつもの報告がなされている[1-5]。しかし、各生物のプロテオーム内でタンパク質のアミノ酸組成がどのように分布しているかということについては全く報告されていなかった。

今回の検討では、NCBIのサイトで公開されているプロテオームの情報を用いて各プロテオーム内でのタンパク質のアミノ酸組成分布の解析を行った。

対象と方法１

NCBIのサイトに公開されているヒト[6]、および大腸菌（K-12株）[7]それぞれのリファレンスプロテオームデータに登録されているタンパク質のアミノ酸配列情報をダウンロードして解析の対象とした。

それぞれのプロテオームの各タンパク質について、各アミノ酸残基をカウントし、20種類のアミノ酸の組成の和が1となるように組成を計算した。この際、両方のプロテオームに含まれるアミノ酸であるセレノシステインについては、含まれるタンパク質数、タンパク質内の残基数が両方ともごく少数であるため、カウントの対象外とした。

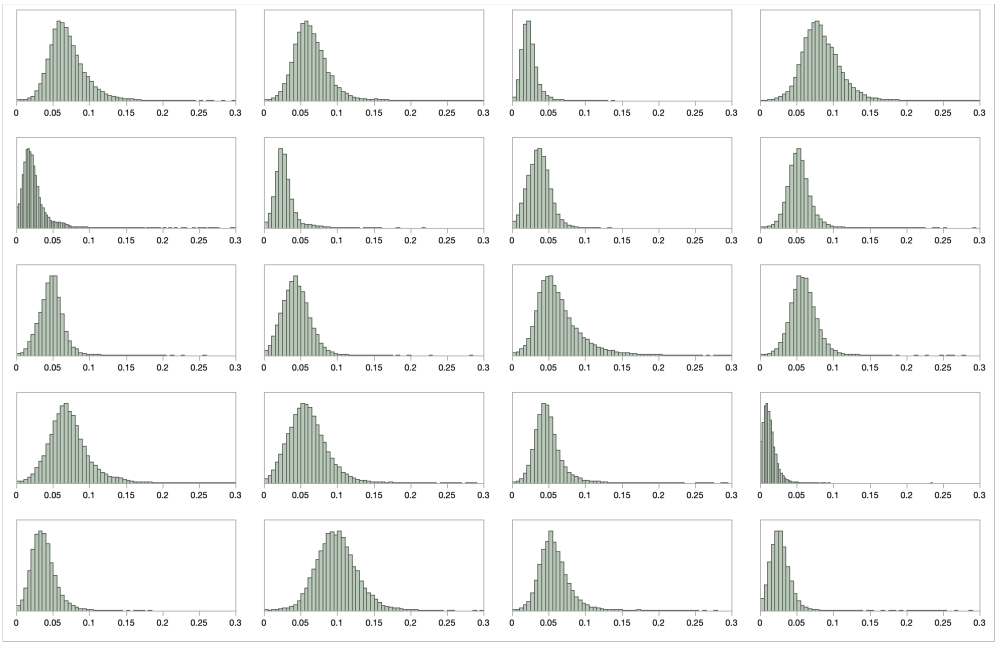
アミノ酸残基数のカウントは Microsoft® Excel for Mac Ver16.61.1 (Microsoft Corporation, USA)を用いて行い、組成の計算と分布の解析は JMP® 16.2.0 (SAS Institute Inc. USA)を用いて行った。

結果１

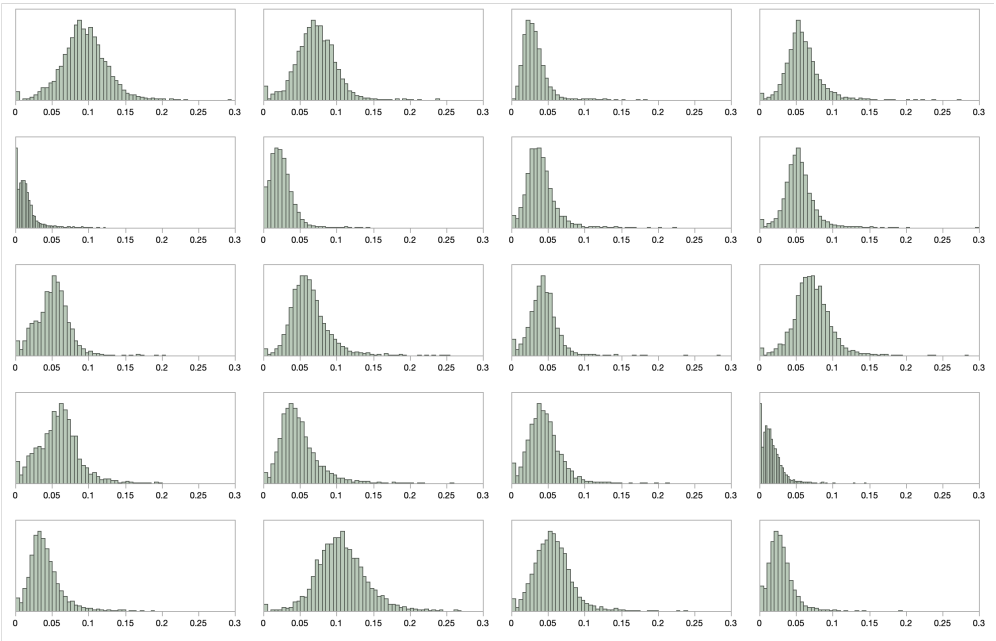
ヒト、および、大腸菌のプロテオームの各アミノ酸組成の分布をFigure 1に示す。今回計算したアミノ酸組成は、全て典型的な釣鐘型分布を呈した。

Figure1

Distributions of amino acid compositions in the Human proteome



Distributions of amino acid compositions in the E. coli proteome



Ala	Gly	Met	Ser
Cys	His	Asn	Thr
Asp	Ile	Pro	Val
Glu	Lys	Gln	Trp
Phe	Leu	Arg	Tyr

※上段：ヒトプロテオーム（n=130184）、および、中段：大腸菌プロテオーム（n=4298）のアミノ酸組成の分布[6,7]。
各アミノ酸の配列は左の通り。

考察１

アミノ酸組成が一様に釣鐘型分布をとることは、プロテオームのタンパク質が合成時に利用可能なアミノ酸の組成に制約されている可能性を示唆すると推測された。仮に、プロテオームの組成が合成時のアミノ酸源に制約されていると考えた場合、上流のアミノ酸源の組成の平均値とプロテオームの組成の平均値は同一値に収束すると推測される[8]。このため、プロテオームにおいて計算された平均値と、過去に報告されているアミノ酸組成の実測値との比較を行うことにした。

アミノ酸合成における上流のアミノ酸源としては、①新規に合成されるもの、②細胞外から取り込まれるもの、③細胞内に遊離アミノ酸として蓄えられているもの、④細胞内のタンパク質が分解され再利用されるもの、が考えられる。アミノ酸利用が生物進化によって最適化されているはずであることを考えれば、おそらく④が最大の供給源である（引用文献なし）。

第1の比較対象として各生物の細胞内の細胞質タンパク質のみのアミノ酸組成を実測した報告を検索したが、比較可能なものは見つからなかった。第2の比較対象として細胞全体のアミノ酸組成を検討した報告を検索したところ、ヒトの培養細胞（腫瘍細胞）、および大腸菌の細胞全体のアミノ酸組成を高速液体クロマトグラフ（HPLC）にて実測した報告のデータ利用可能であったため、このデータとの比較を行う方針とした[9]。

対象と方法２

結果1の組成の平均値を計算した。（Table S1）

過去に報告されている細胞のアミノ酸組成の実測値では、測定時の処理により、アスパラギン（Asn）はアスパラギン酸（Asp）に、グルタミン（Gln）はグルタミン酸（Glu）に合算されていた。また、トリプトファン（Trp）は測定困難であり除外されていた[9]。これと比較を行うため、プロテオームの平均値についてはも、アスパラギン、グルタミンはそれぞれアスパラギン酸、グルタミン酸に合算し、トリプトファンは除外した。この状態で組成の和が 1 となるように組成を再計算し、比較を行った。結果的に、20次元のプロテオームデータは17次元に変換され、17次元の実測データと比較する形となった。

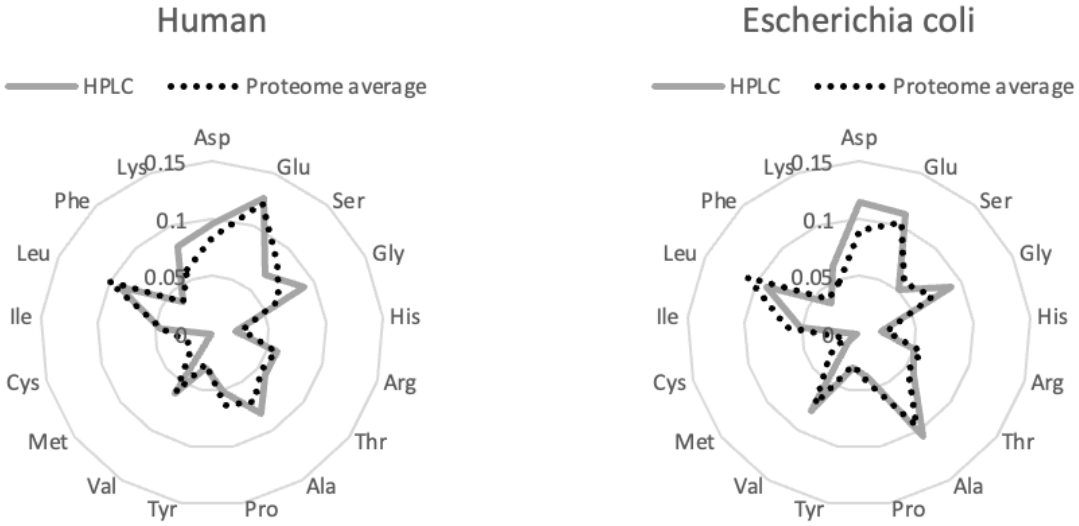
まず、レーダーチャートに上記データを表示し、次にそれぞれの組成の距離（L1 distance）を計算して比較を行った[10]。

結果2-1

プロテオームのアミノ酸組成平均値、および、細胞のアミノ酸組成の実測値[9]を、ヒト、大腸菌それぞれ同一のレーダーチャート上に表示して比較した（Figure 2、Table S2）。主観的ではあるが、それぞれのプロテオームの組成はそれぞれの細胞の実測組成とよく対応しているように見えた。

Figure 2

※プロテオーム組成の平均値（Proteome average）、および細胞の実測値（HPLC）をレーダーチャートにて比較した[6,7,9]。



結果2-2

Figure 2で用いたデータについて、プロテオームの平均値と細胞の実測値間の組成距離（L1 distance）を計算し、表にて比較した（Table 1）[10]。

ヒトプロテオーム組成はヒト細胞組成に近く（L1 distanceが小さく）、大腸菌プロテオーム組成は大腸菌細胞組成に近い（L1 distanceが小さい）結果となった。

Table 1

L1 distance between Proteome average and measured compositions (HPLC)

		Proteome average	
		Human	E. coli
HPLC	Human	0.17535707	0.21438607
	E. coli	0.25734864	0.18111733

考察２

ヒトと大腸菌について、プロテオームと細胞のアミノ酸組成を比較したところ、主観的ではあるが、その組成はグラフ上でよく相関していた。また、組成のL1 distanceを計算したところ、生物種ごとにプロテオームと細胞の実測組成が近く、相互に相関し対応している結果となった。今回の解析ではL1 distanceを用いて解析を行ったが、実際にはCos θ distanceやユークリッド距離を用いても同様の結果が得られた（データ提示なし）。

今回のヒトプロテオーム組成の平均値とヒト細胞実測値との距離である0.175が距離として大きいのか小さいのかを検討する必要があった。このため、ヒトプロテオームの各タンパク質の組成と組成の平均値のL1 distanceを上記17次元で計算し、その分布と比較したところ、L1 distanceの10%tileは0.178となり、上記0.175よりも大きい値となった。従って細胞の実測組成はヒトプロテオーム全体の9割よりも平均値に近いと考えられ、L1 distance 0.175は17次元のアミノ酸組成距離としては十分に小さい（近い）と推測された。（Figure S1）

その他の比較対象として、文部科学省が公開している食品成分表上のアミノ酸組成との比較も試みた[11]。食品成分表のアミノ酸組成リストに掲載されている鶏卵全卵のアミノ酸組成を上記の通り17次元として、今回のヒトプロテオームの平均値とのL1 distanceを計算したところ、その値は0.131であった（データ提示なし）。これは、今回のヒト培養細胞の組成とプロテオームの平均の距離よりも小さい値である。鶏の卵のアミノ酸組成が、ヒトの培養細胞の組成よりもヒトプロテオームの平均値との距離が小さくなった理由としては、使用したのが腫瘍の培養細胞であり平均的な体細胞ではなかった可能性、ヒトとニワトリのプロテオームの組成の平均がそもそも近い可能性、そして、今回の仮説（プロテオームと細胞のアミノ酸組成が近い）が間違っている可能性、などが考えられた。一方で、鶏卵とヒトプロテオーム間のL1 distance 0.131は非常に小さい値であると考えられる。プロテオームと生体アミノ酸組成の間に何らかの相関があることは間違いないと推測された。

プロテオームが上流のアミノ酸源に制約され収束しているのであれば、その収束の度合を量るため、プロテオームの取りうるアミノ酸組成のバリエーションの量（幅）を検討する必要がある。この検討のため、次に、上記のヒト、大腸菌の全タンパク質のアミノ酸組成についてそれぞれのプロテオームの組成の平均値からのL1 distanceを計算し、その分布の比較を行った。

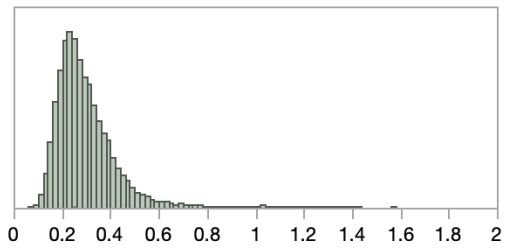
結果３

プロテオームの各タンパク質のアミノ酸組成と、そのプロテオームの組成の平均値とのL1 distanceを計算し、その分布をグラフにして比較した。（Figure 3, Figure 4）

Figure 3

Distribution of L1 distances between amino acid compositions of human proteins and their mean[6].
ヒトタンパク質の組成とプロテオーム組成平均とのアミノ酸距離（L1 distance）の分布

L1 Distance



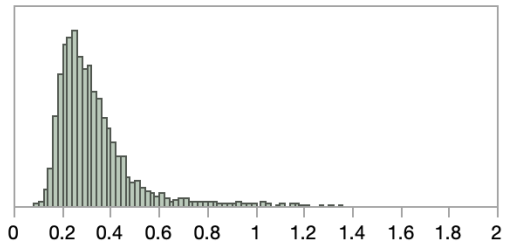
Quantiles		
100.0%	maximum	1.5736391
99.5%		0.738485
97.5%		0.5529608
90.0%		0.4252399
75.0%	quartile	0.3423849
50.0%	median	0.2671467
25.0%	quartile	0.2127641
10.0%		0.1744344
2.5%		0.1392095
0.5%		0.1140867
0.0%	minimum	0.0707237

Summary Statistics	
Mean	0.2889416
Std Dev	0.1112565
Std Err Mean	0.0003084
Upper 95% Mean	0.2895459
Lower 95% Mean	0.2883372
N	130184

Figure 4

Distribution of L1 distances between amino acid compositions of E. coli proteins and their mean[7].
大腸菌タンパク質の組成とプロテオーム組成平均とのアミノ酸距離（L1 distance）の分布

L1 Distance



Quantiles		
100.0%	maximum	1.3403444
99.5%		1.0298631
97.5%		0.7074065
90.0%		0.472578
75.0%	quartile	0.3701714
50.0%	median	0.2870032
25.0%	quartile	0.2266303
10.0%		0.1885097
2.5%		0.1570269
0.5%		0.128201
0.0%	minimum	0.0911439

Summary Statistics	
Mean	0.3188215
Std Dev	0.1414969
Std Err Mean	0.0021583
Upper 95% Mean	0.3230529
Lower 95% Mean	0.3145901
N	4298

考察３

ヒト、および、大腸菌のタンパク質の組成とそれぞれのプロテオームのL1 distanceの分布を検討したところ、ヒト、および、大腸菌の分布はほぼ同一の分布形態を呈した。同様に他生物（マラリア原虫、放線菌）のプロテオームでも同じ検討を行ったところ、それらについてもほぼ同一の分布形態を呈した（Figure S2, Figure S3）。プロテオーム内でのアミノ酸組成のバリエーションの幅・分布形態は、生物種に関わらず一定である可能性が考えられた。

また、ヒト[6]、大腸菌[7]、マラリア原虫[12]、放線菌[13]の４種類のプロテオームについて相互の平均組成間の距離を計算したところ、マラリア原虫と放線菌の間の距離が最も大きく、0.728であった。（Table S3, Table S4）。プロテオームの平均間の距離が0.728という値を取りうるのであれば、プロテオームの半分が平均からのL1 distance 0.2〜0.3以内に分布し、9割がL1 distance0.5以内に分布しているという状況は、個々のプロテオームのアミノ酸組成がかなり狭い範囲に集中していることを意味すると考えられた。

本検討は、プロテオームのタンパク質組成を対象としており、そこには実際の発現量に関する情報は含まれない。にもかかわらず今回の分布解析では、相互制約に伴う収束状態にあることを推測させる状態が観測された。これは、生物進化の過程でランダムな突然変異により様々なアミノ酸組成のタンパク質候補が登場する中で、プロテオームとして採用されるものに対して、アミノ酸組成に対する強い選択圧がかかっているということを示唆すると考えられた。

結論

今回認められたプロテオームアミノ酸組成の釣鐘型分布、組成平均値と細胞実測組成との相関、およびプロテオームアミノ酸組成の中央集中分布は、プロテオームタンパク質のアミノ酸組成が細胞内タンパク質のアミノ酸組成と相互に制約しあい、比較的狭い範囲に収束している状態にあることを示唆すると考えた。

引用文献・引用サイト

1. Tekaia, F., & Yeramian, E. (2006). Evolution of proteomes: Fundamental signatures and global trends in amino acid compositions. *BMC Genomics*, *7*. https://doi.org/10.1186/1471-2164-7-307
2. Brüne, D., Andrade-Navarro, M. A., & Mier, P. (2018). Proteome-wide comparison between the amino acid composition of domains and linkers. *BMC Research Notes*, *11*(1). https://doi.org/10.1186/s13104-018-3221-0
3. Kreil, D. P. (2001). Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Research*, *29*(7), 1608–1615. https://doi.org/10.1093/nar/29.7.1608
4. Moura, A., Savageau, M. A., & Alves, R. (2013). Relative Amino Acid Composition Signatures of Organisms and Environments. *PLoS ONE*, *8*(10). https://doi.org/10.1371/journal.pone.0077319
5. Schmidt, A., Rzanny, M., Schmidt, A., Hagen, M., Schütze, E., & Kothe, E. (2012). GC content-independent amino acid patterns in Bacteria and Archaea. *Journal of Basic Microbiology*, *52*(2), 195–205. https://doi.org/10.1002/jobm.201100067
6. Genome assembly GRCh38.p14 on the NCBI website. https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000001405.40/
7. Genome assembly ASM584v2 on the NCBI website. https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000005845.2/
8. Youtube動画「遺伝暗号縮重の意味」（7:25付近～「おまけ」にて解説）

   https://www.youtube.com/watch?v=GgfigZIc8q8
9. Sorimachi, K. (1999). Evolutionary changes reflected by the cellular amino acid composition. *Amino Acids, 17*(2), 207–226. https://doi.org/10.1007/BF01361883
10. Website. “Comparisons of amino acid compositions between proteins x and y.” http://www0.cs.ucl.ac.uk/staff/L.McGuffin/aminocomp.html
11. 日本食品標準成分表2020年版（八訂）

    https://www.mext.go.jp/a_menu/syokuhinseibun/mext_01110.html
12. Genome assembly GCA_000002765 on the NCBI website. https://www.ncbi.nlm.nih.gov/data-hub/genome/GCA_000002765.3/
13. Genome assembly ASM893130v1 on the NCBI website. https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_008931305.1/

Supplemental data

Table S1

Averages of proteome protein amino acid compositions.

	Human	E. Coli
Ala	0.06906053	0.09344661
Cys	0.02282891	0.01313405
Asp	0.04710301	0.04997623
Glu	0.07107606	0.05770236
Phe	0.03651012	0.03944156
Gly	0.06443533	0.07035089
His	0.02672008	0.02331543
Ile	0.04351055	0.061668
Lys	0.05912225	0.04737596
Leu	0.09879924	0.10642249
Met	0.02272135	0.03051404
Asn	0.03615612	0.03871453
Pro	0.06260936	0.04272531
Gln	0.04831508	0.04384307
Arg	0.05741121	0.05595412
Ser	0.08287089	0.05809553
Thr	0.05271161	0.05333534
Val	0.05895113	0.07100088
Trp	0.0123972	0.01512361
Tyr	0.02668714	0.0278592

Supplemental data
Table S2

Comparison between HPLC data and Proteome composition means.

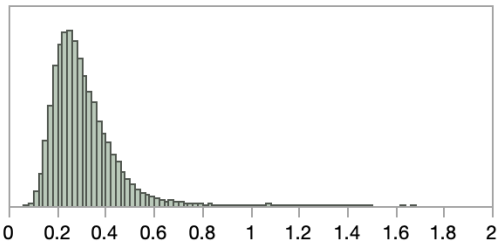
	Human		E. coli	
	HPLC	Proteome	HPLC	Proteome
Asp	0.09592733	0.08430451	0.11461146	0.09005276
Glu	0.12577361	0.12089018	0.11121112	0.10310483
Ser	0.0687762	0.0839114	0.05160516	0.05898768
Gly	0.08973847	0.06524436	0.08960896	0.07143124
His	0.02016371	0.02705557	0.0190019	0.02367347
Arg	0.05859453	0.05813205	0.04910491	0.05681339
Thr	0.05879417	0.05337344	0.05960596	0.05415439
Ala	0.08185267	0.06992763	0.10541054	0.09488163
Pro	0.05190657	0.06339546	0.03840384	0.04338143
Tyr	0.03074466	0.02702222	0.03010301	0.02828703
Val	0.0618886	0.0596913	0.07970797	0.07209121
Met	0.00119784	0.02300663	0.01120112	0.03098263
Cys	0.00069874	0.02311555	0.00150015	0.01333574
Ile	0.04561789	0.04405686	0.05070507	0.06261502
Leu	0.09033739	0.10003974	0.08960896	0.10805679
Phe	0.03673388	0.03696853	0.03620362	0.04004725
Lys	0.08125374	0.05986457	0.06240624	0.0481035

Supplemental data

Figure S1

Distribution of L1 distances of human proteome proteins from the mean composition, calculated in the 17th dimension to fit for HPLC data comparison.

L1 17th



Quantiles		
100.0%	maximum	1.6774681
99.5%		0.7805405
97.5%		0.5824034
90.0%		0.4459349
75.0%	quartile	0.3567037
50.0%	median	0.2768715
25.0%	quartile	0.2186088
10.0%		0.1778481
2.5%		0.140977
0.5%		0.1123265
0.0%	minimum	0.0733264

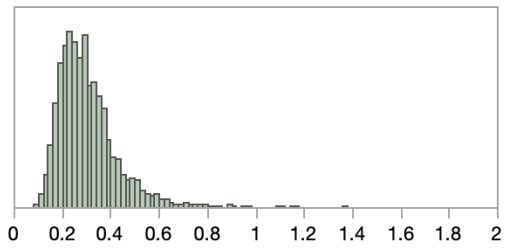
Summary Statistics	
Mean	0.2999318
Std Dev	0.1185777
Std Err Mean	0.0003286
Upper 95% Mean	0.300576
Lower 95% Mean	0.2992877
N	130184

Figure S2

Distribution of L1 distances of P. falciparum proteome proteins from the mean.

Distributions

L1 Distance



Quantiles		
100.0%	maximum	1.3778379
99.5%		0.7523781
97.5%		0.579202
90.0%		0.4416677
75.0%	quartile	0.3535676
50.0%	median	0.2792402
25.0%	quartile	0.2179186
10.0%		0.1782294
2.5%		0.1412435
0.5%		0.112141
0.0%	minimum	0.0888558

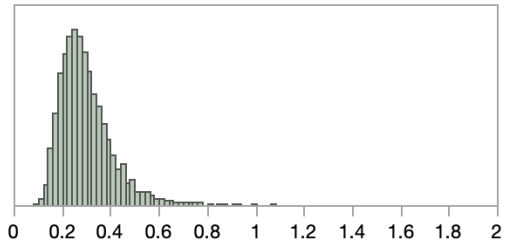
Summary Statistics	
Mean	0.2981281
Std Dev	0.1149172
Std Err Mean	0.0015657
Upper 95% Mean	0.3011975
Lower 95% Mean	0.2950586
N	5387

Figure S3

Distribution of L1 distances of S. coelicolor proteome proteins from the mean.

Distributions

L1 Distance



Quantiles		
100.0%	maximum	1.06211
99.5%		0.6793635
97.5%		0.5441018
90.0%		0.4251238
75.0%	quartile	0.3448697
50.0%	median	0.2750159
25.0%	quartile	0.2212667
10.0%		0.1818068
2.5%		0.1488098
0.5%		0.1248028
0.0%	minimum	0.0959586

Summary Statistics	
Mean	0.2930619
Std Dev	0.1023083
Std Err Mean	0.0011805
Upper 95% Mean	0.295376
Lower 95% Mean	0.2907478
N	7511

Supplemental data

Table S3

Mean amino acid compositions of the proteomes of each organism.

	Human	E. coli	P. falciparum	S. coelicolor
Ala	0.06906053	0.09344661	0.02531304	0.13640701
Cys	0.02282891	0.01313405	0.01830385	0.00836881
Asp	0.04710301	0.04997623	0.05974709	0.06096588
Glu	0.07107606	0.05770236	0.06988169	0.0580959
Phe	0.03651012	0.03944156	0.04698609	0.02658511
Gly	0.06443533	0.07035089	0.03194682	0.09474899
His	0.02672008	0.02331543	0.02274009	0.02343441
Ile	0.04351055	0.061668	0.09327141	0.02925133
Lys	0.05912225	0.04737596	0.11656755	0.02091947
Leu	0.09879924	0.10642249	0.08264144	0.10217055
Met	0.02272135	0.03051404	0.02269231	0.01820463
Asn	0.03615612	0.03871453	0.12115248	0.01647969
Pro	0.06260936	0.04272531	0.02225285	0.06095429
Gln	0.04831508	0.04384307	0.02815444	0.02625286
Arg	0.05741121	0.05595412	0.0299709	0.08440261
Ser	0.08287089	0.05809553	0.06373072	0.04964266
Thr	0.05271161	0.05333534	0.04243422	0.06102103
Val	0.05895113	0.07100088	0.04154341	0.08686675
Trp	0.0123972	0.01512361	0.00588738	0.0152616
Tyr	0.02668714	0.0278592	0.0547779	0.01996643

Table S4

L1 distances between amino acid compositions of each proteome.

	Human	E. coli	P. falciparum	S. coelicolor
Human	0	0.1776181	0.48682816	0.36194749
E. coli	0.1776181	0	0.50090623	0.29845461
P. falciparum	0.48682816	0.50090623	0	0.72350743
S. coelicolor	0.36194749	0.29845461	0.72350743	0