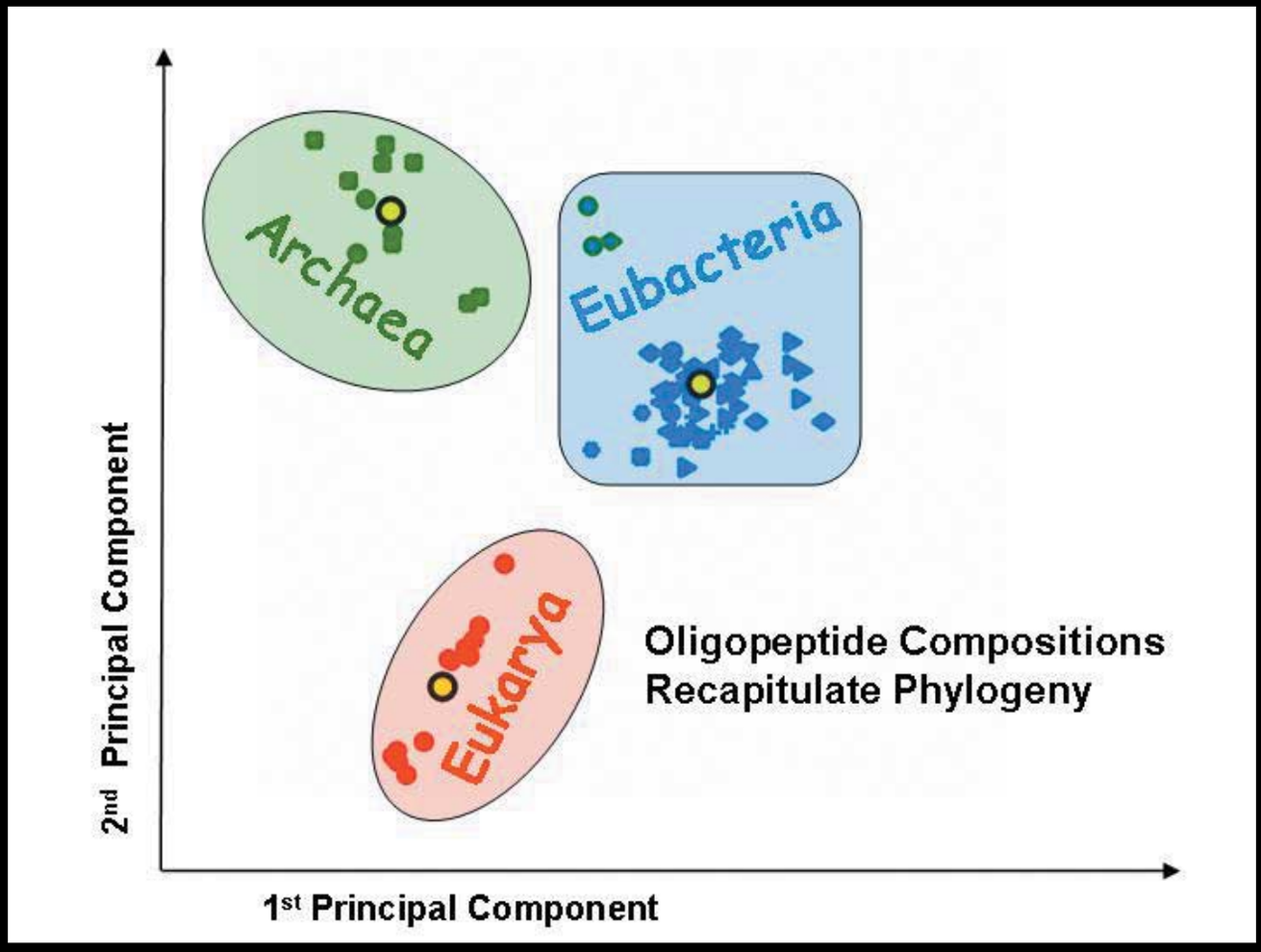
Volume 54, Number 1, January 1, 2004

Articles published online in Wiley InterScience, 3 September 2003–12 December 2003

PROTEINS

Structure, Function, and Bioinformatics



WILEY-LISS
ISSN 0887-3585

PROTEINS: Structure, Function, and Bioinformatics Volume 54, Number 1, January 1, 2004 Pages 1–178 Wiley-Liss

PROTEINS: Structure, Function, and Bioinformatics 54:20–40 (2004)

Proteomic Signatures: Amino Acid and Oligopeptide Compositions Differentiate Among Phyla

Itsik Pe'er,[1¶] Clifford E. Felder,[2] Orna Man,[1,2] Israel Silman,[3§] Joel L. Sussman,[2†*] and Jacques S. Beckmann[1‡*]
[1]*Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel*
[2]*Department of Structural Biology, Weizmann Institute of Science, Rehovot, Israel*
[3]*Department of Neurobiology, Weizmann Institute of Science, Rehovot, Israel*

ABSTRACT **Availability of complete genome sequences allows in-depth comparison of single-residue and oligopeptide compositions of the corresponding proteomes. We have used principal component analysis (PCA) to study the landscape of compositional motifs across more than 70 genera from all three superkingdoms. Unexpectedly, the first two principal components clearly differentiate archaea, eubacteria, and eukaryota from each other. In particular, we contrast compositional patterns typical of the three superkingdoms and characterize differences between species and phyla, as well as among patterns shared by all compositional proteomic signatures. These species-specific patterns may even extend to subsets of the entire proteome, such as proteins pertaining to individual yeast chromosomes. We identify factors that affect compositional signatures, such as living habitat, and detect strong eukaryotic preference for homopeptides and palindromic tripeptides. We further detect oligopeptides that are either universally over- or underabundant across the whole proteomic landscape, as well as oligopeptides whose over- or underabundance is phylum- or species-specific. Finally, we report that species composition signatures preserve evolutionary memory, providing a new method to compare phylogenetic relationships among species that avoids problems of sequence alignment and ortholog detection. Proteins 2004;54:20–40.**

Key words: phylogenetics; principal component analysis; proteome composition

INTRODUCTION

Over 100 genomes have been fully sequenced to date, providing an opportunity for comprehensive comparison and analysis of their organization, similarity, uniqueness, and variability at the sequence level. Comparative analysis of the proteomes derived from these genomes has already proven powerful in gene identification and in prediction of structure, function, and active sites of proteins, as well as in phylogenetic analysis. These analyses are usually based on a per-sequence comparison (e.g., see Gribaldo and Philippe[1]). However, such studies suffer from a major difficulty imposed by the requirement for orthologs of the analyzed proteins from all compared species. Even when orthologs are present, their detection often is prone to error. Furthermore, the resemblance of a specific (or putative) protein may not be representative of species relatedness because of ancestral gene duplication, pseudogenization, or lateral gene transfer (LGT). Finally, the success of such comparative analyses greatly depends on the quality of the sequence alignment, which is hard to control automatically for large data sets.

With available complete-genome data sets, one can pursue complementary per proteome approaches and address general, global properties. In contrast to gene-based approaches, whole-proteome analyses can be performed in the absence of any ortholog knowledge of the encoded products. Indeed, such approaches can be powerful, as illustrated by recent studies.[2–5]

Prominent approaches of this type consider high-level organization, such as chromosomal gene order and compo-

Grant sponsor: Israel Ministry of Science and Technology Grant for Interdisciplinary Studies

Grant sponsor: The Israel Ministry of Science and Technology grant for the Israel Structural Proteomics Center.

Grant sponsor: European Commission Fifth Framework "Quality of Life and Management of Living Resources" program under contract number: QLK3-2000-00650

Grant sponsor: European Commission Fifth Framework "Quality of Life and Management of Living Resources" 'SPINE' Project; Grant number: QLG2-CT-2002-00988

Grant sponsor: Helen & Milton A. Kimmelman Center for Biomolecular Structure and Assembly (Rehovot, Israel)

Grant sponsor: Benoziyo Center for Neurosciences (Rehovot, Israel).

Grant sponsor: Kalman and Ida Wolens Foundation (Rehovot, Israel).

Grant sponsor: Jean and Jula Goldwurm Memorial Foundation (Rehovot, Israel).

Grant sponsor: Divadol Foundation (Rehovot, Israel).

¶Itsik Pe'er is a recipient of the ESHKOL fellowship by the Israeli Ministry of Science and Technology.

§Israel Silman is the Bernstein–Mason Professor of Neurochemistry.

†Joel L. Sussman is the Morton and Gladys Pickman Professor of Structural Biology.

‡Jacques S. Beckmann is the Hermann Mayer Professor of Molecular Genetics.

*Correspondence to: Joel L. Sussman, Department of Structural Biology, Weizmann Institute of Science, Rehovot, 76100 Israel. E-mail: joel.sussman@weizmann.ac.il, or Jacques S. Beckmann, Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, 76100 Israel. E-mail: jacqui.beckmann@weizmann.ac.il

Received 8 April 2003; Accepted 17 June 2003

sition of the proteome.[6] The latter approach, in which similarity between two genomes is reflected by the fraction of shared orthologous genes, usually considers vectors denoting presence or absence of many proteins and compares them across several proteomes.[7] This relieves the dependence on sequence alignment but retains other requirements for analysis, such as correct orthology assignment, LGT avoidance, and accounting for differences in genome size. Higher order analyses compare overall abundance of functional classes of proteins across species.[8] This strategy also alleviates the need for accurately annotated orthologs. The higher order content of different species is indeed found to be different and to distinguish among phyla.[8,9]

In this study, we employ alignment-free, system-level methods for examination of evolutionary differences at the molecular level by comparing amino acid or oligopeptide compositions of known proteomes. The underlying hypothesis is that closely related proteomes tend also to resemble each other at the basic compositional level. Various methods of multivariate analysis have been used to study amino acid residue proteomic composition, leading to the identification of species-specific compositional patterns.[5] These approaches can distinguish composition characteristics of prokaryotic habitats[3,10] and superkingdoms.[4] Recent work[11] suggests analysis of lexical elements that takes into account sequence context, but this work is orthology-dependent and applicable only to specific data sets, such as mitochondrially encoded proteins. All these analyses suggest the use of species-specific compositions as proteomic signatures, analogous to genomic signatures.[12]

We apply and extend the paradigm of comparative analysis of proteomic compositions, and improve analysis of residue composition by focusing on discriminative proteomic features extracted from rationally selected sets of proteomes (training sets). This selection avoids overrepresentation of prokaryotes (due to their overabundance in genome sequence databases), which, to date, has hampered compositional characterization of the eukaryotic superkingdom by an unsupervised (blind) learning method. We further adopt similar methodology for systematic compositional analysis of dipeptides and tripeptides.

Our data demonstrate that not only amino acid composition but also oligopeptide frequencies recapitulate phylogeny (i.e., reflect independent segregation between species and phyla). We identify several distinct factors that shape the landscape of proteomic composition. These results are inferred in an unsupervised manner, namely, by analysis that a priori is blind to taxonomy.

METHODS

We downloaded sequence sets of over 90 complete nuclear proteomes.[i] Inclusion of several other complete or near-complete eukaryotic proteomes, and exclusion of redundant representatives of genera, resulted in 72 different proteomes, listed in Table I. All analyses subsequently described were implemented in MATLAB.

We derived the observed raw counts (the number of occurrences) of all compositional elements (both residues and all overlapping di- and tripeptide fragments) and computed the observed percent frequency of each compositional element along the entire proteome. For single residues, we analyzed the observed frequencies. For di- and tripeptides, we compared the observed frequencies with the frequency that one would expect: Dipeptide expected frequencies were based on single amino acid counts, whereas tripeptide expected frequencies were based on dipeptide counts. To rule out artifactual results due to different classes of sequences having distinct expected frequencies, the evaluation of expected frequencies was performed on a per-sequence basis, then tallied over the entire proteome (see Appendix for a detailed description of the computational procedures). We computed the deviations between each of the observed frequencies given its expectation and a Z score for these deviations (see Appendix).

The frequencies of the 20 amino acids, 400 dipeptides, and 8000 tripeptides define, respectively, 20-dimensional (20-D), 400-D, and 8000-D spaces, in which the frequency vector of each species is a point. To examine each such virtual multidimensional space, and to visualize its salient aspects in 2-D plots, we chose to utilize the powerful tool of principal component analysis (PCA),[13] a multivariate analysis technique that projects the data points onto the coordinate system thus allowing optimal discrimination among them.

To maintain fair focus on variation in each superkingdom, we employed a training set strategy, giving equal weight to all superkingdoms. This allows an unbiased representation of each superkingdom for determination of the best discriminating coordinate system (see Appendix). Once superkingdom separation had been established in an unsupervised manner, we highlighted superkingdom differences by averaging the vectors representing each superkingdom. A similarity tree was computed by hierarchical clustering (see Appendix).

RESULTS

Amino Acid Composition of Proteomes

Amino acid frequencies vary among archaea, eubacteria, and eukaryotes (see Fig. 1). However, intrasuperkingdom variation of these frequencies (see error bars in Fig. 1) obscures the significance of overall compositional differences and warrants application of more rigorous analytic methods.

Species in Residue Frequency Space

The frequencies of the 20 amino acids define a 20-D space, in which the frequency vector of each species is a point. To examine this virtual multidimensional space and to visualize its salient aspects in 2-D plots, we chose to utilize the powerful tool of PCA.[13] We considered the 20-D vectors of amino acid frequencies across different species and visualized this space by PCA (see Methods section), highlighting its various facets by viewpoints determined by different training sets [Fig. 2(a, b, d, and e)].

Figure 2(a) presents the projection of the amino acid frequency vectors of species, for eukaryotes and eubacteria, projected onto the plane defined by the first two principal components as computed for a small subset of species from these superkingdoms (five species per superkingdom, see Table I). These two components capture

TABLE I. Species Used for the Data Set

Superkingdom	Classification	Species	acronym	size[1]
Eukaryota	●	*Mus musculus*	Mmu	20.1(8.1)
		Arabidopsis thaliana	Ath	26.0(11.2)
		Schizosaccharomyces pombe[2]	Spo	5.1(2.4)
		Saccharomyces cerevisiae	Sce	6.2(2.9)
		Caenorhabditis elegans	Cel	20.0(8.8)
		Drosophila melanogaster[2]	Dme	15.3(7.9)
		Homo sapiens[2]	Hsa	25.9(12.3)
		Rattus norvegicus[2]	Rno	7.5(3.2)
		Anopheles gambiae[2]	Aga	15.2(6.6)
		Ciona intestinalis	Cin	15.9(6.1)
		Oryza sativa	Osa	15.1(5.8)
		Takifugu rubtipes	Tru	82.0(23.8)
Archaea	Euryarchaeota ■	*Methanocaldococcus jannaschii*[3]	Mja	1.8(0.5)
		Methanosarcina mazei	Mma	3.4(1.0)
		Archaeoglobus fulgidus[3]	Afu	2.4(0.7)
		Pyrococcus furiosus	Pfu	2.1(0.6)
		Halobacterium sp. NRC-1[3]	Hsp	2.4(0.7)
		Thermoplasma acidophilum	Tac	1.5(0.5)
		Methanopyrus kandleri	Mka	1.7(0.5)
		Methanothermobacter thermautotrophicus[3]	Mth	1.9(0.5)
	Crenarchaeota ●	*Aeropyrum pernix*[3]	Ape	2.7(0.6)
		Sulfolobus solfataricus	Sso	2.9(0.8)
		Pyrobaculum aerophilum	Pae	2.6(0.7)
Eubacteria	Firmicutes ◆	*Streptococcus pyogenes*	Spy	1.7(0.5)
		Lactococcus lactis subsp. lactis	Lla	2.2(0.7)
		Bacillus subtilis	Bsu	4.1(1.2)
		Clostridium acetobutylicum	Cac	3.8(1.2)
		Listeria monocytogenes	Lmo	2.8(0.9)
		Mycoplasma genitalium	Mge	0.5(0.2)
		Thermoanaerobacter tengcongensis ◆	Tte	2.5(0.8)
		Ureaplasma parvum	Upa	0.6(0.2)
		Staphylococcus aureus subsp. aureus Mu50	Sau	2.7(0.8)
		Oceanobacillus iheyensis	Oih	3.5(1.0)
	Alpha proteobacteria ▲	*Mesorhizobium loti*	Mlo	7.3(2.2)
		Sinorhizobium meliloti[4]	Sme	6.1(1.9)
		Rickettsia conorii	Rco	1.4(0.3)
		Brucella melitensis[4]	Bme	3.2(0.9)
		Caulobacter crescentus	Ccr	3.7(1.2)
		Agrobacterium tumefaciens str. C58 (U. Washington)	Atu	5.4(1.7)
	Beta proteobacteria ▶	*Ralstonia solanacearum*	Rso	5.0(1.7)
		Neisseria meningitidis serogroup B	Nme	2.0(0.6)
	Gamma proteobacteria ▼	*Pseudomonas aeruginosa*	Pae	5.6(1.9)
		Xanthomonas campestris pv. campestris	Xca	4.2(1.4)
		Salmonella typhi	Sty	4.7(1.4)
		Yersinia pestis	Ype	3.9(1.2)
		Vibrio cholerae	Vch	3.8(1.2)
		Haemophilus influenzae	Hin	1.7(0.5)
		Pasteurella multocida	Pmu	2.0(0.7)
		Xylella fastidiosa	Xfa	2.8(0.7)
		Escherichia coli K12	Eco	4.4(1.4)
		Buchnera aphidicola (Acyrthosiphon pisum)[4]	Bap	0.6(0.2)
		Wigglesworthia brevipalpis	Wbr	0.6(0.2)

almost all of the variation among species, together accounting for almost 90% of the observed variance of the 20-D data. The separation between eukaryotes and eubacteria is evident, although the method of analysis a priori was blind to the taxonomy of the proteomes analyzed. Furthermore, the principal plane we ascertained using the training set also clearly separates the remaining species (i.e., those not used in the training set). It should be observed,

TABLE I. (Continued)

Superkingdom	Classification	Species	acronym	size[1]
		Shewanella oneidensis	Son	4.8(1.4)
	Proteobacteria-delta/epsilon subdivisions ◀	*Campylobacter jejuni*	Cje	1.6(0.5)
		Helicobacter pylori 26695	Hpy	1.6(0.5)
	Cyanobacteria (blue-green algae) +	*Synechocystis sp. PCC 6803*	Ssp	3.1(1.0)
		Nostoc sp. PCC 7120	Nsp	6.1(2.0)
		Synechococcus elongatus[4]	Sel	2.5(0.8)
	Actinobacteria ★	*Mycobacterium leprae*	Mle	1.6(0.5)
		Streptomyces coelicolor	Sco	7.8(2.5)
		Corynebacterium glutamicum ATCC 13032	Cgl	3.1(1)
		Bifidobacterium longum	Blo	1.7(0.6)
	Spirochaetes ■	*Borrelia burgdorferi*[d]	Bbu	1.3(0.4)
		Treponema pallidum	Tpa	1.0(0.3)
		Leptospira interrogans	Lin	4.7(1.2)
	Chlamydiae ×	*Chlamydia trachomatis*	Ctr	0.9(0.3)
		Chlamydophila pneumoniae AR39	Cpn	1.1(0.4)
	●	*Deinococcus radiodurans*	Dra	3.1(0.9)
		Chlorobium tepidum	Cte	2.2(0.6)
		Thermotoga maritima	Tma	1.9(0.6)
		Aquifex aeolicus	Aae	1.6(0.5)
		Fusobacterium nucleatum subsp. nucleatum	Fnu	2.1(0.6)
Subsets	*S. cerevisiae* subsets	Chromosomes 1-16		0.1-0.9 (0.048-0.39)
	Human subsets	100 proteins (100 subsets)		0.1(0.036-0.063)
		1000 proteins (100 subsets)		1(0.43-0.52)

[a]In thousands of proteins (millions of amino acids).
[b]Used in training set for Figures 2(a–d), 3(a), and 5(a).
[c]Used in training set for Figures 2(b–d), 3(a), and 5(a).
[d]Used in training set for Figures 2(a), 2(d), 3(a), and 5(a).
Note: The data set included complete or partial proteomes of 12 eukaryotes, 11 archaea, and 49 eubacteria. The archaea and eubacteria were further subdivided taxonomically.[iii] The Classification column shows the symbol given to the species in the various figures, as well as its subdivision, when necessary. Special symbols were used for thermophilic eubacteria. The Acronym column indicates the three-letter abbreviation for the species employed. All complete proteomes were downloaded from EBI.[i] Eukaryotic proteomes whose genomes are only partially sequenced, or still uncurated by EBI, were also downloaded: *Rattus norvegicus,*[iv] *Anopheles gambiae,*[v] *Takifugu rubripes,*[vi] *Oryza sativa,*[vii] and *Ciona intestinalis.*[viii]

however, that the first principal component (x axis) is dominated by the variance within eubacteria, whereas the second (y axis) separates them from the eukaryotes. This is concordant with the observation relying on alignment-based comparative methods that eukaryotes are also less divergent. It should be noted that subphyla of eubacteria also tend to segregate: Chlamidiae, actinobacteria, cyanobacteria, spirochaetes, and α-, β-, and γ-proteobacteria each reveal an observable cluster in composition space.

Figure 2(b) presents a corresponding analysis for archaea versus eukaryotes (Table I lists the species used for training). Here, too, the superkingdom separation is clear-cut and predictive with respect to species not used for training. In this case, both principal components contribute to the separation and account for 52% and 31%, respectively, of the variance.

The clustering of superkingdoms is evident irrespective of the training set chosen (see Supplementary Material[ix]), indicating that it is an inherent feature of amino acid composition. Interestingly, the principal plane is discriminative also for eubacteria not used for training [Fig. 2(c)].

The three superkingdoms are also separated upon selection of a training set consisting of five arbitrarily selected species per superkingdom [see Table I, Fig. 2(d)]. Whereas

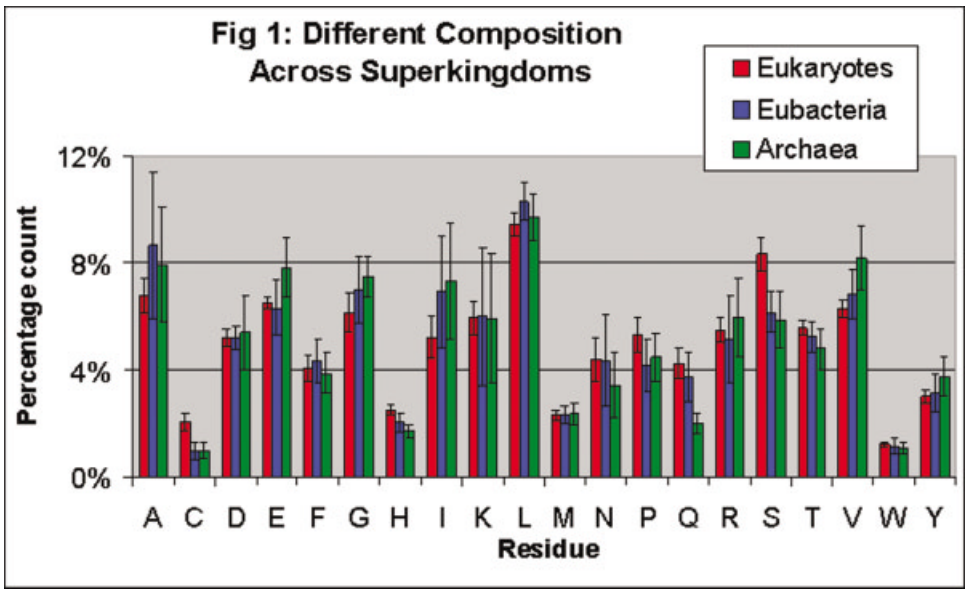


Fig. 1. Average amino acid compositions across superkingdoms. Amino acid percentage counts were recorded for each species in Table I. Bars represent average amino acid frequency for all our eukaryotic (red), eubacterial (blue), and archaeal (green) proteomic data sets. Error bars represent the empirical standard deviation of the recorded percentage counts in each of the three superkingdom-specific data sets.

the first principal component accounts primarily for intra-superkingdom variability (specifically, variation among eubacteria), the second layers eubacteria midway between eukaryotes and archaea. A different perspective is obtained if the average compositions of the proteomes in each superkingdom are used to define the projection plane [Fig. 2(e)]. This viewpoint not only presents three-way superkingdom separation more clearly but also highlights its two-dimensionality, demonstrating that no single axis is best to tell these three superkingdoms apart. It should be noted that the three eubacterial species (*Thermoanaerobacter tengcongensis*, *Thermtoga maritima*, and *Aquifex aeolicus*) observed to cluster with archaea are thermophilic, suggesting that the compositional differences between the two prokaryotic superkingdoms may be related to habitats (e.g., growth temperature[2,3]). We observe that the eukaryote–eubacteria ("organism complexity") axis is almost orthogonal to the eubacteria–archaea ("habitat") axis. Together, these axes make up the eukaryote–archaea ("combined") axis.

We show that our analysis is robust even upon substantial reduction of sample size. Thus, the distinctness of eukaryotic composition is significant enough to be observed when examining a set of proteins much smaller than a complete proteome: The compositional vectors of random small subsets of the human genome (even down to 100 proteins per set) cluster well around the compositional vector for the entire human proteome. The same holds true for the compositional vectors of all the individual yeast chromosomes [Fig. 2(d′)], which individually encode from 105 to 828 (chromosomes 1 and 4, respectively) different proteins.

Oligopeptide Composition

Although single-residue composition is very informative in itself, additional and complementary information may be hidden in more complex lexical structures. This is evident, for example, in the significant frequency differences among almost every dipeptide *XY*, and the reverse-order dipeptide, *YX* (see Supplementary Material[ix]). In this article, we studied also di- and tripeptide compositions. This required careful analysis and normalization (see Methods section) to exclude compositional bias already observed at the single–amino acid level from obscuring potential additional differences in the composition that are present at the dipeptide level. This "expected" composition was very sensitive to whether it was evaluated on a per-proteome or a per-protein basis (see Methods section). Similar normalization for exclusion of bias due to dipeptide compositional differences was performed during tripeptide compositional analysis.

The deviation observed versus expected di- and tripeptide counts greatly exceeded its theoretical variance for many oligopeptides (see Supplementary Material[ix]). Furthermore, even when normalized by their empirical variance (computed from all oligopeptides in the respective species), these deviations still highlight extremely outlying oligopeptides, the identity of which differs significantly from species to species. Table II presents these outlying oligopeptides in proteomes representative of the three superkingdoms (human, *Escherichia coli* and *Pyrococcus furiosus*). It can be seen that many homodipeptides are overabundant in the human proteome. Furthermore, several homotripeptides are overabundant, even when taking dipeptide distribution into account. Such features are shared by all eukaryotes (see Supplementary Material[ix]). This is in agreement with the overabundance of long-residue runs in eukaryotic genomes.[5] It is interesting to note that even the prokaryotes have a skewed distribution

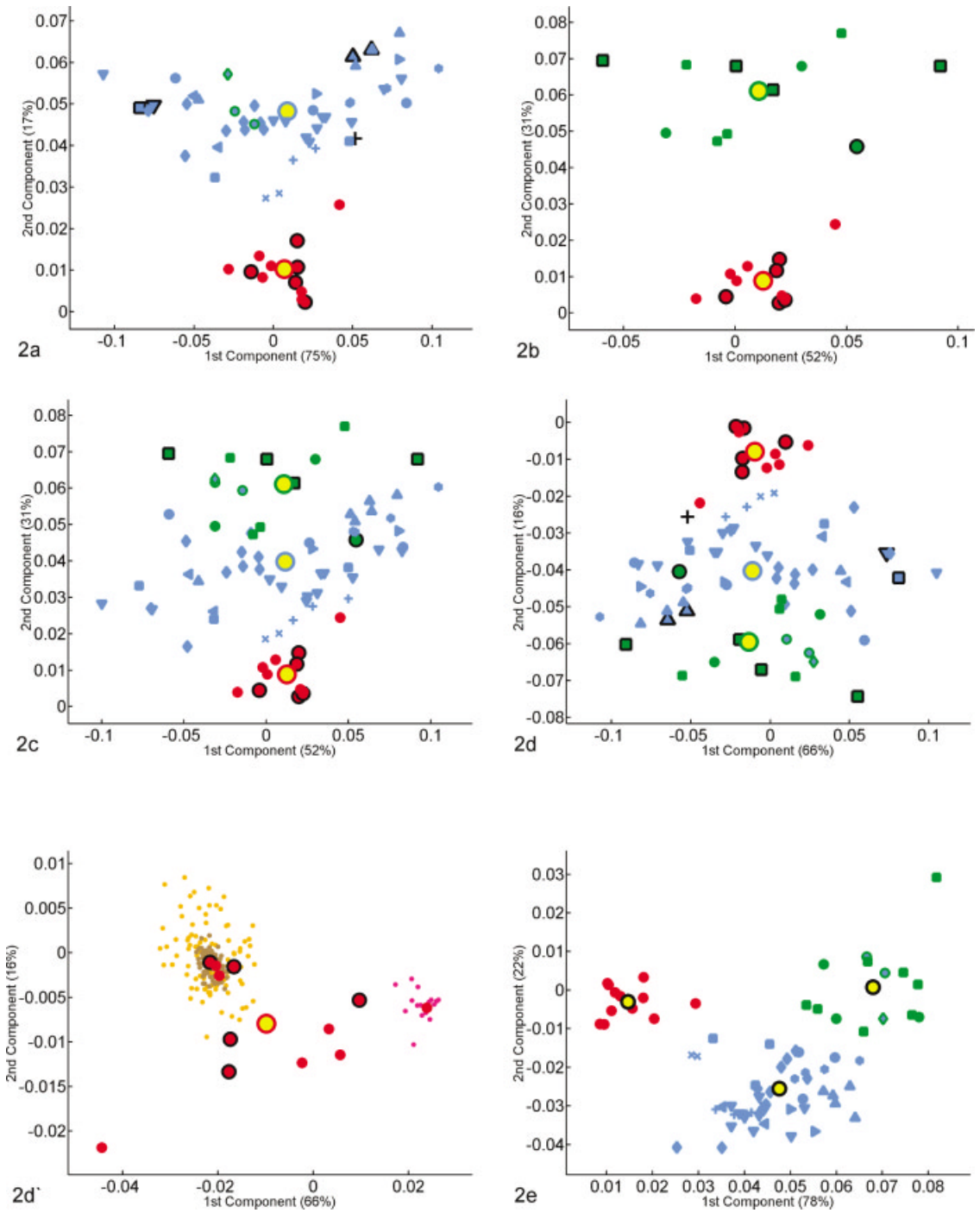


Fig. 2. Projection of the amino acid frequency vectors of species onto the factorial plane formed by the first two principal components. The *x* and *y* axes represent, respectively, the first and second principal components obtained by PCA of the 20-D amino acid composition space of a species training set. The various factorial planes (**a–e**) are obtained by using different training sets (**a–d**: unsupervised classification; **e**: supervised), but species not used for training are also projected onto the same factorial plane. The training set is highlighted by an increased size of symbols and a black border around them. The percentage contribution of a given component to the overall variability within the training set is indicated on its axis. The symbols representing the various species are as in Table I. Yellow circles indicate the average amino acid compositions of the three superkingdoms. The borders around the yellow circles indicate which superkingdom they represent: red for eukaryotes, blue for eubacteria, and green for archaea. (**a**) All eukaryotes and all eubacteria in the data set are projected onto the factorial plane created by using five eukaryotes (*Dme*, *Aga*, *Hsa*, *Rno*, and *Spo*) and five eubacteria (*Bbu*, *Bme*, *Bap*, *Sel*, and *Sme*) as a training set. (**b**) All eukaryotes and all archaea in the data set are projected onto the factorial plane created by using five eukaryotes (see **a**) and five archaea (*Afu*, *Ape*, *Mth*, *Mja*, and *Hsp*) as a training set. (**c**) Projection of all species in the dataset onto the factorial plane of **b**. (**d**) All species in the data set are projected onto the factorial plane created by using five eukaryotes (see **a**), five eubacteria (see **a**), and five archaea (see **b**) as a training set. (**d′**) A magnification of the factorial plane is shown around the eukaryotes. Pink circles represent the chromosomes of *S. cerevisiae*. Brown and orange circles represent randomly picked sets of 100 and 1000 human proteins, respectively. (**e**) All species in the data set are projected onto the factorial plane created by PCA analysis, with the average protein compositions of the three superkingdoms as a training set (supervised set).

TABLE II. Over-/Underabundant Oligopeptides in Representative Proteomes

Species	Top or bottom	Peptide length	Peptide	Z-score		p-value	
				Theoretical	Empirical		
P. furiosus	Most underabundant	3	IEI	-7.21	-5.06	0.003	0.0002
			LEL	-6.67	-4.67	0.02	0.0012
			IAI	-6.59	-4.62	0.03	0.002
			GKE	-6.54	-4.58	0.04	
			EKE	-6.38	-4.48	0.06	0.003
			IEL	-6.38	-4.47	0.06	
			LEI	-6.26	-4.39	0.09	
			EEE	-5.90	-4.14	0.28	0.0007
			EED	-5.71	-4.01	0.50	0.05
			LKL	-5.70	-4.00	0.51	0.03
	Most overabundant		LEE	6.34	4.44	0.07	0.007
			IKE	6.52	4.56	0.04	
			KAL	6.60	4.62	0.03	
			EEA	6.62	4.64	0.03	0.003
			LKE	7.61	5.33	0.0008	
			EKL	7.99	5.60	0.0002	
			LKK	8.08	5.66	0.0001	1.E-05
			KEL	8.21	5.75	7.E-05	
			EEL	8.72	6.11	8.E-06	8.E-07
			GKT	9.26	6.49	7.E-07	
		2	NP	19.52	4.10	0.02	
E. coli	Most under-abundant	3	LRL	-8.09	-5.13	0.002	0.0001
			LAL	-7.98	-5.06	0.003	0.0002
			DPG	-7.71	-4.88	0.008	
			LKL	-7.41	-4.69	0.02	0.0011
			LEL	-7.21	-4.57	0.04	0.002
			LQL	-6.79	-4.30	0.13	0.007
	Most overabundant		LEN	6.66	4.22	0.20	
			LPP	6.74	4.27	0.16	0.02
			DEP	6.86	4.34	0.11	
			LAE	7.01	4.44	0.07	
			LRE	7.6	4.86	0.009	
			EKL	7.70	4.88	0.009	
			IVG	7.74	4.90	0.008	
			QAL	7.96	5.04	0.004	
			GKT	8.13	5.15	0.002	
			MKK	8.84	5.60	0.0002	2.E-05
		2	QQ	27.48	3.64	0.11	0.006
Homo sapiens	Most under-abundant	3	DPP	-21.72	-5.27	0.0011	0.0001
			CGE	-21.31	-5.17	0.002	
			GEC	-21.06	-5.11	0.003	
			LQL	-17.65	-4.28	0.15	0.008
			LKL	-16.64	-4.03	0.44	0.02
			LEL	-16.41	-3.98	0.55	0.03
	Most overabundant		AAA	29.60	7.19	5.E-09	1.E-11
			QQQ	30.14	7.32	2.E-09	5.E-12
			PLP	30.74	7.47	7.E-10	3.E-11
			PAP	32.80	7.97	1.E-11	7.E-13
			EEE	34.75	8.44	3.E-13	6.E-16
			TGE	35.52	8.63	5.E-14	
			PPP	36.12	8.77	1.E-14	3.E-17
			HTG	43.41	10.54	4.E-22	
			YKC	44.35	10.77	4.E-23	
			CGK	47.70	11.59	4.E-27	
		2	PP	72.35	3.76	0.07	0.003
			SS	77.27	4.02	0.02	0.0012
			AA	79.81	4.15	0.013	0.0007
			EE	109.27	5.68	5.E-06	3.E-07

Note: Dipeptide and tripeptide frequency counts were recorded for *H. sapiens, E. coli,* and *P. furiosus,* and theoretical Z scores were computed (see Methods section) to reflect deviation from the counts expected given observed lower order (single-residue and dipeptide, respectively) compositions. Empirical Z scores were computed based on the observed variance among peptides of the respective proteome. Using the empirical Z scores, normal distribution *P* values for over-/underabundance were computed, correcting for two-sidedness and multiple tests performed (400 dipeptides or 8000 tripeptides). For homopeptides (red), palindromes (green), and semihomopeptides (XXY/YXX tripeptides, shown in black), *p* values corrected for less multiple tests (20, 400, or 800, respectively) are also quoted, whereas only a single *p* value is listed for heteropeptides (blue). Only significant entries are listed and, at most, 10 tripeptides in each category.

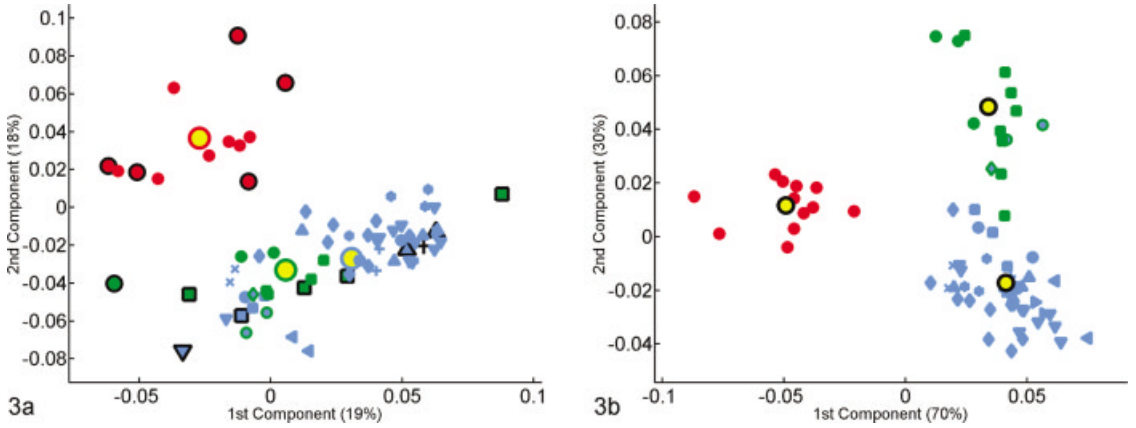


Fig. 3. Projection of the dipeptide overabundance vectors of species onto the factorial plane formed by the first two principal components. The overabundance of each dipeptide in each proteome was computed on a per-sequence basis, taking into account observed single-residue frequencies (see Methods section). The *x* and *y* axes represent, respectively, the first and second principal components obtained by principal component analysis of the 400-D space of dipeptide composition. All species in the data set are projected onto the factorial plane that is (**a**) defined by using the training set of Figure 2(d); (**b**) defined by the compositional averages of the three superkingdoms, as in Figure 2(e). Plotted symbols are as in Figure 2.

of lexical elements with overabundance of homodipeptides and -tripeptides, though very much less than for the eukaryotes (data not shown).

Species in Oligopeptide Frequency Space

The vectors of frequency deviation from expectation for di- and tripeptides are entities in 400- and 8000-D spaces, respectively. These higher dimensional spaces provided an even stronger motivation for the use of PCA than the amino acid compositions.

Figure 3 presents the dipeptide compositional vectors of species onto the factorial plane formed by the first two principal components. Figure 3(a) presents the dipeptide space from a viewpoint determined by the same training set utilized to generate Figure 2(d) (see Table I). There is a striking differentiation between eukaryotes and prokaryotes. Within prokaryotes, archaea are clearly clustered, but not as well separated from eubacteria as in composition analysis of single amino acids (see Fig. 2). Eukaryotes also display significant variation. Furthermore, compared to single-residue composition, less of the variation is visible in the first two principal components of dipeptide composition (19% and 18%, respectively). A more informative viewpoint, based on average dipeptide composition of superkingdoms [Figure 3(b)], allows their separation, with the exception of the three thermophilic eubacteria present in our data set. We note that even though frequencies of reversed-order dipeptides (*XY* vs. *YX*) may differ significantly within each pair, in the 210-D space in which reversed-order dipeptides are lumped together, phyla separation is equally visible (data not shown).

Comparative analysis for tripeptides showed similar superkingdom clustering based on superkingdom averages (Fig. 4), although tripeptide composition is subject to more species-specific variation, and when determining principal components by unsupervised training sets, less of the variation is accounted for (data not shown). For the tripeptides, the three thermophilic eubacteria are visibly separated from both archaea and mesophilic eubacteria.

Residue Contribution to Principal Components

Because amino acid, di-, and tripeptide compositions all separate the three kingdoms so clearly with the use of PCA methodology, we went on to examine the underlying compositional differences responsible for these separations. To this end, we considered the coordinate system that defines PCA plots, such as the one presented in Figure 2(d). The weight or impact of each residue's frequency on the first and second principal components is best visualized by plotting the residues by using these weights as coordinates [Fig. 5(a)]. The first principal component in this view (*x* axis) accounts for 58% of the variability. This component seems to be related to the GC content of the codons for amino acids:

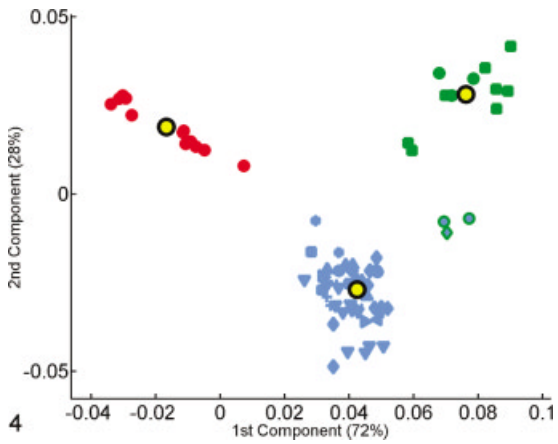


Fig. 4. Projection of the tripeptide frequency vectors of species onto the factorial plane formed by the first two principal components. The overabundance of each tripeptide in each proteome was computed on a per-sequence basis, taking into account observed dipeptide frequencies (see Methods section). The *x* and *y* axes represent, respectively, the first and second principal components obtained by PCA of the 8000-D space of tripeptide composition. All species in the data set are projected onto the factorial plane defined by the compositional averages of the three superkingdoms, as in Figure 2(e). Symbols are as in Figure 2.

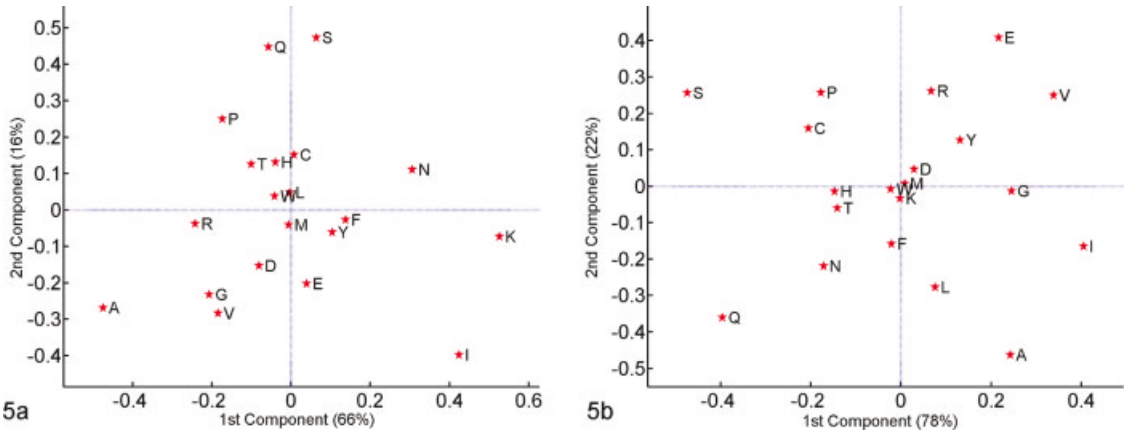


Fig. 5. Amino acid contributions to principal components of species compositions. The contributions of each amino acid to the first (x) and second (y) component axes of the coordinate system defined by PCA are plotted. (**a**) The first and second principal components for the unsupervised training set presented in Figure 2(d); (**b**) the first and second principal components for the supervised training set presented in Figure 2(e). The farther away the residue from the origin of axes, the more it contributes to the difference between the observed compositions. The direction (with respect to the origin of axes) in which the residue is plotted corresponds to the direction in the corresponding species plot in which it distinguishes compositions.

Amino acids that are coded by GC-rich codons, such as G, A, R, and P, are concentrated at low x values, whereas those that are coded for by AT-rich codons, such as F, I, Y, K, and N, display high x values. We cross-validated this against the open reading frame (ORF) GC content of yeast chromosomes that shows 90% correlation to their x-axis position in Figure 2(d′). This axis is mainly responsible for variation within the individual superkingdoms—particularly within eubacteria [see Fig. 2(d)]. In contrast, the second principal component (y axis) separates eukaryotes, eubacteria, and archaea. This separation is better analyzed in Figure 5(b), which presents the coordinate axes for the principal components that separate superkingdom average compositions [Figure 2(e)]. The dissection of intersuperkingdom space into "habitat" and "organism-complexity" [Figure 2(e)] allows characterization of these axes in terms of amino acids. The most prominent "habitat"-dependent difference is Q versus its charged counterpart E, which has been hypothesized to play a role in stabilizing oligomeric protein structures in thermophiles.[14] On the "organism-complexity" axis, we can identify overabundance of aliphatic residues in prokaryotes. The "combined" axis highlights overabundance of uncharged polar residues in eukaryotes, especially compared to archaea.

Oligopeptide Contribution to Principal Components

Oligopeptide compositions can be analyzed similarly. Figure 6 shows the dipeptide contributions to the PCA presented in Figure 3(b). The striking phenomenon is the strong contribution of homodipeptides to the negative direction of the first principal component (i.e., overabundance of homodipeptides among eukaryotes). It should be noted that this phenomenon is observed when single-residue effects are eliminated. This means that the eukaryotic overabundance of, for example, *AA* is apparent only with respect to its expected frequency, which is much lower in eukaryotes (where *A* is more rare). A contrasted example is the homodipeptide *SS*, whose excess in eukaryotes is significant even when the eukaryotic overabundance of *S* is taken into account.

It is interesting to ask whether the only difference between dipeptide compositions of eukaryotes and prokaryotes is due to bias toward homodipeptides. To answer this question, we performed PCA, analogous to that presented in Figures 3(b) and 6, but taking into account the overall homodipeptide bias (see Methods section). The unaccounted variation can be separated into its two components: variation in homo- versus heterodipeptide compositions. As can be seen, both homodipeptide [Fig. 7(a, b)] and heterodipeptide [Fig. 7(c, d)] frequency information is sufficient for separating eukaryotes from prokaryotes. We further verified that eukaryotes' homodipeptide preference is not an artifact due to selected proteins with long homopeptide runs, because their removal from the analysis did not qualitatively affect the results (data not shown).

Tripeptide analysis (Fig. 8) reveals that homotripeptides are strongly characteristic of eukaryotic proteomes, and is consistent with the homodipeptide overabundance mentioned earlier. Furthermore, palindromic tripeptides are also characteristic of eukaryotes. Again, even when species-specific homopeptide preference is accounted for, the relative frequency information for either homotripeptides [Fig. 9(a, b)] or heterotripeptides [Fig. 9(c, d)] is sufficient to separate eukaryotes from prokaryotes. Just as amino acid composition was stripped from the data for dipeptide composition analysis, the dipeptide bias was removed for tripeptide composition.

Universally Outlying Oligopeptides

Use of PCA permits meaningful analysis of the average compositions of the three superkingdoms in relation to each other. Furthermore, one can make absolute assertions regarding the significance of over- or underabundant

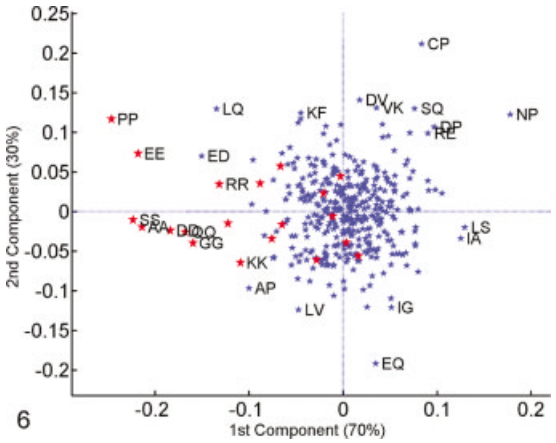


Fig. 6. Dipeptide contribution to principal components of species compositions. The contributions of each dipeptide to the first (x) and second (y) component axes of the coordinate system defined by PCA, using the average dipeptide compositions of the three superkingdoms as a training set (as in Fig. 3(b)), are plotted. Homodipeptides are plotted in red, and heterodipeptides in blue.

oligopeptides in superkingdoms. We observe (see Supplementary Material[ix]) that the counts of dozens of di- and tripeptides significantly deviate from their expectation. These absolute data determine that the comparative preference for homopeptides in eukaryotes relative to prokaryotes is primarily due to their overabundance in eukaryotes, and only to a lesser extent to their underabundance in prokaryotes. PCA representation illustrates an overall trend. A closer look at the specific peptide distributions shows finer details (see Table III and Supplementary Material[ix]). For instance, the homodipeptides *LL*, *II*, and *DD* are less frequently encountered than expected in both eubacteria and eukaryotes, whereas in archaea, which seems to be neutral with respect to *II*, a number of dipeptides are popular, including *LL*, *DD*, *KK*, *EE*, *AA*, and *SS* (see Supplementary Material[ix]).

Figure 10 displays the distribution of rank preferences of all the 400 palindromic tripeptides (including homotripeptides) in the three superkingdoms. It can be seen that both eubacteria and archaea display a similar very low preference for a substantial number of palindromes, and a similar effect, but less marked, can be seen for eukaryotes.

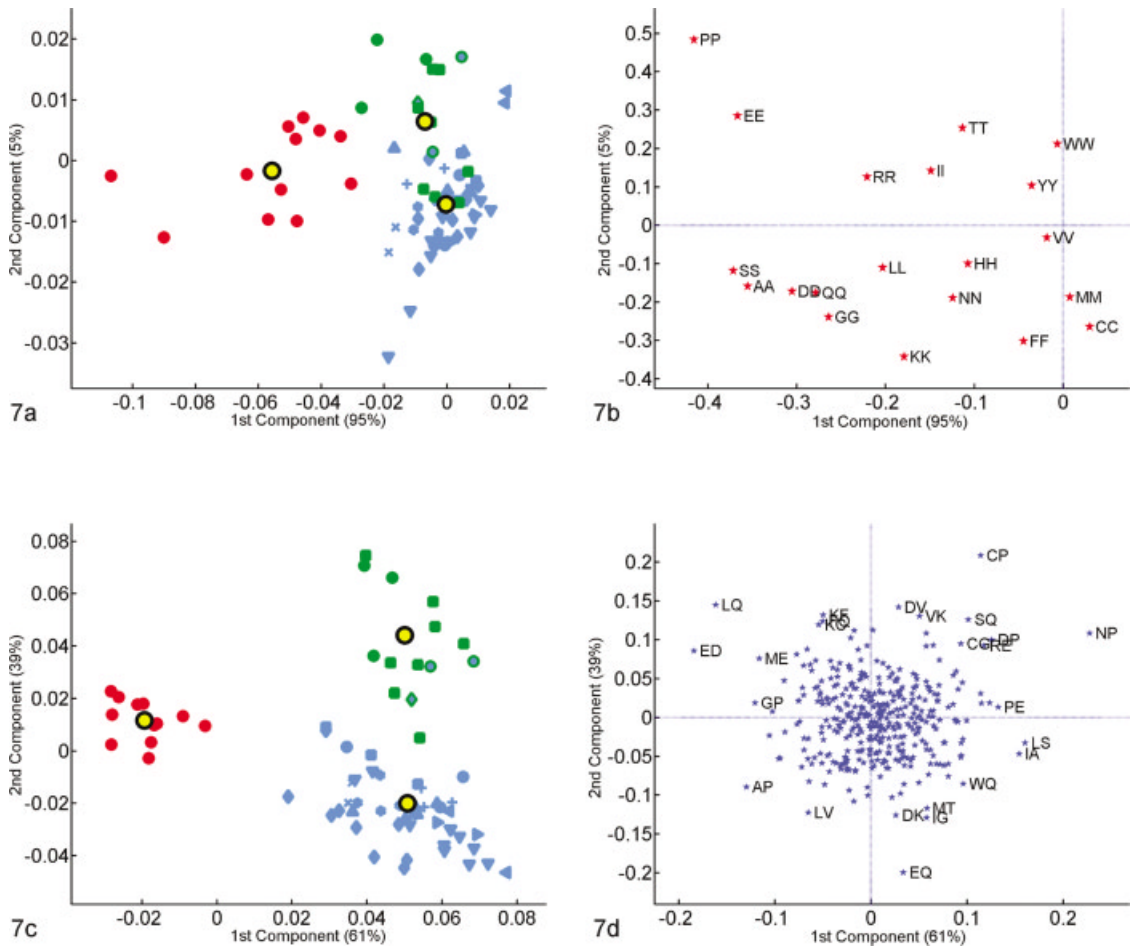


Fig. 7. Homo- versus heterodipeptide contribution to principal components of species compositions. Principal components were computed for homodipeptide frequencies relative to one another (**a**, **b**), as well as for relative heterodipeptide frequencies (**c**, **d**). Average superkingdom compositions served as a training set. As in Figure 3(b), species are plotted (**a**, **c**) according to the two principal components in the respective analyses. As in Figure 6, the dipeptide contribution to the first two principal components is plotted (**b**, **d**).

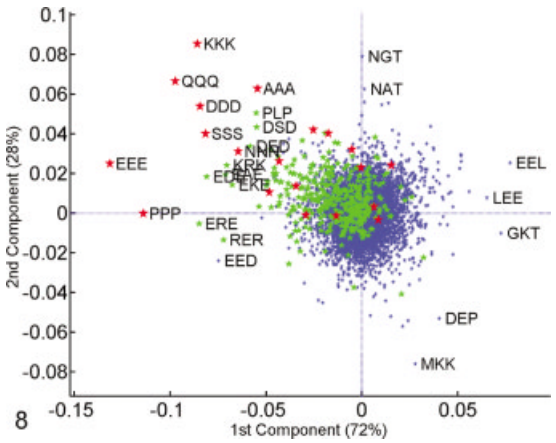


Fig. 8. Tripeptide contribution to principal components of species compositions. The contributions of each tripeptide to the first (x) and second (y) component axes of the coordinate system defined by PCA, using the average tripeptide compositions of the three superkingdoms as a training set (as in Fig. 4), are plotted. Homotripeptides are plotted in red, palindromic tripeptides in green, and other heterotripeptides in blue.

At the other end of the scale, no strong preference of either of the two prokaryote superkingdoms exists for palindromic sequences, but a highly significant preference exists in the case of eukaryotes. Thus, 124 (31%) of all possible palindromes appear in the top 500 (6.25%) preferred tripeptide sequences ($p < 10^{-50}$). If one narrows the windows to 100 at both ends of the scale, the similarity of the two prokaryotes to each other and their difference from eukaryotes is even more clearly pronounced. Of the 100 least favored tripeptides for eukaryotes, 11 are palindromes, whereas for both prokaryote superkingdoms, the corresponding number is 29. Of the 100 most favored tripeptides, 50 are palindromes in the eukaryotes, whereas only 4 and 3, respectively, are palindromes for eubacteria and archaea. Unrelated to the possible biologic significance of our data (see Discussion section), the remarkable similarity of the results for eubacteria and archaea strongly supports the robustness of our approach. It should be noted that although neither of the prokaryote superkingdoms shows significant overall preference for palindromic tripeptides, a significant preference exists for homotripeptides, though much weaker than for eukaryotes, for which

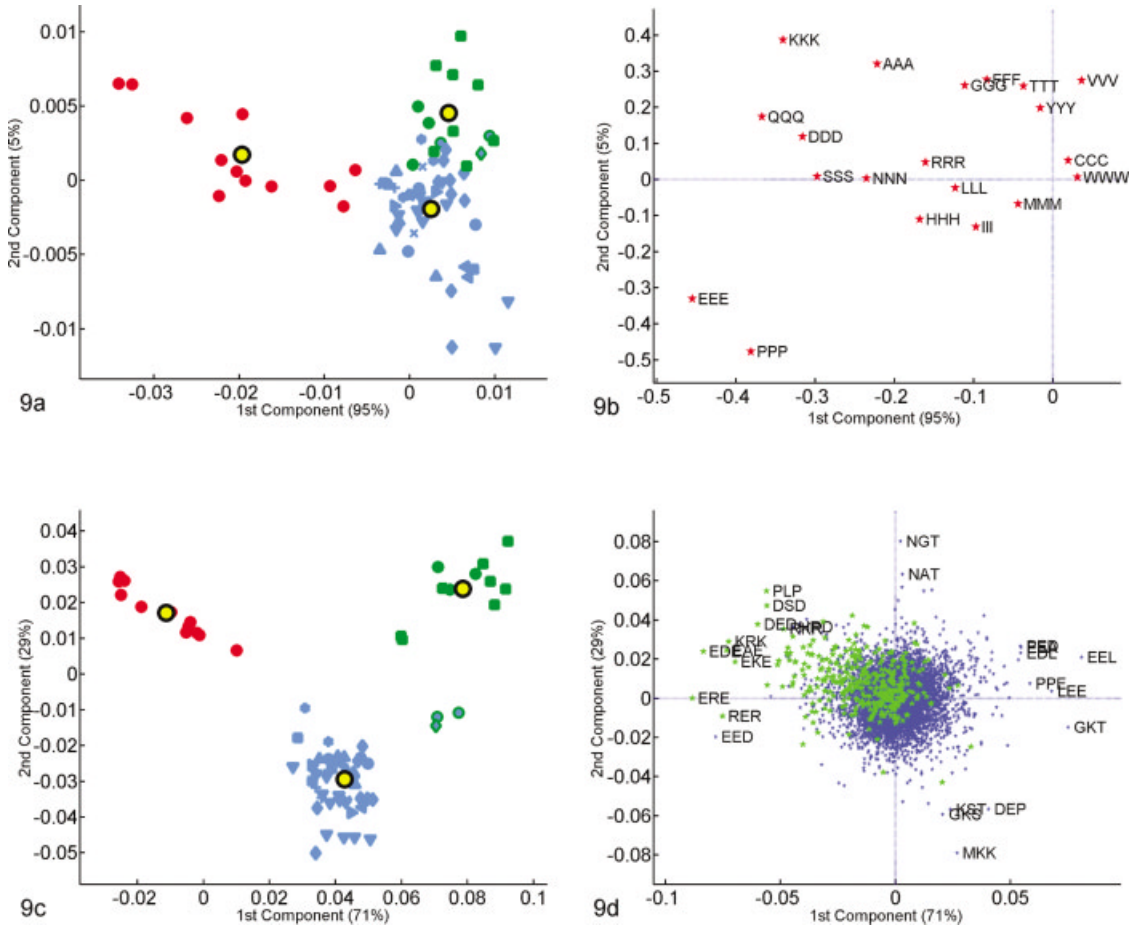


Fig. 9. Homo- versus heterotripeptide contribution to principal components of species compositions. Principal components were computed for homotripeptide frequencies relative to one another (**a**, **b**), as well as for relative heterotripeptide frequencies (**c**, **d**). Average superkingdom compositions served as a training set. As in Figure 4, species are plotted (**a**, **c**) according to the two principal components in the respective analyses. As in Figure 8, the tripeptide contribution to the first two principal components is plotted (**b**, **d**). Homotripeptides are plotted in red, palindromic tripeptides in green, and other heterotripeptides in blue.

12 of the 20 homotripeptides appear in the top 100 most preferred tripeptides, and 10 in the top 20.

At a more specific level, both *LLL* and *III* seem disfavored in all three superkingdoms. Prokaryotes are also deficient in *KKK* and *EEE*, which are particularly abundant in eukaryotes. We further observe a universal, cross-phyla bias for certain classes of peptides. The most prominent such phenomenon is underabundance of *LxL* and *IxI* palindromes. Once again, one can identify highly contrasted palindromes such as *EAE*, *EKE*, and *KEK*, which are all very abundant in eukaryotes and quite rare in prokaryotes, whereas *GPG* shows the opposite behavior, to cite but a few of the discriminating palindromic tripeptides. A closer look at individual species-specific preferential usage of oligopeptides shows that wide variations exist. For instance, whereas *III* ranks low in all sampled eukaryotes, some of them, such as the mammalian proteomes, seem to show no bias, if not a slight preference, against *LLL* (see Supplementary Material[ix]), which stands in contrast with most prokaryotes or even yeasts. Although, in general, eubacteria and archaea are similar in their overall preferences for particular oligopeptides, and differ strongly from eukaryotes, there are individual cases in which they differ strongly between themselves; for example, *PPP* is strongly favored by both eukaryotes and eubacteria, but strongly disfavored by archaea.

Composition Similarity Tree

The compositional vector is a species-specific attribute of proteomes, which suggests that it might serve as an additional useful tool for understanding taxonomic differences. We define a distance function based on composition, and measure this distance between species (see Supplementary Material[ix]). This distance matrix permits construction of a "phylogenetic" or "similarity" tree by standard methods, such as hierarchical clustering (Fig. 11)[15] or neighbor-joining[16] (data not shown). Proteomic signatures capture environmental and genomic effects (e.g., the clustering of thermophilic eubacteria with archaea), as well as phylogenetic or taxonomic signals. Eukaryotes, and vertebrates within them, cluster as monophyletic clades, as do many eubacterial classes (chlamidiae, cyanobacteria, δ/ε-proteobacteria). Other eubacterial classes have one or two outliers (firmicutes, actinobacteria, α-proteobacteria, γ-proteobacteria). Similar relationships can be demonstrated using distance matrices for di- or tripeptides. As a matter of fact, combining the amino acid distance matrix with a weighted matrix for dipeptides generates a tree with perhaps even fewer discordant assignments (see Supplementary Material[ix]). The discordances noted may be attributed, to a large extent, to the nonphylogenetic signals (GC content, thermophilicity) that we detect.

DISCUSSION

Rigorous Compositional Analysis

We have used the powerful technique of PCA to analyze the species- and phyla-specific signatures of amino acid and oligopeptide compositions. PCA had been used previously to compare amino acid compositions in a relatively narrow range of species,[2,3] but in this study, we have covered a large number of proteomes from all three superkingdoms. We have performed a comprehensive, systematic, and quantitative investigation using unbiased PCA to dissect compositional differences into several distinct factors.

Because we examined all superkingdoms and extended our analysis to di- and tripeptides, we took care to avoid methodologic pitfalls taking the following measures:

1. We maintained balanced representation of phyla in the training sets.
2. We assigned equal weights to proteomes of different sizes by expressing relative abundance of constituents as percentages.
3. We discovered phyla separation in an unsupervised procedure, which is a priori ignorant of taxonomy and of the specific compositions of the training sets.
4. We made sure that the separating criteria were predictive for species not used during the inference procedure.
5. We verified robustness to the choice of training species; we also showed robustness to small sample size and, by inference, to confounding effects caused by gene families or lateral gene transfer.
6. We accounted for lower order variability in higher order composition.
7. We accounted for differences between proteins by per-protein computation of expectation.
8. For precise and standardized analysis, we switched to supervised analysis based on phyla average compositions.

Whole, uncurated proteome data were used in their crude form, as they appear in the database: We examined the distribution in the composition space of the species-specific proteomes and of the constituent residues, and attempted to identify the driving forces accounting for these observations. Despite including hypothetical proteins (namely, proteins whose existence was inferred based on ORFs), we found consistent results. Our method of analysis does not rely on ortholog identification or sequence alignment, requirements that often hamper comparative genomics.

Identifiable Factors Influence Proteome Composition

As has been noted by others,[10] species can be differentiated by amino acid composition. Our analysis goes one step further by showing that each of the superkingdoms is compositionally distinct. Furthermore, these compositional differences are evident not only at the amino acid level, but also at least at the di- and tripeptide levels independently.

The vast majority of species cluster around the average of each superkingdom, with the exception of the three thermophilic eubacteria. These are consistently found in the cloud of archaea for amino acid and dipeptide composi-

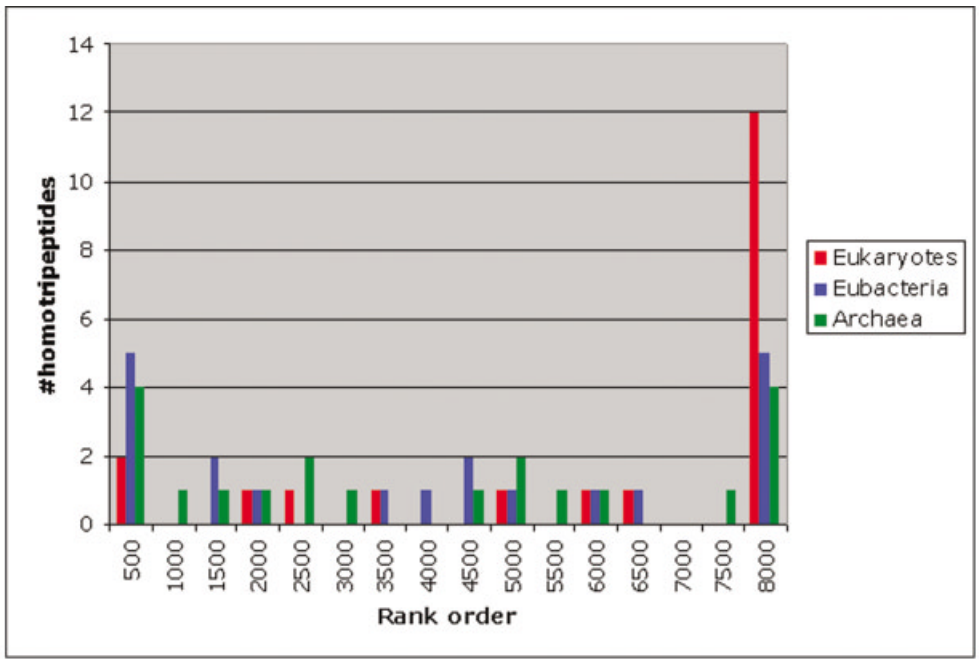
TABLE III. Rank Ordering of Oligopeptides by Over-/Underabundance in Different Superkingdoms

Sort by	Top or bottom	Peptide	Dipeptide analysis: Rank order by Eukaryota	Eubacteria	Archaea
Archaea	Most underabundant	LV	27	20	1
		GP	6	1	2
		AP	70	8	3
		GA	36	5	4
		EP	2	10	5
		ES	1	7	6
		LI	7	14	7
		VG	25	27	8
		TE	38	23	9
		LL	385	3	10
	Most overabundant	LK	371	382	400
		AL	189	398	399
		NP	291	390	398
		EK	391	397	397
		LA	279	386	396
		EI	337	385	395
		PE	346	394	394
		DV	357	280	393
		EE	398	329	392
		LS	126	400	391
Eubacteria	Most underabundant	GP	6	1	2
		KF	51	2	58
		LL	385	3	10
		II	209	4	19
		GA	36	5	4
		IL	202	6	64
		ES	1	7	6
		AP	70	8	3
		EG	10	9	13
		EP	2	10	5
	Most overabundant	LS	126	400	391
		IA	129	399	386
		AL	189	398	399
		EK	391	397	397
		SG	366	396	382
		TP	379	395	388
		PE	346	394	394
		FS	351	393	365
		ID	327	392	357
		KN	320	391	353
Eukaryota	Most underabundant	ES	1	7	6
		EP	2	10	5
		SE	3	34	34
		QS	4	51	88
		KS	5	19	23
		GP	6	1	2
		LI	7	14	7
		DK	8	83	12
		KG	9	22	27
		EG	10	9	13
	Most overabundant	SS	400	338	349
		AA	399	377	379
		EE	398	329	392
		PP	397	41	334
		RR	396	342	380
		QQ	395	368	314
		GG	394	164	103
		KK	393	379	300
		LQ	392	24	333
		EK	391	397	397

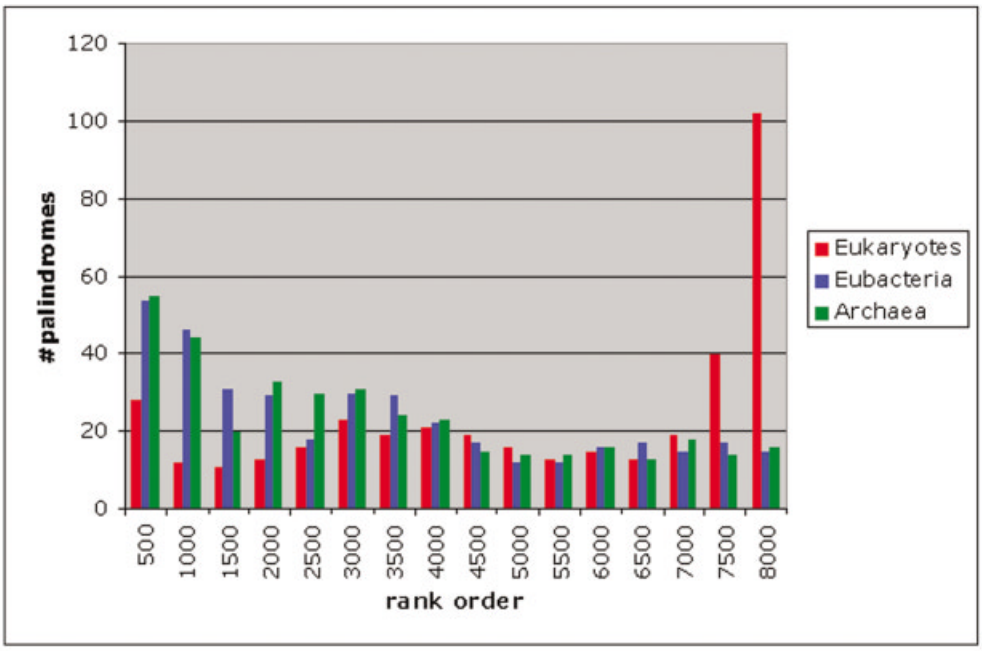
TABLE III. (Continued)

Sort by	Top or bottom	Peptide	Tripeptide analysis Rank order by Eukaryota	Eubacteria	Archaea
Archaea	Most underabundant	LEL	2	3	1
		LAL	10	1	2
		EEE	7995	47	3
		LEI	8	7	4
		LRL	4	4	5
		LKL	1	5	6
		IEL	23	16	7
		IEI	111	10	8
		ILI	98	21	9
		EKK	133	15	10
	Most overabundant	EEL	7949	7998	8000
		LEE	7702	7991	7999
		LRE	7960	7990	7998
		EKL	7990	7999	7997
		LKE	7980	7996	7996
		EAL	7978	7995	7995
		GKT	7957	8000	7994
		LAE	7886	7993	7993
		ERL	7971	7985	7992
		KEL	7976	7983	7991
Eubacteria	Most underabundant	LAL	10	1	2
		KKK	7953	2	31
		LEL	2	3	1
		LRL	4	4	5
		LKL	1	5	6
		LKI	22	6	19
		LEI	8	7	4
		IKI	158	8	25
		LLL	481	9	26
		IEI	111	10	8
	Most overabundant	GKT	7957	8000	7994
		EKL	7990	7999	7997
		EEL	7949	7998	8000
		LEK	7988	7997	7987
		LKE	7980	7996	7996
		EAL	7978	7995	7995
		AIA	7846	7994	7642
		LAE	7886	7993	7993
		KKI	7561	7992	7627
		LEE	7702	7991	7999
Eukaryota	Most underabundant	LKL	1	5	6
		LEL	2	3	1
		LQL	3	19	195
		LRL	4	4	5
		DPP	5	130	149
		LSL	6	11	22
		LLI	7	23	51
		LEI	8	7	4
		EES	9	230	73
		LAL	10	1	2
	Most overabundant	SSS	8000	7788	7386
		AAA	7999	7555	7912
		PPP	7998	7754	2064
		QQQ	7997	5528	5869
		GGG	7996	7873	7944
		EEE	7995	47	3
		DDD	7994	1185	1595
		PLP	7993	7452	7543
		PAP	7992	7936	7542
		GSG	7991	7865	7532

Note: Dipeptide (**a**) and tripeptide (**b**) frequency counts were compared to their expected values based on residue and dipeptide compositions, respectively, and averaged across phyla to evaluate oligopeptide over- or underabundance. These values were used to rank order dipeptides and tripeptides according to each superkingdom: archaea (A), eubacteria (B), or eukaryota (E). Peptides are color-coded, as in Table II. The top 10 are highlighted in blue, the top 50 in turquoise, and in tripeptides, the top 250 in each ranking in green, while bright and light yellow highlight the bottom 10 and 50 peptides, respectively, in each ranking.



(a)



(b)

Fig. 10. The distribution of the ranks of palindromes for the three superkingdoms. In each superkingdom, we ranked all possible tripeptides according to their Z values, with the tripeptide having the lowest Z value receiving rank 1, and so on. For nonoverlapping windows of 500 ranks, we then counted for each superkingdom how many homotripeptides (**a**) and how many palindromes (**b**) are in each window.

tions, but segregate from both archaea and other eubacteria in the case of tripeptides [see Fig. 4(b)].

We went on to analyze the different residues whose frequencies account for variation between species and phyla. The principal axis in compositions of individual species reflects the wide variability in GC content of eubacterial genomes [Figs. 2(d) and 5(a)], a conclusion similar to that reached by others.[2,3] Known, species-

TABLE IV. Characteristics of the Archaeal Species in the Data Set in Terms of Extreme Conditions

Species	Characteristics
Methanocaldococcus jannaschii	Thermophilic
Methanosarcina mazei	Halotolerant
Archaeoglobus fulgidus	Hyperthermophilic
Pyrococcus furiosus	Hyperthermophilic, radiation resistant
Halobacterium sp. NRC-1	Halophilic
Thermoplasma acidophilu	Thermophilic, acidophilic
Methanopyrus kandleri	Hyperthermophilic
Methanothermobacter thermautotrophicus	Thermophilic
Aeropyrum pernix	Hyperthermophilic
Sulfolobus solfataricus	Hyperthermophilic, acidophilic
Pyrobaculum aerophilum	Hyperthermophilic

Note: Ten out of the 11 archaeal species in the data set are considered extremophiles, and one, *Methanosarcina mazei,* is halotolerant. The Characteristics column describes these species in terms of extreme conditions.

specific dinucleotide patterns in genomic DNA[17] suggest that these might be involved in determination of the proteomic signature, but such a connection remains to be established. When supervised training is employed to coalesce all intrasuperkingdom variability [Figs. 2(e) and 5(b)], the principal plane observed is focused at intersuperkingdom separation. Thus, we can identify two orthogonal axes that separate eubacteria-like compositions from either archaea-like or eukaryote-like compositions.

Compositional differences between eubacteria and archaea may be related to the extreme habitats occupied by the latter (see Table IV). Single-residue compositional preferences related to thermophilicity, common to both thermophilic eubacteria and archaea have been identified.[2,3] We validate this single residue result and establish its analog for dipeptide compositional analysis, which clusters eubacterial thermophiles within the archaeal cloud. In tripeptide analysis, however, eubacterial thermophiles segregate as an individual cluster distinct from both mesophilic eubacteria and archaea [Fig. 4(b)]. We do not find GC content to be a major determinant of thermophiles' whole genome composition, as it is for their ribosomal and transfer RNAs.[18,19] The recently established existence of mesophilic and even psychrophilic archaea[20] motivates further investigation of thermophilicity-related composition once their genomes become available.

Systematic Compositional Bias Versus Differential Protein Content

The individuality of species- and phyla-specific proteomic signatures may be due to several factors. Eukaryote-specific composition may originate in the existence of whole families of structural and functional proteins present in eukaryotes, and especially in multicellular organisms, which are totally lacking in prokaryotes. Specific examples would include structural proteins, such as collagen, and proteins of the nervous system and other signaling entities. A recent development that may be pertinent is the recognition[21,22] that protein sequences may be "natively unfolded," namely, devoid of intrinsic structure in isolation. It has been suggested that such proteins possess the capacity to interact with other proteins to form scaffold structures or signaling complexes[23] characteristic of higher organisms (e.g., in their nervous systems). Indeed, Dunker and Obradovic[22] have indicated a much higher percentage of unfolded sequences in eukaryotes than in prokaryotes. They have designated certain amino acids, primarily hydrophobic, as "order promoting," and others, primarily hydrophilic, as "disorder promoting."[24] Our initial analyses of the amino acid compositional differences underlying the PCA separation according to superkingdoms (see below) do not appear to be directly correlated with this classification, but a more rigorous treatment will be required to address this issue.

The di- and tripeptide analyses show clear separation between the three superkingdoms, just as for the monomers. Whereas simple arguments such as GC composition, tRNA abundance, ability to synthesize amino acids, and/or habitat temperature can be invoked to explain the separations seen at the amino acid–level, the driving force(s) acting at the dipeptide level must be different, because we were careful to eliminate the bias contributed by the amino acid composition in our dipeptide analysis (see Methods section) and, similarly, took care to remove the dipeptide bias in our tripeptide analysis. In any event, the clearest observation is the strong representation of homodipeptide and -tripeptide sequences in eukaryotes relative to both prokaryotic superkingdoms, even though prokaryotes also show some preference for such sequences. The overabundance of amino acid runs in eukaryotic proteomes, probably due to trinucleotide replication slippage, has been analyzed by Karlin et al.,[5,25] who pointed out that many such runs are observed in proteins associated with human disease conditions (e.g., the huntingtin protein associated with Huntington's disease[5,26]). However, amino acid runs in eukaryotes relative to prokaryotes may be a result of their occurrence in eukaryotic-specific domains or motifs.[27] For example, proline-rich motifs occur in polyproline-II sequences, which participate in signaling pathways[28] and in attachment domains involved in assembly of acetylcholinesterase subunits to tetramers at cholinergic synapses.[29,30] Similarly, the higher frequency of palindromes in eukaryotes might be due in part to the presence of the alternating charge runs noted in such proteins as the human triplet disease protein, atrophin 1, and periodic histidine-containing sequences found in, for example, various *Drosophila* developmental proteins.[5] Purely as a conjecture, one might envisage that the very different nature of lipids in archaea,[31] relative to both eukaryotes and eubacteria, might result in certain protein motifs recognizing them preferentially and thus being selected for, and thus preferred, in archaea. Note that eukaryotic specificity of hexanucleotide repeats due to replication slippage cannot fully explain the observed overabundance of tripeptide palindromes: Had such repeats been the dominant feature in the data, any outlying

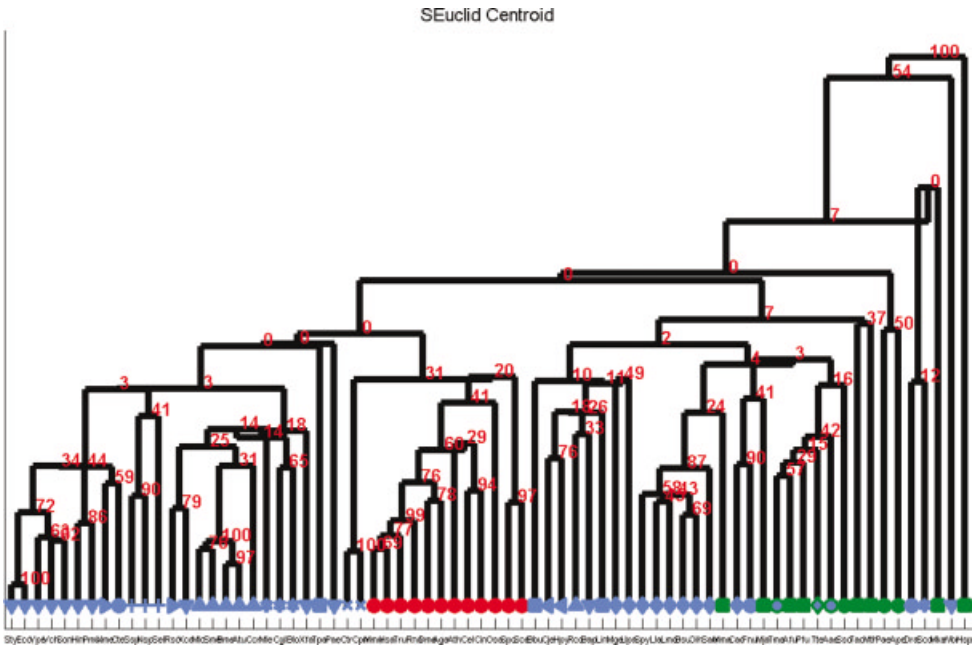


Fig. 11. Species similarity tree based upon amino acid compositions. The tree was obtained by hierarchical clustering (see Methods section). Species are denoted by their three-letter acronym, and phyla by a characteristic symbol (see Table I). Note that this clustering method may result in negative-length branches. Bootstrap values are indicated.

XYX palindrome would be roughly as abundant as its *YXY* counterpart, in contrast to what we observe.

We detect many universally underabundant oligopeptides, including a substantial number of palindromes (see Fig. 10, Table III, and Supplementary Material[ix]). These negative preferences might reflect general structural constraints. Examination of computer molecular models shows that steric hindrance cannot explain this aversion shared by all superkingdoms (data not shown), so more complex explanations must be sought. One possible factor affecting preference might be some steric restriction of the ribosomal machinery, which might bias against certain sequences.

Therefore, it will be of interest to determine next which of these dynamics is responsible for specific compositional characteristics of phyla, such as overabundance of uncharged polar residues, and palindromic or homopeptide runs in eukaryotes. Are these due to a systematic bias or functional class enrichment? These questions can be approached, for instance, by examination of sets of orthologous proteins from different phyla, thus controlling for functional enrichment, with each protein being equally represented within each species. Preliminary results wherein orthologous sets of proteins from *E. coli* and *S. cerevisiae* were extracted from the COG database[b] and compared, suggest that species-specific attributes leave a bigger mark than functional relatedness, because the two ortholog sets are closer to their proteomes of origin than to each other (data not shown).

Similarity Trees From Proteomic Signatures

Beyond its descriptive nature, exploration of proteomic residue compositions can have important practical implications. One fascinating aspect of this work is that one can, in contrast to alternative methods, perform comparative analyses across genomes or proteomes, to define relationships even for functionally and structurally unrelated proteins. There is no need to compare ortho- or paralogs. All that is required are the residue compositions of a statistically significant set of proteins.

The compositionally derived similarity trees (e.g., for amino acids, or di- or tripeptides) demonstrate that closely related proteomes display similar compositions. This evolutionary memory is so strong that these trees stand in surprisingly good agreement with recently published phylogenetic trees, whose topologies substantially depend on the method used for tree reconstruction (see recent reviews[1,9,32]). Our study adopts the paradigm shift in phylogenetic reconstruction from an individual-gene approach towards genome-wide methods in order to avoid the usual vagaries caused by unrecognized horizontal gene transfer, unrecognized paralogy, highly variable rates of gene evolution, or misalignment, common in phylogenies based on single genes. How accurately evolutionary history can be recovered based on compositional analysis of longer polypeptides remains to be examined.

Future Prospects

The distribution of amino acids and oligopeptides is a key element in bioinformatics applications across the board. Sequence alignment, functional annotation, and

phylogenetic analysis, to name a few, all employ null-models for protein compositions. This study highlights the sensitivity of such models to the identity of the organism under study and suggests improved, species-aware computational methods.

Furthermore, based on these results, it is intriguing to analyze compositional features of various protein subsets both within and between species. For instance, one can compare the vocabularies of transcription factors or extracellular proteins across species, and see whether they also manifest specific recognizable functional, compartmental, and cell-, tissue-, or species-specific signatures. One can also investigate to what extent species-specific sequences differ from the rest of the proteome. The yeast-like character was not lost on partition of the yeast proteome into chromosome-specific subsets, the smallest of which contained only 105 proteins, at either the monomer [Fig. 2(d′)] or dipeptide level (data not shown). This robustness of analysis permits the application of this method to partial proteomic sets.Therefore, we are currently examining whether compositional similarities can be used as an additional means to assign phylogenetic proximity for relatively small sets of proteins (e.g., those encoded by viral or organellar DNA). In a similar vein, this approach may prove useful to scan genomes for the presence (and perhaps dating) of evidence for horizontal gene transfer.[33]

ACKNOWLEDGMENTS

Our thanks to Beni Yakir, Uri Einav, and Dedi Segal for preliminary work that motivated this study, to Ed Trifonov for thoughtful discussions, and to Dan Graur, Tzachi Pilpel, Ron Shamir, and Ron Unger for helpful comments on the manuscript.

REFERENCES

1. Gribaldo S, Philippe H. Ancient phylogenetic relationships. Theor Popul Biol 2002;61:391–408.
2. Kreil DP, Ouzounis CA. Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res 2001;29:1608–1615.
3. Tekaia F, Yeramian E, Dujon B. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: A global picture with correspondence analysis. Gene 2002;297:51–60.
4. Karlin S, Brocchieri L, Trent J, Blaisdell BE, Mrazek J. Heterogeneity of genome and proteome content in bacteria, archaea, and eukaryotes. Theor Popul Biol 2002;61:367–390.
5. Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ. Amino acid runs in eukaryotic proteomes and disease associations. Proc Natl Acad Sci U S A 2002;99:333–338.
6. Korbel JO, Snel B, Huynen MA, Bork P. SHOT: A web server for the construction of genome phylogenies. Trends Genet 2002;18: 158–162.
7. Tekaia F, Lazcano A, Dujon B. The genomic tree as revealed from whole proteome comparisons. Genome Res 1999;9:550–557.
8. Ling L, Wang J, Cui Y, Li W, Chen R. Proteome-wide analysis of protein function composition reveals the clustering and phylogenetic properties of organisms. Mol Phylogenet Evol 2002;25:101–111.
9. Wolf YI, Rogozin IB, Grishin NV, Koonin EV. Genome trees and the tree of life. Trends Genet 2002;18:472–479.
10. Echols N, Harrison P, Balasubramanian S, Luscombe NM, Bertone P, Zhang Z, Gerstein M. Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. Nucleic Acids Res 2002;30:2515–2523.
11. Stuart GW, Moffett K, Leader JJ. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. Mol Biol Evol 2002;19:554–562.
12. Karlin S, Burge C. Dinucleotide relative abundance extremes: A genomic signature. Trends Genet 1995;11:283–290.
13. Joliffe IT, Morgan BJ. Principal component analysis and exploratory factor analysis. Stat Methods Med Res 1992;1:69–95.
14. Vieille C, Zeikus GJ. Hyperthermophilic enzymes: Sources, uses, and molecular mechanisms for thermostability. Microbiol Mol Biol Rev 2001;65:1–43.
15. Johnson SC. Hierarchical clustering schemes. Psychometrika 1967;32:241–254.
16. Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4:406–425.
17. Gentles AJ, Karlin S. Genome-scale compositional comparisons in eukaryotes. Genome Res 2001;11:540–546.
18. Galtier N, Lobry JR. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol 1997;44:632–636.
19. Buckley DH, Graber JR, Schmidt TM. Phylogenetic analysis of nonthermophilic members of the kingdom crenarchaeota and their diversity and abundance in soils. Appl Environ Microbiol 1998;64:4333–4339.
20. Preston CM, Wu KY, Molinski TF, DeLong EF. A psychrophilic crenarchaeon inhabits a marine sponge: *Cenarchaeum symbiosum* gen. nov., sp. nov. Proc Natl Acad Sci U S A 1996;93:6241–6246.
21. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 2000;41:415–427.
22. Dunker AK, Obradovic Z. The protein trinity—linking function and disorder. Nat Biotechnol 2001;19:805–806.
23. Wright PE, Dyson HJ. Intrinsically unstructured proteins: Reassessing the protein structure–function paradigm. J Mol Biol 1999;293:321–331.
24. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered proteins. Proteins 2001;42:38–48.
25. Karlin S, Chen C, Gentles AJ, Cleary M. Associations between human disease genes and overlapping gene groups and multiple amino acid runs. Proc Natl Acad Sci U S A 2002;99:17008–17013.
26. Perutz MF. Glutamine repeats and neurodegenerative diseases: Molecular aspects. Trends Biochem Sci 1999;24:58–63.
27. Gerstein M. A structural census of genomes: Comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. J Mol Biol 1997;274:562–576.
28. Kay BK, Williamson MP, and Sudol M. The importance of being proline: The interaction of proline-rich motifs in signaling proteins with their cognate domains. FASEB J 2000;14:231–241.
29. Simon S, Krejci E, Massoulie J. A four-to-one association between peptide motifs: Four C-terminal domains from cholinesterase assemble with one proline-rich attachment domain (PRAD) in the secretory pathway. EMBO J 1998;17:6178–6187.
30. Perrier AL, Massoulie J, Krejci E. PRiMA: The membrane anchor of acetylcholinesterase in the brain. Neuron 2002;33:275–285.
31. van de Vossenberg JL, Driessen AJ, Konings WN. The essence of being extremophilic: The role of the unique archaeal membrane lipids. Extremophiles 1998;2:163–170.
32. Hedges SB. The origin and evolution of model organisms. Nat Rev Genet 2002;3:838–849.
33. Brown JR. Ancient horizontal gene transfer. Nat Rev Genet 2003;4:121–132.

Website References

i. EBI proteomic database at ftp://ftp.ebi.ac.uk/pub/databases/spproteomes/fasta_files/proteomes/, December 2002.
ii. The COG database at http://www.ncbi.nlm.nih.gov/COG/, November 2002.
iii. NCBI taxonomy browser at http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/, December 2002.
iv. The rat genome database at ftp://ftp.ncbi.nih.gov/genomes/R_norvegicus/rn_protnr.gz, July 2002.
v. The mosquito genome database at ftp://ftp.ensembl.org/pub/current_mosquito/data/fasta/pep/Anopheles_gambiae.MOZ2.pep.fa.gz
vi. The fugu genome database at ftp://ftp.ensembl.org/pub/current_fugu/data/fasta/pep/Fugu_rubripes.FUGU2.pep.fa.gz, (fugu-9.1), November 2002.

vii. The rice genome database at ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/OSA1.pep, December 2002.
viii. The ciona genome database at ftp://ftp.jgi-psf.org/pub/JGI_data/Ciona/v1.0/ciona.prot.fasta.gz, December 2002.
ix. http://www.weizmann.ac.il/~joel/papers/suppl/peer_2003

APPENDIX: MATHEMATICAL DERIVATIONS AND PROCEDURES

Observed Frequencies

We derived the observed raw counts (the number of occurrences) of all residues, as well as of all overlapping di- and tripeptide fragments comprising the protein. Residues labeled "unknown" in the data and fragments containing them were discarded. Let $N_1(s)$,$N_2(s)$, and $N_3(s)$, respectively, be the total number of residues, dipeptides, and tripeptides along each protein sequence s. Obviously, if there are no unknown residues, then

$$N_1(s) = N_2(s) + 1 = N_3(s) + 2 \tag{1}$$

For each compositional element c (which can be a residue, a dipeptide, or a tripeptide) let $l(c)$ be its length (which can be 1, 2, or 3, respectively). Denote the observed raw count of c along the sequence s by $f_o(c,s)$. For a complete proteome, S, we divided the total raw count of c along all the sequences of S by their total length, thus computing the observed percent frequency of c along the entire proteome:

$$P_o(c, S) = \frac{\sum_{s \in S} f_o(c, s)}{\sum_{s \in S} N_{l(c)}(s)}. \tag{2}$$

Expected Overall Frequencies

For single residues, we analyzed their observed frequencies. For dipeptides and tripeptides, we compared the observed frequencies with the frequency $P_e(c,S)$ that one would expect: Dipeptide expected frequencies were based on single–amino acid counts, whereas tripeptide expected frequencies were based on dipeptide counts.

We now elaborate on the computation of the expected overall frequencies $P_e(c,S)$ from lower order counts. We argue that lower order frequencies may significantly vary between one protein sequence and another (see detailed discussion of this issue below). The computation of expected frequencies should, therefore, account for such variation by computing the expected count $f_e(c,s)$ for each sequence s, and summing per-sequence expectancies:

$$P_e(c, S) = \frac{\sum_{s \in S} f_e(c, s)}{\sum_{s \in S} N_{l(c)}(s)}. \tag{3}$$

Expected Dipeptide Counts Per Sequence

It remains to be shown how we computed the expected per-sequence counts, $f_e(c,s)$. For a dipeptide AB, we can compute this quantity by summation of the AB-occurrence probability along s. Formally, denote the amino acids in s by $s[1], s[2],\ldots, s[N_1(s)]$. Then*:

$$\begin{aligned} f_e(AB, s) &= \sum_{pair\ s[i],s[i+1]} \text{Prob}(s[i] = A, s[i+1] = B) \\ &= N_2(s) \times \frac{\#\text{ possible } AB \text{ pairs}}{\#\text{ possible pairs}} \\ &= \frac{N_2(s)}{N_1(s)(N_1(s) - 1)} \times (\#\text{ possible } AB \text{ pairs}). \end{aligned} \tag{4}$$

We now distinguish two cases.

1. For homodipeptides:

$$\#\text{ possible } AA \text{ pairs} = f_o(A, s)(f_o[A, s] - 1) \tag{5}$$

2. For heterodipeptides:

$$\#\text{ possible } AB \text{ pairs} = f_o(A, s) f_o(B, s) \tag{6}$$

Expected Tripeptide Counts Per Sequence

In the following subsection, all formulae refer to a protein sequence s, which is omitted from notation.

For ABC tripeptides, we rely on the number of runs (of length ≥ 1) of an amino acid B, $R(B) = f_o(B) - f_o(BB)$. Let $U(B)$ be the number of singleton B's (B runs consisting of a single residue). If $f_o(B) = 1$, then also $U(B) = 1$. Otherwise, $U(B)$ is a random variable. Let $E(U[B])$ be the expectation of this random variable. Let $U_1(B)$, $U_2(B)$,..., $U_{R(B)}(B)$ be the binary random variables that are 1 if the respective run is a singleton:

$$\begin{aligned} E(U[B]) &= \sum_{i=1}^{R(B)} E(U_i[B]) \\ &= \sum_{i=1}^{R(B)} \text{Prob}(U_i[B] = 1) \\ &= \sum_{i=1}^{R(B)} \text{Prob}(\text{the } i\text{th run is a singleton}) \\ &= R(B) \times \text{Prob}(\text{a specific run is a singleton}). \end{aligned} \tag{7}$$

The process of choosing a random sequence that preserves the $f_o(B)$ and $f_o(BB)$ counts involves partitioning the BB dimers into the $R(B)$ runs. The number of such partitions is $\binom{R(B) + f_o(BB) - 1}{R(B) - 1}$, out of which $\binom{R(B) + f_o(BB) - 2}{R(B) - 2}$ partitions have a specific singleton run. Therefore, the probability of a specific singleton run is

*Formulae that are slightly more accurate than Eqs. (4–6) are obtained when distinguishing the number $f_o^L(A, s)$ of times A is observed as the left (first) residue within a dipeptide, from the number $f_o^R(A, s)$ of times A is observed as the right (second) residue within a dipeptide. This distinction involves separate treatment of the head and tail of the sequence. In our computer code, we have implemented this accurate computation.

$$\frac{\binom{R(B)+f_o(BB)-2}{R(B)-2}}{\binom{R(B)+f_o(BB)-1}{R(B)-1}} = \frac{R(B)-1}{R(B)+f_o(BB)-1} = \frac{R(B)-1}{f_o(B)-1}. \quad (8)$$

For each residue **B** with $(f_o[B]) > 1)$ the expected number of singletons is, therefore, implied by Eqs. (7) and (8):

$$E(U[B]) = \frac{R(B)(R[B]-1)}{f_o(B)-1}. \quad (9)$$

Analogously to the distinction between homo- and heterodipeptides, we now need to distinguish several cases, as follows:

1. For each heterotripeptide ABC ($A \neq B$ and $C \neq B$), the expected count $f_e(ABC)$ is the product of $E(U[B])$ by the estimated probability of C following a run of B's and A preceding such a run*:

$$f_e(ABC) = E(U[B]) \times \frac{f_o(AB)}{f_o(B)-f_o(BB)} \times \frac{f_o(BC)}{f_o(B)-f_o(BB)}. \quad (10)$$

2. For semihomotripeptides† ABB or BBC, one needs to consider only positions that are beginnings or ends, respectively, of nonsingleton runs. There are expected to be $R(B) - E(U[B])$ such positions. Thus, one needs to multiply $R(B) - E(U[B])$ by the probability of encountering the non-B residue:

$$f_e(ABB) = \{R - E(U[B])\} \times \frac{f_o(AB)}{f_o(B)-f_o(BB)}$$

$$f_e(BBC) = \{R - E(U[B])\} \times \frac{f_o(BC)}{f_o(B)-f_o(BB)}. \quad (11)$$

3. For homotripetides BBB, the raw count is a direct function of $U(B)$, $R(B)$, and $f_o(BB)$:

$$f_o(BBB) = f_o(BB) - (R(B) - U(B)) = 2f_o(BB) - f_o(B) + U(B).$$

Therefore,

$$f_e(BBB) = 2f_o(BB) - f_o(B) + E(U(B)) \quad (12)$$

Per-Sequence Expectations Versus Naive Computation

We note that naive computation of expectation may be based on frequencies observed along the complete proteome, that is,

$$P_e^{naive}(AB, S) = P_o(A, S) \times P_o(B, S);$$

$$P_e^{naive}(ABC, S) = \frac{P_o(AB, S) \times P_o(BC, S)}{P_o(B, S)}. \quad (13)$$

This computation may produce erroneous predictions as a result of protein-dependent frequency. Such an error may be best understood by an example: Suppose elements A and B are very common, but one is present only in membranal proteins, and the other, only in globular ones. The naive computation would erroneously expect many AB combinations, whereas the per-protein computation, which we employ, avoids such culprits.

Computing *Z* Scores

For every compositional element c and proteomic set S, we computed a surprise score for the observed frequency given its expectation:

$$Z\ \text{Score}(c, S) = \frac{\sqrt{\overline{N_{l(c)}(S)}}(P_o[c, S] - P_e[c, S])}{\sqrt{\overline{P_e(c, S)}}}. \quad (14)$$

Under the null hypothesis of independent residue distribution, this Z-score quantity has a standard-normal distribution and can thus measure our degree of surprise on observing a given frequency.

Only Homopeptides or Heteropeptides

Relative overabundance of amino acid runs in the human genome,[25] as well as in our own data (presented in part in Tables II, III), motivated additional examination of homodipeptide/heterodipeptide counts separately. Percent frequencies for each homodipeptide AA among all homodipeptides were computed by

$$P_o^{=}(AA, S) = \frac{f_o(AA, S)}{N_2^{=}(S)}, \quad (15)$$

where $f_o(AA, j)$ is the raw count and $N^{=}(j)$ is the total count of all homodipeptides in the given proteome. Similarly,

$$P_o^{\neq}(AB, S) = \frac{f_o(AB, S)}{N_2^{\neq}(S)} \quad (16)$$

was computed for heterodipeptides. We calculated the expected probabilities for these quantities by

$$P_e^{=}(AA, S) = \frac{P_e(AA, S)}{\sum_A P_e(AA, S)};$$

$$P_e^{\neq}(AB, S) = \frac{P_e(AB, S)}{\sum_{A \neq B} P_e(AB, S)}. \quad (17)$$

Similar reasoning was used to considering only homotripeptides or heterotripeptides.

Computing Deviations From Expectancy

To compare observed and expected probabilities, we used the formula,

*As in the case of dipeptides, slightly more accurate formulae are obtained by separate handling of the head and the tail of the sequence.

†Tripeptides that have one homodimer, with an additional, different residue.

$$Normalize(c, S) = \frac{P_o(c, S) - P_e(c, S)}{\sqrt{P_e(c, S)}}, \qquad (18)$$

where the scale factor $\sqrt{P_e(c, S)}$ eliminates the distortion that arises from magnifying large differences between small probabilities versus small differences between large ones; but in contrast to the Z score above, the proteome size was not used, in order to enforce equal representation for each proteome in the comparative analysis. This normalized value was used for downstream comparative analyses, by PCA and hierarchical clustering.

Principal Component Analysis

To systematically represent the sequence compositions for all the protein sets, we applied PCA to each data set (e.g., single residues, dipeptides, or tripeptides). The input to PCA is a set of vectors, one per proteomic set (species). Vector entries are either all single-residue frequencies, or all normalized dipeptide or tripeptide frequencies [see Eq. (18)]for the corresponding proteomic set, comprising 20-, 400-, or 8000-D vectors in respective analyses*.

PCA is aimed at summarizing multidimensional data into a series of mutually orthogonal vectors in residue or peptide space. The first principal component is a multidimensional axis along which the species are most diverse. Each species is positioned along this axis by a weighted sum of its (normalized) frequency values for each of the dimensions (residues, dipeptides, or tripeptides). The weight contributed by each dimension is indicated by its position along this principal axis. Positions of species along the principal axis define a single attribute identifying the composition of species-specific protein sets that deviate most substantially from the others, and also which compositional elements are responsible for this deviation (e.g. to what extent each sequence element discriminates among the different proteome sets). These quantities define the direction of the primary axis in a high dimension. The second component identifies the significant outliers after the first component axis is subtracted out, and so on.

The empirical variance in the projection of the training set on each axis, divided by the total training-data variance, defined the relative contribution of that axis to the overall variance.

Phylogenetically Unbiased, Unsupervised Training

We note that the comparison of frequencies allows each proteome set to contribute equally to the analyses, and avoids bias toward larger proteomes. The downside of this feature is imbalance between superkingdoms, because most available genomes belong to prokaryotes. Indeed, composition analyses of complete genomes are dominated by phenomena in the prokaryotic world.[3] To keep an equal balance between eukaryotic and prokaryotic genomes, we devised a two-step procedure: We first established the principal axes of the space, using two selected training sets constituted, respectively, of an equal number (5 in this article) of arbitrarily chosen species from the analyzed superkingdoms. Second, we considered the complete set of proteomes and computed their position along these principal axes. The procedure was repeated for comparative analysis of eukaryotes versus eubacteria, archaea, or both. The procedure was repeated with different training sets of randomly chosen species (data not shown).

This unsupervised training procedure was applied to various compositional elements, such as residues, dipeptides, tripeptides, and homo-/heterodipeptides.

Supervised Training

Once superkingdom separation had been established in an unsupervised manner, we highlighted superkingdom differences by averaging the vectors representing each superkingdom. We used the three averages as the training set for establishment of principal component coordinates. This supervised training procedure was applied to the same compositional elements as in the unsupervised training procedure.

Similarity Tree by Hierarchical Clustering

We computed pairwise distances between amino acid frequency vectors of each species. Distances were standardized Euclidean distances (normalized for the variation of each coordinate). We applied hierarchical clustering to the distance matrix by iteratively joining the closest pair and replacing the vectors in a joint cluster with their centroid (see MATLAB 12 help for details). Bootstrapping values were obtained by 100 iterations, in each of which the 20 dimensions of the input were resampled, with replacement, and the clustering procedure was applied to the resampled data set.

*Homopeptide analyses employed 20-D vectors of homopeptide frequencies, computed and normalized by Eqs. (15), (17), and (18). Heterodipeptide/heterotripeptide analyses employed 380- and 7980-D vectors, respectively, computed and normalized by Eqs. (15), (17), and (18).