GBE

Genome Biology and Evolution

SMBE

Global Shifts in Genome and Proteome Composition Are Very Tightly Coupled

Maria Brbić[1,2], Tobias Warnecke[3], Anita Kriško[2], and Fran Supek[1,4,*]

[1]Division of Electronics, Rudjer Boskovic Institute, Zagreb, Croatia

[2]Molecular Basis of Ageing, Mediterranean Institute for Life Sciences (MedILS), Split, Croatia

[3]MRC Clinical Sciences Centre, Imperial College, Hammersmith Campus, London, United Kingdom

[4]EMBL/CRG Systems Biology Unit, Centre for Genomic Regulation, Barcelona, Spain

*Corresponding author: E-mail: fran.supek@irb.hr.

Accepted: May 9, 2015

Abstract

The amino acid composition (AAC) of proteomes differs greatly between microorganisms and is associated with the environmental niche they inhabit, suggesting that these changes may be adaptive. Similarly, the oligonucleotide composition of genomes varies and may confer advantages at the DNA/RNA level. These influences overlap in protein-coding sequences, making it difficult to gauge their relative contributions. We disentangle these effects by systematically evaluating the correspondence between intergenic nucleotide composition, where protein-level selection is absent, the AAC, and ecological parameters of 909 prokaryotes. We find that G + C content, the most frequently used measure of genomic composition, cannot capture diversity in AAC and across ecological contexts. However, di-/trinucleotide composition in intergenic DNA predicts amino acid frequencies of proteomes to the point where very little cross-species variability remains unexplained (91% of variance accounted for). Qualitatively similar results were obtained for 49 fungal genomes, where 80% of the variability in AAC could be explained by the composition of introns and intergenic regions. Upon factoring out oligonucleotide composition and phylogenetic inertia, the residual AAC is poorly predictive of the microbes' ecological preferences, in stark contrast with the original AAC. Moreover, highly expressed genes do not exhibit more prominent environment-related AAC signatures than lowly expressed genes, despite contributing more to the effective proteome. Thus, evolutionary shifts in overall AAC appear to occur almost exclusively through factors shaping the global oligonucleotide content of the genome. We discuss these results in light of contravening evidence from biophysical data and further reading frame-specific analyses that suggest that adaptation takes place at the protein level.

Key words: amino acid composition, oligonucleotide composition, intergenic DNA, ecological preferences, prokaryotic genome, fungal genome, support vector regression.

Introduction

Amino acid composition (AAC) differs widely among microbial proteomes. Indeed, compositional differences are sufficiently pronounced that they were already noted by biochemical studies in the pregenomic era (Stokes and Gunness 1946; Freeland and Gale 1947) and allow discrimination of major taxonomic groups (Smole et al. 2011). In addition, differences in AAC can be used to predict whether organisms inhabit different ecological niches (Zeldovich et al. 2007; Smole et al. 2011), with characteristic compositional signatures associated with, for example, thermophilic (Tekaia and Yeramian 2006; Zeldovich et al. 2007) and pathogenic (Vidovic et al. 2014) lifestyles. This suggests that proteome-wide differences in AAC reflect not just historical contingencies but may to some extent constitute adaptive responses to specific ecological niches. In line with this notion, mechanistic explanations have been advanced for why the elevated or reduced abundance of certain amino acids might be beneficial for protein structure and function in a particular environment (Greaves and Warwicker 2007; Graziano and Merlino 2014; Vidovic et al. 2014). For instance, halophiles have an abundance of negatively charged residues on the protein surface, presumably to disfavor misfolded conformations through repulsive interactions (Graziano and Merlino 2014), and the proteomes of pathogenic bacteria avoid secondary structure-destabilizing amino acids, thus being protected from oxidative stress inflicted by the host defenses (Vidovic et al. 2014).

Nucleotide composition too varies widely across microbial genomes, especially across prokaryotes (Rocha and Feil 2010). This variation is not only evident in coding but also intergenic DNA, suggesting that compositional heterogeneity is not a trivial consequence of differential amino acid usage (Karlin 1998; Hershberg and Petrov 2010). Directional mutation pressures (mutational biases) are thought to be a significant contribution to such compositional heterogeneity within and between genomes (Sueoka 1988; Nekrutenko and Li 2000). However, both G + C content and more complex measures of oligonucleotide composition have been examined across genomes (Deschavanne et al. 1999) and—echoing findings from AAC—were found to carry a substantial environmental signal (Willner et al. 2009). This suggested that a specific composition might make nucleic acid molecules better suited to withstand high temperature, salinity, oxygen or other challenges, prompting further mechanistic models. For example, an increase in ApG dinucleotides with growth temperature (Zeldovich et al. 2007) may be the consequence of selection to strengthen nucleobase stacking interactions. Similarly, purine loading may contribute to thermal adaptation of mRNAs (Lambros et al. 2003).

The presence of selective pressures at the DNA/RNA level that might plausibly lead to skewed oligonucleotide composition complicates the interpretation of AAC signals associated with the environmental niche of organisms. With both nucleotide-level and amino acid-level selection acting on coding sequences, any ecologically predictive AAC signal may reflect selection at the amino acid level, selection at the nucleotide level or a mixture of the two—or indeed mutational processes symptomatic of a particular environment (Gu et al. 1998). Previous observations that noncoding G + C composition is predictive of coding sequence composition (Muto and Osawa 1987; Hershberg and Petrov 2010) further suggest that AAC signals may not exclusively spotlight constraints on protein biophysics but constitute composite signals that, at least in part, reflect DNA-level processes. However, the precise quantitative nature of this relationship, the relative merits of oligonucleotide composition and AAC as predictors of microbial ecology, and—ultimately—the relative importance of DNA-level and protein-level adaptations to different environments remain largely uncharacterized.

Here, in an effort to disentangle nucleotide- and amino acid-level contributions to ecological adaptations, we assess the correspondence between oligonucleotide composition in intergenic DNA (where there are no protein-related selective constraints), AAC and microbial ecology in a systematic, quantitative way using nonlinear support vector machine (SVM) regression. First, we highlight that simple G + C content variation—widely used to examine dependencies between genome and proteome composition (Singer and Hickey 2000; Bohlin et al. 2013)—lacks sufficient degrees of freedom to capture AAC and ecological diversity and can therefore lead to misleading conclusions about the true correspondence between AAC and nucleotide composition. We then demonstrate that joint consideration of mono-, di-, and trinucleotide composition of intergenic DNA yields an excellent predictor of AAC, explaining almost all (~91%) of AAC variability across prokaryotes when phylogenetic inertia is taken into account. Importantly, we find that AAC is a much poorer predictor of ecology once oligonucleotide composition (and phylogeny) has been controlled for. Intuitively, this might be taken to suggest that ecologically informative AAC signatures predominantly reflect selection at the nucleotide rather than amino acid level. However, reading frame-specific analyses continue to support selection at the amino acid level, highlighting a complex relationship between ecology and global shifts in nucleotide and AAC.

Materials and Methods

Data Collection

We downloaded 909 prokaryotic genomes (825 bacteria and 84 archaea) from the National Center for Biotechnology Information (NCBI) Genomes database. After excluding plasmids and chromosomes with less than 200 kb, the remaining 1,119 chromosomes/plasmids were used as instances in the regression analyses predicting the AAC of proteomes. For the classification task of recognizing environmental preferences, we collected data from the NCBI Genome Projects "*lproks0*" table.

Additional 600 genomes used as a test set are draft (incomplete) genomes downloaded from the NCBI Genomes database. As these genomes were not assembled and thus information on chromosomes/plasmids was not supplied, each genome here corresponds to one instance in the regression.

Eukaryotic genomes were collected from the website of the Genozymes project (Berka et al. 2011) and the NCBI Genomes database (49 organisms in total). Here, each organism was considered as one instance in the regression analysis. We labeled 13 organisms as thermophiles, and the remaining 36 organisms as nonthermophiles. These organisms were labeled manually by collecting information from different biological data sources.

In order to examine highly and lowly expressed genes separately, we used previously compiled data for 911 prokaryotic genomes (Krisko et al. 2014), where a statistical test was used to assign a binary high/low expression label to genes (Supek et al. 2010) based on similarity of their codon usages to a reference set of known highly expressed genes (ribosomal protein genes, chaperones, and translation factors). Due to the smaller number of highly expressed genes, we performed a rarefaction procedure where the same number of amino acid sites was sampled from the lowly expressed gene set as the number of amino acids available for the highly expressed genes on a given chromosome.

Lists of putatively horizontally transferred genes were obtained from the Horizontal gene transfer database (HGT-DB) (Garcia-Vallve et al. 2003). As information about horizontally transferred segments is available only for a subset of our initial data set, for this analysis we used 316 genomes, which resulted in 393 learning examples.

Regression Analysis

Each plasmid and chromosome with more than 200 kb was considered as a learning example for the regression task of predicting the AAC of prokaryotic proteomes. In the analysis of eukaryotic organisms, each organism was considered as one learning example. We sequentially introduced four different sets of features: 1) genomic G + C, 2) dinucleotide composition, 3) trinucleotide composition and 4) phylogenetic labels, and trained the regression models on sets (1), (1 + 2), (1 + 2 + 3), and (1 + 2 + 3 + 4).

The oligonucleotide frequencies were calculated only from the regions in the genomes that were not annotated as protein-coding regions, nor as RNA genes. Furthermore, we excluded 20 nucleotides upstream of the gene start codons, known to be under selective pressures due to translation initiation signals (Molina and Nimwegen 2008). The dinucleotide and trinucleotide frequencies were normalized to observed/expected ratios (O/E) by dividing by the product of the corresponding mononucleotide frequencies found from G + C content. Therefore, the features we employ throughout our analyses carry information orthogonal to that contained in G + C. In practice, this normalization to (O/E) has little effect on the outcome of regression when the G + C is used together with the di/trinucleotide features (supplementary fig. S8, Supplementary Material online), as expected from the ability of SVM to handle feature interactions relevant for the target variable.

The oligonucleotide frequencies were determined strand-symmetrically: For each oligonucleotide, we summed its frequency with a frequency of its reverse complement, resulting in 10 features for the dinucleotides and 32 for the trinucleotide composition. Phylogeny was encoded as the set of 188 binary features indicating phylum-, order-, and class-level membership of the organisms in the data set. The dependent variable in our regression model was the frequency of a single amino acid, and separate regression models were fit for each amino acid. To measure accuracy of models, we recorded average and standard deviation over ten runs of 10-fold cross-validation for the coefficient of determination (R^2) and RMSE. Due to the small number of learning instances, we used leave-one-out cross-validation when performing experiments for eukaryotic organisms. For prokaryotic genomes, model performance was also tested on a separate test set.

To test whether phylogenetic relatedness of our data instances artificially inflates model performance reported by cross-validation, we reduced the original data set by allowing only one instance (randomly chosen) for each species. This resulted in a diversified data set with 480 species, which we compared with a same-size data set of randomly chosen instances, i.e., generated without taking into consideration the phylogeny. A comparison was performed using 10-fold cross-validation, as well as on a separate test set consisting of the 639 excluded instances. Due to the random choice of organisms, we repeated this process three times and recorded mean R^2 and RMSE value for each amino acid.

SVM regression was performed using the LibSVM library (Chang and Lin 2011). We used epsilon-SVR implementation with the epsilon parameter in loss function of 0.001 and the radial basis function (Gaussian) kernel. The use of the kernel trick enables SVMs to map the data into high-dimensional space and very efficiently perform nonlinear classification and regression (Ben-Hur et al. 2008). The regularization parameters C and gamma were optimized using a grid search ($C = 2^{-5}, 2^{-4}, \ldots, 2^{10}, \gamma = 2^{-15}, 2^{-14}, \ldots, 2^{5}$) in increments of R^2 in five runs of 4-fold cross-validation, per recommendation of LibSVM authors (Hsu et al. 2010). We normalized all feature values to the unit interval; all other parameters were set to their default values. The same algorithm was used for performing regression analysis separately for highly and lowly expressed genes.

As an alternative nonlinear regression method to the SVM, we also considered Random Forest regression. In particular, we employed "predictive clustering trees" (PCTs) (Blockeel et al. 1998), a generalization of standard decision trees, where leaves correspond to clusters and the tree can be viewed as a hierarchy of clusters. PCTs have been successfully applied to multitarget prediction tasks (both regression and classification), such as hierarchical classification of gene functions (Schietgat et al. 2010; Škunca et al. 2013) and gene expression time series analysis (Slavkov et al. 2010). Here, the use of PCTs enabled us to predict all amino acid frequencies simultaneously. To fit PCTs, we used system CLUS version 2.12 (freely available at http://dtai.cs.kuleuven.be/clus/, last accessed May 21, 2015) in a Random Forests setting, where an ensemble of trees is used to increase predictive performance (Breiman 2001). We used 1,000 unpruned trees, and used variance reduction as a heuristic to select splits.

In addition to the SVM and Random Forests, we also performed an experiment with a simple linear regression method, Ordinary Least Squares. We reduced set of features using a greedy backward feature elimination and AIC as a model performance measure. This experiment was performed using the Weka library.

Prediction of Environmental Preferences

Predictions of the amino acid frequencies obtained from the SVM regression were then used to find residuals defined as the difference between the observed and predicted amino acid frequencies. As a result, each genome was described

with 20 residuals, one per each amino acid. These residuals represent the variance of the AAC not explained by the oligonucleotide frequencies and phylogeny and they were used to classify organisms according to their environmental preferences. Thus, the features in our learning set were 20 amino acid residuals and the dependent variable was 1 if the organism lives in the particular environmental niche, and 0 otherwise. To observe the difference in environmental preferences prediction between oligonucleotide-phylogeny-normalized amino acid frequencies and true amino acid frequencies, we created another learning set where features were original amino acid frequencies. We also predicted environments directly from the oligonucleotide composition of noncoding DNA with genomic G + C, dinucleotide and trinucleotide frequencies and phylogenetic categories encoded as features.

In order to classify organisms according to the environmental niche, we used the SVM classifier implemented in LibSVM library (Chang and Lin 2011). We employed a C-SVC with a Gaussian kernel, where the *C* and *gamma* parameters were again optimized using a grid search as described above, all feature values were normalized to the unit interval and with probability estimates parameter set, whereas other parameters were set to their default values. We recorded the average value and standard deviation of the AUROC score over ten runs of 10-fold cross-validation. Predictions of the SVM classifier resulting from a single run of 10-fold cross-validation were used to visualize ROC curves. When considering highly and lowly expressed genes separately, the same algorithm was used and the same procedure repeated.

The complete data set is given in supplementary table S1, Supplementary Material online. It consists of i) nucleotide frequencies in intergenic DNA of all examined genomes, ii) amino acid frequencies of the corresponding proteomes, iii) environmental preferences of the organisms, and iv) the AAC residuals from SVM regression.

Pseudogenes Detection

In order to remove potential pseudogenes from the intergenic regions, we detected all ORFs with length greater or equal than a set length. The threshold was set to 30 codons, which corresponds to a false positive rate of 0.25 at the G + C content of 0.514 (median of the G + C content in our data set) (Pohl et al. 2012). We created a new data set in which oligonucleotide frequencies were calculated from the intergenic regions, after having excluded the parts which were identified as potential pseudogenes.

Bootstrapping Analysis

To obtain the bias-corrected estimate of amino acids' explained variance, we performed a bootstrap adjustment. For each amino acid, data points (bacterial chromosomes) were resampled using the Weka library. This bootstrap was repeated 100 times, and each time a SVM regression model was trained on the unique subset of the resampled data, containing on average approximately 63% of the total number of instances. Parameters *C* and *gamma* for SVM were kept as determined for the original data set (for a particular amino acid). Ten-fold cross-validation was used to calculate bootstrap estimates of coefficient of determination (R^2). The bias in R^2 was then estimated to be the difference between the mean of the bootstrap estimates and R^2 calculated from the original data set.

Direction of the Environment-Associated Change in DNA Composition

We separately analyzed different positions in coding DNA and calculated the G + C and dinucleotide frequencies for each of three codon positions. Under "first codon position," we assume the first and the second nucleotide in the codon, the "second position" are the second and the third nucleotides, whereas the "third position" are the third nucleotide and the first one in the next codon. For each environment and each dinucleotide frequency we determined the Mann–Whitney statistic separately (using *R*), and normalized it to the readily interpretable AUROC score by dividing with the product of the sample sizes for the two classes. The analyses in figure 6 implicitly account for phylogenetic relatedness, as the first sites are compared with second sites (and 2nd vs. 3rd, and 3rd vs. 1st) in the exact same set of genomes. In other words, if a high AUROC score is purely due to phylogenetic signal confounded with the environmental labels, it should be equally so at all codon sites, and no significant difference in AUROC scores will be found.

Other Statistical Analyses

Statistical analyses were performed using R and MATLAB. Differences between ROC curves were calculated using the DeLong method (DeLong et al. 1988) implemented in package pROC for R. In order to correct for multiple tests, all *P* values were adjusted using Benjamini–Hochberg method with false discovery rate (FDR) $\leq$ 10%. The pROC package was also used to calculate 95% CI of ROC curves, using the default setting of 2,000 stratified bootstrap replicates. Principal component analyses were performed using MATLAB R2013a.

Results

Composition of Noncoding DNA Almost Fully Explains the AAC of Proteomes

We examined the genome-encoded protein sequences from 909 bacterial and archaeal genomes, where each organism was represented by the relative frequencies of 20 amino acids in the complete set of proteins. Then, for each amino acid, we predicted the change in its relative frequency across genomes from the composition of intergenic DNA of these genomes using nonlinear SVM regression, and evaluated the fit using

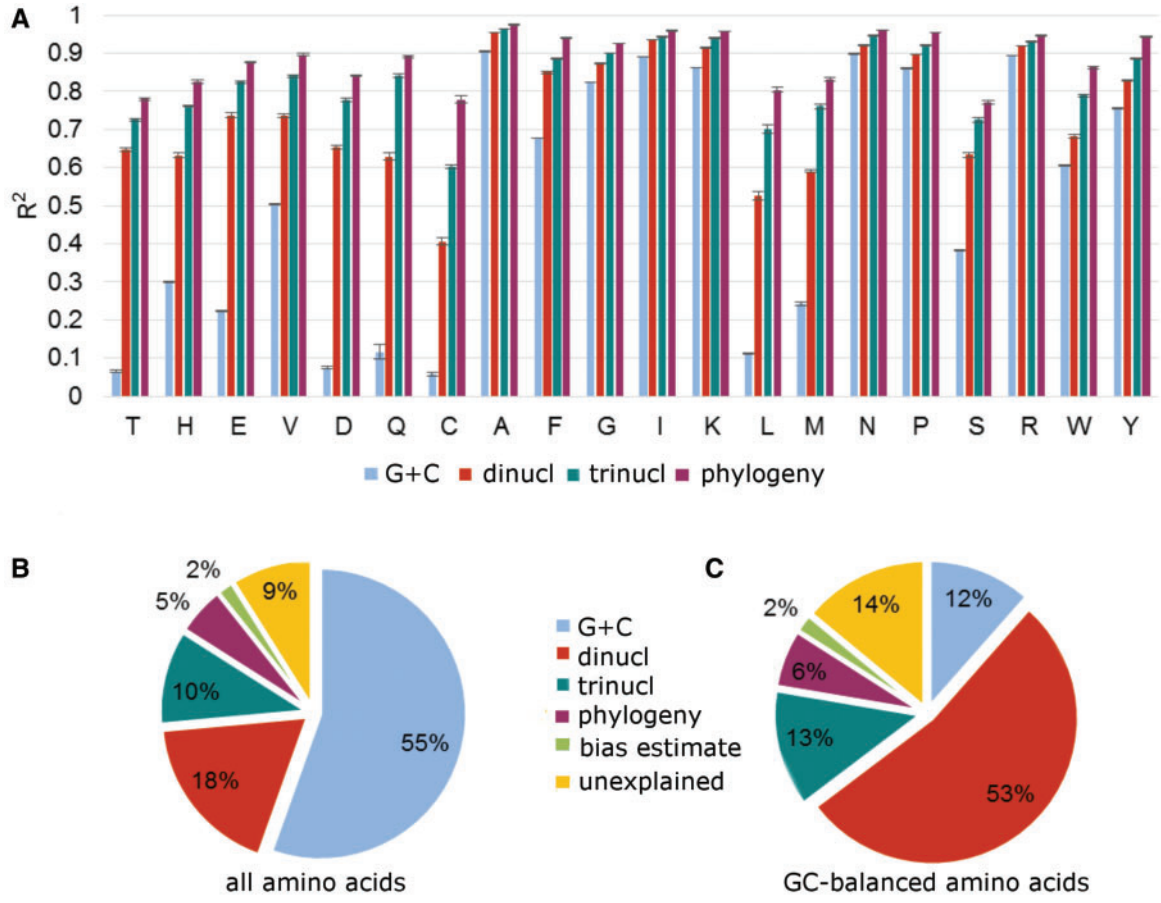

Fig. 1.—The oligonucleotide frequencies in the noncoding DNA of prokaryotes are highly predictive of their proteome compositions. (*A*) Explained variance (as squared Pearson correlation coefficient, R^2) in the amino acid usage of proteomes in a multiple regression against different sets of features; by considering only the G + C content (blue bars), and by progressively including also the dinucleotide frequencies (red), the trinucleotides (teal), and phylogenetic groups (purple). Error bars are standard deviations from ten runs of cross-validation. (*B*, *C*) The median variance explained using the same sets of features over all 20 amino acids (*B*) or only over the seven G + C balanced amino acids (THEVDQC) (*C*). The "bias estimate" is from bootstrapping (Materials and Methods).

cross-validation (see Materials and Methods). Intergenic DNA was defined as the sequence not annotated as harboring an RNA or protein-coding gene.

Consistent with previous work (Singer and Hickey 2000; Lightfield et al. 2011), we find that G + C content alone can explain some of the AAC variation between genomes (fig. 1; median R^2 over amino acids = 0.555) but leaves a substantial fraction of variance unexplained. This is not surprising as G + C variation has a single degree of freedom, insufficient to capture the diversity in AAC (and ecological preferences) among microbes, as illustrated by the seven amino acids with balanced G + C across codons (THEVDQC): AAC for this subset of amino acids is poorly predictable from G + C alone (fig. 1; median $R^2 = 0.115$). In more general terms, we estimate that our data set has at least 6 and 7 degrees of freedom for the AAC and ecological preference, respectively (supplementary fig. S1, Supplementary Material online). This is important to note because in cases where AAC correlates with ecological parameters, but G + C does not—such as for thermophilicity (Hurst and Merchant 2001; Zeldovich et al. 2007) and halophilicity (Paul et al. 2008)—this should not be taken as sufficient evidence for adaptation at the amino acid level. Rather, absence of a clear association might reflect intrinsic limitations of G + C content as a predictor.

Introducing the relative frequencies of dinucleotides in intergenic DNA (Materials and Methods) in addition to G + C content considerably improves AAC prediction, for both the G + C balanced amino acids (median $R^2 = 0.647$) and on overall (fig. 1; $R^2 = 0.736$, 0.632–0.879 [median, Q1–Q3 over amino acids]). Observed dinucleotide frequencies were normalized by the frequency expected from G + C content in order to capture orthogonal information (Materials and Methods). We gain further predictive accuracy by adding trinucleotide composition as a predictor (fig. 1; $R^2 = 0.840$,

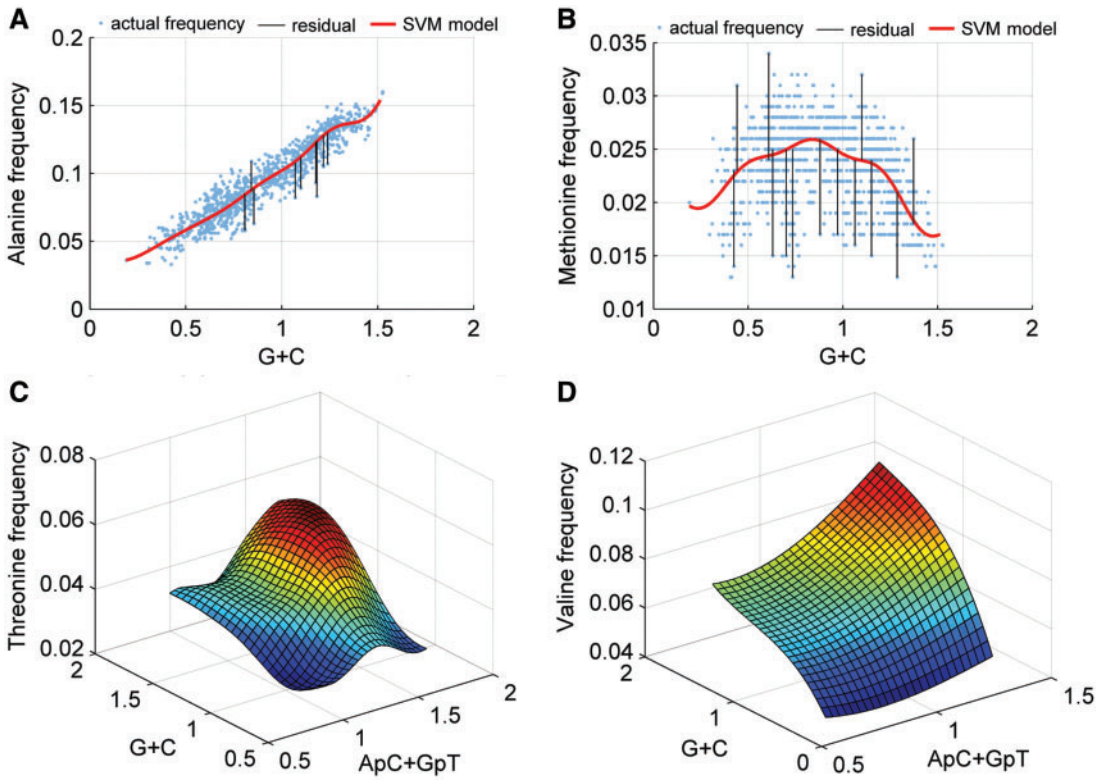

Fig. 2.—Nonlinear SVM regression models that predict amino acid usage in proteomes from G + C and dinucleotide frequencies in noncoding DNA. Dependency of relative frequencies of Ala (*A*) and Met (*B*) in proteomes on the G + C content of DNA, as examples of a linear and nonlinear relationship, respectively. Each dot is a prokaryotic chromosome (>200 kb in size). Red curves show SVM predictions. Several examples which deviate strongly from the dominant trend are highlighted by the vertical lines that show residuals of the regression. SVM regression models that regress the relative frequency of Thr (*C*) and Val (*D*) in proteomes against a combination of the G + C content and the frequency of the ApC + GpT dinucleotide.

0.761–0.905). Testing this regression model on an additional set of 600 genomes yielded similar results (median $R^2 = 0.820$; supplementary fig. S9C, Supplementary Material online). Of note, in some cases the regression models involve complex interactions between features. For instance, in valine and threonine, a combination of G + C content and ApC/GpT dinucleotide frequencies exhibits nonadditive effects in determining the frequency of the amino acids (fig. 2*C* and *D*). A simple linear regression model where the number of free parameters was further penalized (by Akaike information criterion [AIC], see Materials and Methods) still yields a median cross-validation $R^2 = 0.728$ (supplementary fig. S11, Supplementary Material online), suggesting that our SVM estimates of fit are not inflated as a result of overfitting.

We also considered the possibility that, if unannotated pseudogenes were a major contributor to intergenic compositional biases, these biases might simply reflect past selection operating at the amino acid level. However excluding all uninterrupted open reading frames (ORFs) of at least 30 codons from the intergenic DNA, and thus reducing potential contamination from recently pseudogenized genes, has very little impact on the ability of intergenic composition to predict AAC (supplementary fig. S2, Supplementary Material online). Similarly, we found no change in predictive power when excluding genomic segments suspected to be derived from horizontal gene transfer (supplementary fig. S12, Supplementary Material online). Finally, we examined to what extent the above estimates of model fit could be biased due to use of cross-validation on phylogenetically related (and thus not fully independent) points. By excluding multiple species from the same genus or the same family, we found that only a very small fraction of the variance explained (~0.02–0.04) might conceivably be due to phylogenetic nonindependence (Materials and Methods; supplementary fig. S9A-B, Supplementary Material online).

Accounting for Phylogenetic Inertia Leaves Little Variability in AAC Unexplained

Both the genomic oligonucleotide usage (Pride et al. 2003) and the AAC of the proteome are known to display phylogenetic inertia (Bohlin et al. 2013), contributing to the observed variability between organisms. For instance, it is known that Bacteria can be accurately separated from Archaea based on

AAC (Smole et al. 2011). Thus, part of the unexplained variance in AAC might be due to phylogenetic dependencies rather than, for example, amino acid level selection that is decoupled from oligonucleotide composition. To control for this factor, we introduce phylum-, order-, and class-level labels to the set of features used in the regression (Materials and Methods). Doing so further increases predictive accuracy ($R^2 = 0.893$, 0.830–0.944 [median, Q1–Q3]; fig. 1). We verified that these findings are not dependent on a particular regression method, being broadly similar for Random Forests regression (supplementary fig. S3*A*, Supplementary Material online).

The regression model that predicted variability in AAC across genomes most accurately was the one that employed G + C content, dinucleotide and trinucleotide composition, and the phylogenetic categories as features, explaining a remarkable 91% of AAC variability across genomes (fig. 1; bootstrap-adjusted median R^2 over 20 amino acids = 0.911). Furthermore, the amino acids with some residual unexplained variance are either rare in proteomes (supplementary fig. S3*B*, Supplementary Material online; Cys 22%, His 18%, Trp 14% unexplained) or change little in frequency across genomes (supplementary fig. S3*B*, Supplementary Material online; Asp 16%, Ser 23%, Thr 22% unexplained), suggesting that the unexplained variance is, at least in part, attributable to noise due to small sample sizes. A bootstrapping analysis supports this notion (supplementary fig. S3*C*–*F*, Supplementary Material online).

AAC Is Poorly Predictive of Microbial Ecology upon Factoring Out Background Nucleotide Composition

The large fraction of AAC variability explained by variation in oligonucleotide composition suggests that ecological AAC signatures might (perhaps even predominantly) originate from the nucleotide level. We therefore asked to what extent can AAC be used to discriminate organisms by ecological niche, before and after controlling for oligonucleotide composition. The residuals of the best regression model (see above) represent the variance in AAC that was not explained by oligonucleotide frequencies, and can be thought of as DNA composition-normalized amino acid frequencies (examples in fig. 2*A* and *B*).

Previous work demonstrates that AAC can separate thermophilic from mesophilic organisms with very high accuracy (Zeldovich et al. 2007; Smole et al. 2011), a finding replicated by our SVM classifier when considering the area under the receiver operating characteristic (ROC) curve (AUROC) as a measure of classification accuracy (fig. 3*C*; AUROC = 0.990). The AUROC expresses the probability that, in a randomly drawn thermophile–mesophile pair of microbes, the thermophile will be correctly recognized, with a value of 0.5 indicating random guessing. In contrast to the very high classification accuracy obtained when considering AAC prior to nucleotide normalization, we find that AAC residuals could accomplish the thermophile recognition task with a much lower success (AUROC = 0.738; fig. 3*C*). This suggests that a substantial component of the thermal AAC signature is grounded in oligonucleotide content, as becomes evident when comparing the distributions of the AAC residuals of thermophiles and mesophiles, alongside the raw AAC of both groups (fig. 3*A*). We obtain similar results when we try to discriminate halophiles from nonhalophiles (supplementary fig. S4*A*, Supplementary Material online; AAC AUROC = 0.968, AAC residual AUROC = 0.678), or aerotolerant from obligate anaerobe organisms (fig. 3*D*; 0.958 vs. 0.715), or similarly for obligately aerobic, host-associated, soil-dwelling, psychrophilic or radioresistant microbes (supplementary fig. S4*B*–*F*, Supplementary Material online). Consistently, the environment can be predicted from genomic oligonucleotide frequencies of intergenic DNA nearly as accurately as it can be from the AAC of the proteomes (fig. 3 and supplementary fig. S4, Supplementary Material online). This suggests that the contribution to raw AAC signatures made by variation that exclusively pertains to the amino acid level is often limited, at least for the ecological parameters considered here. Of note, although the classification from AAC residuals was severely compromised in comparison to the actual AAC, the AUROC scores were still significantly above the baseline of 0.5 ($P < 0.001$ for all environments; fig. 3 and supplementary fig. S4, Supplementary Material online). Therefore, this analysis does not exclude selection on AAC in different environments, but implies that its signal is subtle when compared against the backdrop of the AAC changes dependent on oligonucleotide composition.

Oligonucleotide Composition Predicts AAC across Eukaryotes

Next, we examined the genomes and proteomes of 49 fungi, of which 13 were thermophilic. Results are broadly consistent with our findings in prokaryotes: The G + C content of noncoding DNA—here encompassing introns and intergenic regions—can explain 60% of the variability in AAC across fungi (fig. 4*A* and *B*). Incorporating di- and trinucleotide composition as features in the regression leads to enhanced predictive power ($R^2 = 0.73$), with the further addition of phylogenetic categories leading to 80% of variance in proteome composition being accounted for. As observed for prokaryotes, thermophilic fungi can be recognized with high accuracy from the AAC of their proteomes (AUROC = 0.940; fig. 4*C*), whereas prediction from AAC residuals after nucleotide composition is factored out is considerably less accurate (AUROC = 0.639; fig. 4*C*). These findings indicate that the putatively adaptive signatures in AAC emanate from the nucleotide level not only in prokaryotes but also in eukaryotes.

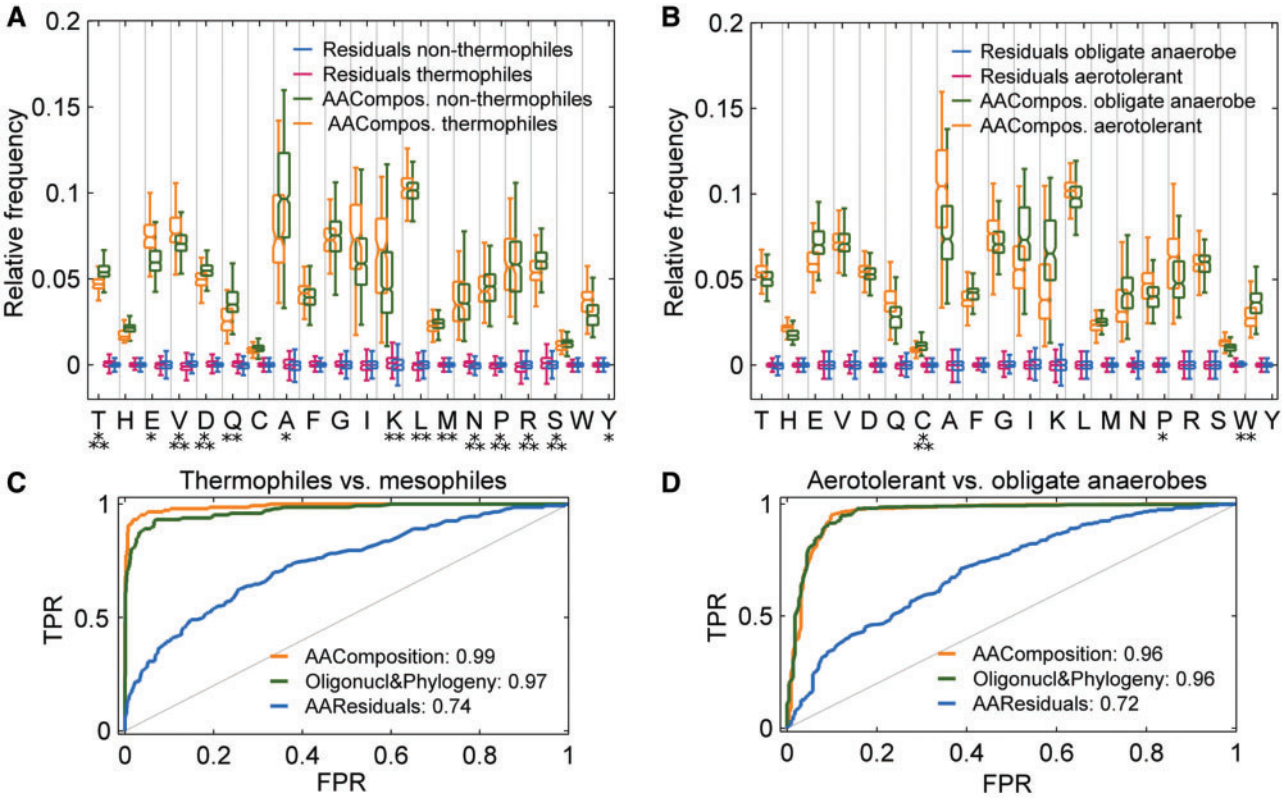

Fig. 3.—Accuracy in classifying prokaryotes by environmental preference from the AAC of proteomes and from oligonucleotide frequencies in noncoding DNA. (*A*, *B*) Distributions of AACs (given as relative frequencies of each amino acid) across proteomes, as well as the residuals of the amino acid composition in SVM regression. Asterisks are Mann–Whitney tests (two-tailed) applied to distributions of residuals. *FDR < 25%; **FDR < 10%; ***FDR < 1%. ROC curves for discriminating thermophiles from mesophiles (*C*) and strict anaerobes from aerotolerant organisms (*D*). Orange curves show predictions from AAC in proteomes, green curves from noncoding DNA (G + C content, di- and trinucleotide frequencies) and phylogenetic descriptors (clade memberships), and blue curves from AAC after a normalization for oligonucleotide frequencies in noncoding DNA and for phylogenetic relatedness (residuals from regression of AAC on these features). AUROC scores are given in plot legends, where 1.0 indicates perfect performance, and 0.5 random guessing (shown as the diagonal line). Predictions in the ROC curves are from an SVM classifier, in 10-fold cross-validation. TPR, true positive rate; FPR, false positive rate. More environments shown in supplementary figure S4, Supplementary Material online.

Highly Expressed Proteins Do not Exhibit More Prominent AAC Signatures

Thus far, we have shown that intergenic oligonucleotide composition is an excellent predictor of AAC and that controlling for nucleotide composition leads to a substantial drop-off in classifier performance. Intuitively, this might imply that a given ecological signal primarily emanates from the nucleotide level and that the AAC is, to a greater or lesser extent, an epiphenomenon that passively tracks nucleotide composition. To further consider the relative contributions of nucleotide versus amino acid level selection, we considered the predictive capacity of the AAC in light of gene expression levels. Selection at the amino acid level should be stronger in highly expressed genes, increasing its relative contribution to the composite AAC signature that reflects both nucleotide and amino acid level processes. Consequently, AAC should be harder to predict from intergenic DNA for highly expressed genes compared with lowly expressed genes. Expression levels of proteins in conditions favorable to growth can be approximated from codon biases in protein-coding genes (Ikemura 1985). To this end, we use previous data for 911 prokaryotic genomes (Krisko et al. 2014), where a statistical test was used to assign a binary high/low expression label to genes (Supek et al. 2010). Using highly and lowly expressed genes separately to predict AAC from oligonucleotide composition, we find no significant difference in prediction accuracy (fig. 5*A*; mean difference of root-mean-square error [RMSE] over 20 amino acids = 0.002%, 95% CI: [−0.016%, 0.020%]). This suggests that higher expression does not lead to a greater preponderance of amino acid-related signatures in the AAC signal. We explicitly test this by examining the predictive power of AAC residuals derived

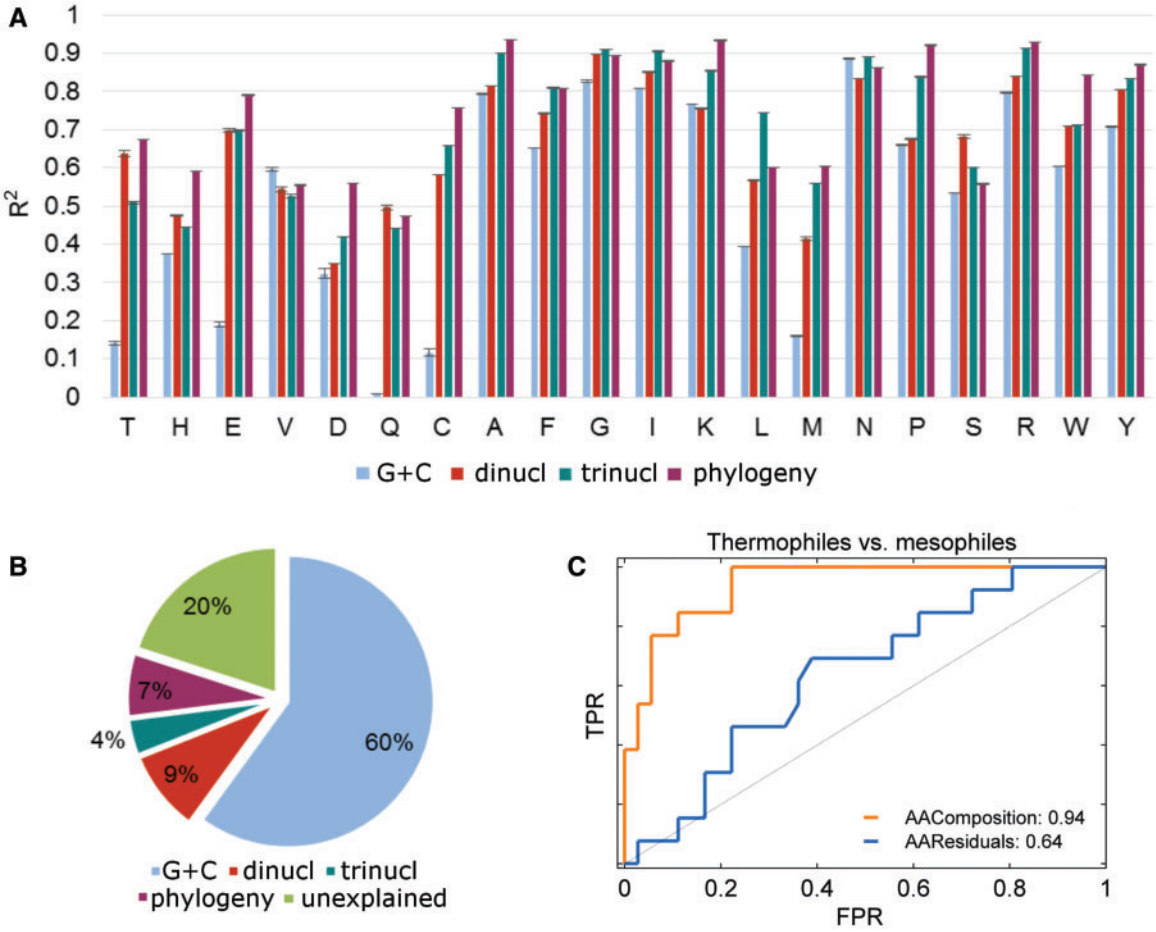

Fig. 4.—Composition of noncoding DNA in 49 fungal genomes is highly predictive of the corresponding proteome composition. (*A*) Explained variance (as squared Pearson correlation coefficient, R^2) in amino acid usage of proteomes in a multiple regression against different sets of features; obtained by considering only the G + C content (blue bars), and by progressively including also the dinucleotide frequencies (red), the trinucleotides (teal), and phylogenetic groups (purple). Error bars are standard deviations from ten runs of cross-validation. (*B*) The median variance explained using the same sets of features over all 20 amino acids. (*C*) Cross-validation ROC curves describing the accuracy of discrimination of 13 thermophilic fungi by their AAC (orange) or by the genome composition-normalized AAC (the "AAC residuals," blue). Inlaid numbers are AUROC scores.

from highly expressed genes for the organismal ecology and find that they are, overall, as poorly predictive as residuals derived from the remainder of the proteome, in contrast to the original AAC (fig. 5*B*). When examining individual environments separately, we again find no significant differences between the highly expressed genes and the rest of the proteome (at FDR < 10%; supplementary fig. S5, Supplementary Material online). This analysis is not affected by the phylogenetic relatedness of the points (organisms) in our regression data (supplementary fig. S10, Supplementary Material online).

A Reading Frame-Specific Analysis Provides Evidence for Selection on AAC

The above observations seem to indicate that AAC correlates with environmental preferences predominantly as a consequence of shifts in the global oligonucleotide frequencies. They, however, do not necessarily imply that AAC is not adaptive. For example, adaptive benefits might systematically correlate between nucleotide composition and the resulting AAC, leading us to underestimate the role of selection at the amino acid level. We therefore carried out further tests to establish whether changes in DNA composition can fully explain the observed variability in AAC.

One means to disentangle these two influences is to separately analyze coding nucleotides in different phases of the reading frame. If a particular change in the nucleotide composition is adaptive for the DNA/RNA, it should be so regardless of the reading frame. However, the same change will have different effects on AAC depending on how the affected sites are positioned in the coding sequence. For instance, an increase in ApG favors Ser and Arg if in the first/second codon position, but favors Gln, Glu and Lys if at the second/third

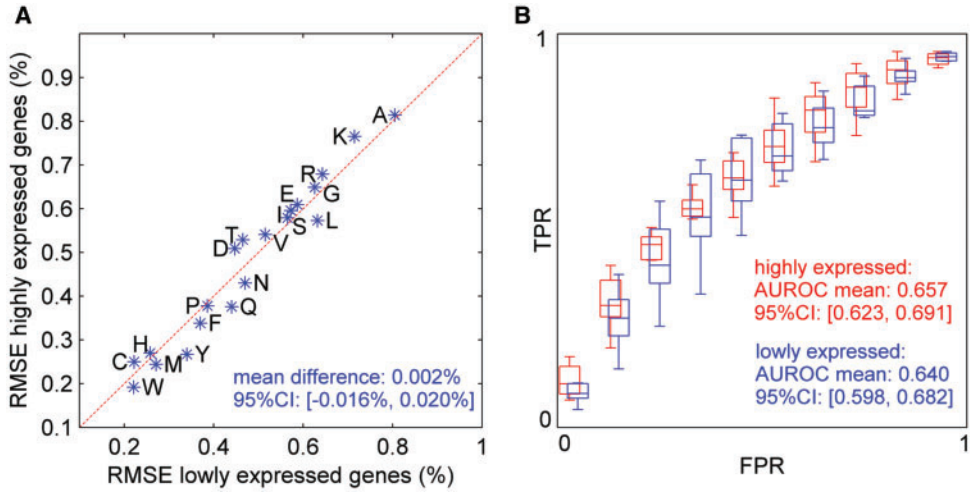

Fig. 5.—Lack of a particular environment-associated signal in the AAC of highly expressed proteins. (*A*) The RMSEs in predicting the frequencies of each amino acid from the composition of noncoding DNA (G + C, di- and trinucleotide content) and phylogenetic relatedness (clade membership) of organisms. RMSEs are compared for lowly versus highly expressed genes across all organisms. (*B*) Binned and pooled ROC curves for classifying the organisms by various environmental preferences from AAC, after having factored out the composition of noncoding DNA and phylogeny. ROC curves shown separately for classification only from highly expressed or only from lowly expressed genes. Full ROC curves for individual environments shown in supplementary figure S5, Supplementary Material online. Average and 95% CI of AUROC scores inlaid on plots.

codon position. Thus, if the codon positions exhibit different shifts in dinucleotide composition between, for example, thermophiles and mesophiles, this suggests such shifts are adaptive—at least in part—due to the associated changes in protein AAC.

Crucially, we would expect a very pronounced general difference in dinucleotide composition between the three codon positions due to universal protein structure constraints. We therefore compare only relative (rather than absolute) dinucleotide usage between, for example, thermophiles and mesophiles. Shifts in the distributions of dinucleotide frequencies were measured using a Mann–Whitney statistic, normalized to an AUROC score. Here, an AUROC of 0.5 signifies no shift in either direction, an AUROC less than 0.5 indicates lower frequencies, and an AUROC greater than 0.5 higher frequencies of a dinucleotide in one environment, all at a particular codon position. We find that for thermophiles, 5/11 tested AUROC scores are significantly different between the first and the second sites, 5/11 AUROC scores between the second and third, and 5 more AUROC scores between the first and third (fig. 6*A*; DeLong test, FDR $\leq$ 10%). One example of such position-specific changes is an increase in the frequency of TpA in thermophiles, characteristic for the first and second, but not the third codon position (fig. 6*A*), or a depletion for ApT which is mostly confined to the second codon position (fig. 6*A*). Similar comparisons of AUROC scores for halophiles (fig. 6*B*), obligate anaerobes (fig. 6*C*), psychrophiles (fig. 6*D*) and other niches (supplementary fig. S6, Supplementary Material online) also reveal pervasive codon position-specific dinucleotide shifts between environments.

Next, we visualize the distributions of selected dinucleotide frequencies of thermophilic and mesophilic protein-coding genes in all three codon positions (fig. 7). Here, the codon positions and can be compared qualitatively, in terms of direction and magnitude of change. Indeed, differences between the codon positions can be observed, where, for instance, the second codon position shifts toward higher GpA/TpC values in thermophiles, whereas this trend is reversed in the third codon position. These differences are not evident in randomized data (supplementary fig. S7, Supplementary Material online). Similar visualizations reveal significant differences between codon positions in dinucleotide frequency shifts between halophiles and non-halophiles (fig. 7*B*), strict anaerobes and aerotolerant organisms (fig. 7*C*), psychrophiles and non-psychrophiles (fig. 7*D*), and other niches (not shown).

The above reading frame-specific analysis suggests selection on AAC changes rather than solely on the nucleotide composition of DNA and/or RNA (discussed below).

Discussion

Microbial proteome composition is well known to vary between species, where such variation often reflects the ecological preferences of the organism. Learning if, and how, these changes in AAC help proteins to function in spite of various physical and chemical insults is paramount for our understanding of robustness and adaptability of life. The variation in G + C content across genomes is long known to strongly influence changes in amino acid usage of proteomes (Singer

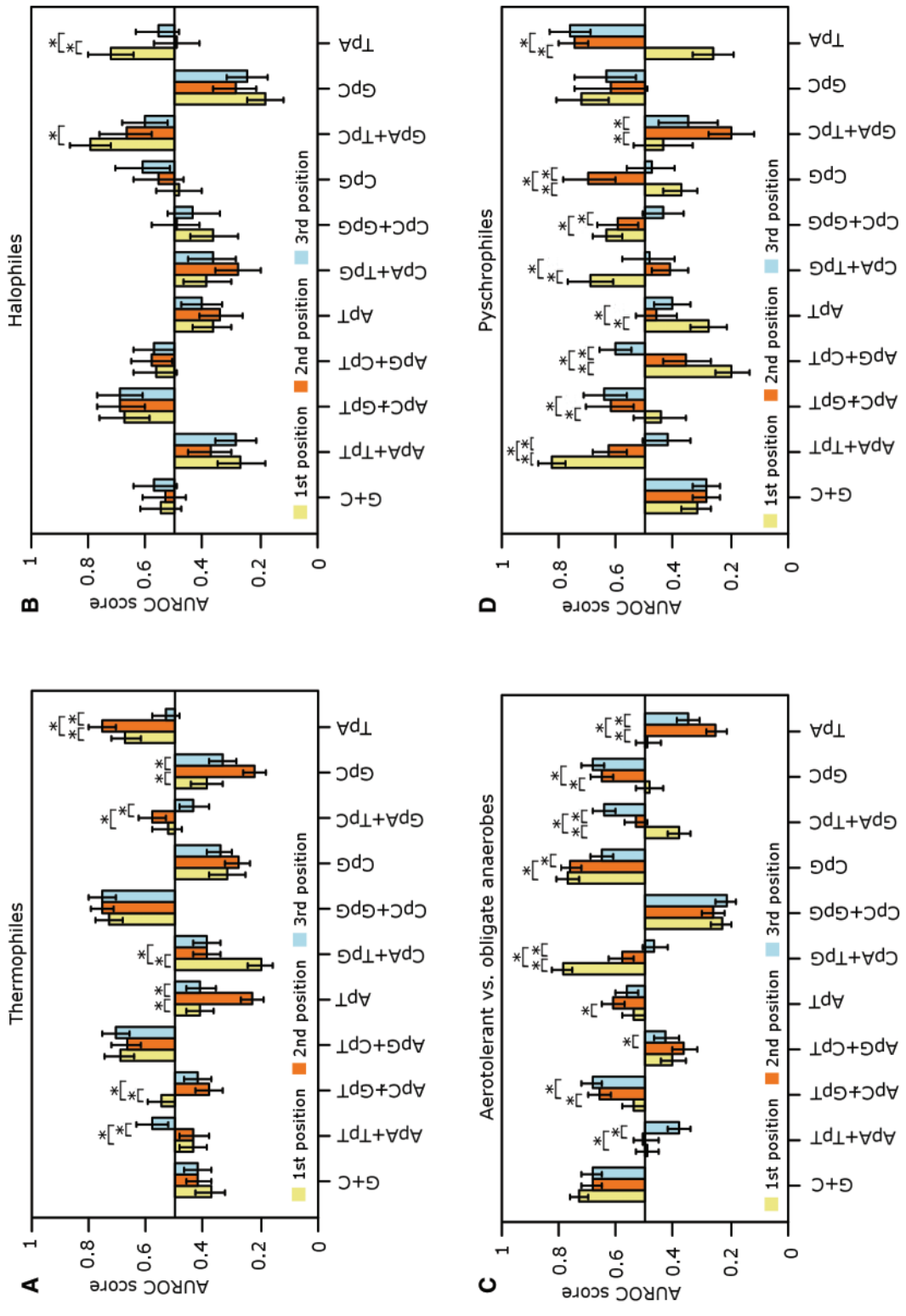

Fig. 6.—Differences in environment-specific trends in dinucleotide composition of 1st, 2nd, and 3rd codon sites in protein-coding genes. Shifts in G+C and dinucleotide frequencies between thermophilic and nonthermophilic (*A*), halophilic and nonhalophilic (*B*), strictly anaerobic and aerotolerant (*C*), and psychrophilic and nonpsychrophilic (*D*) organisms at different codon positions. Bars show AUROC scores, a measure of separability of two distributions by the given feature, where 0.5 signifies maximal overlap, and most extreme values (0 or 1) indicate complete separation of, for example, thermophiles and mesophiles by the frequency of given dinucleotide; values less than 0.5 and greater than 0.5 here indicate opposite directions of the shift. Error bars are 95% CI of the AUROC. Asterisks show significant differences in the environment-associated shifts between codon positions at less than 10% FDR.

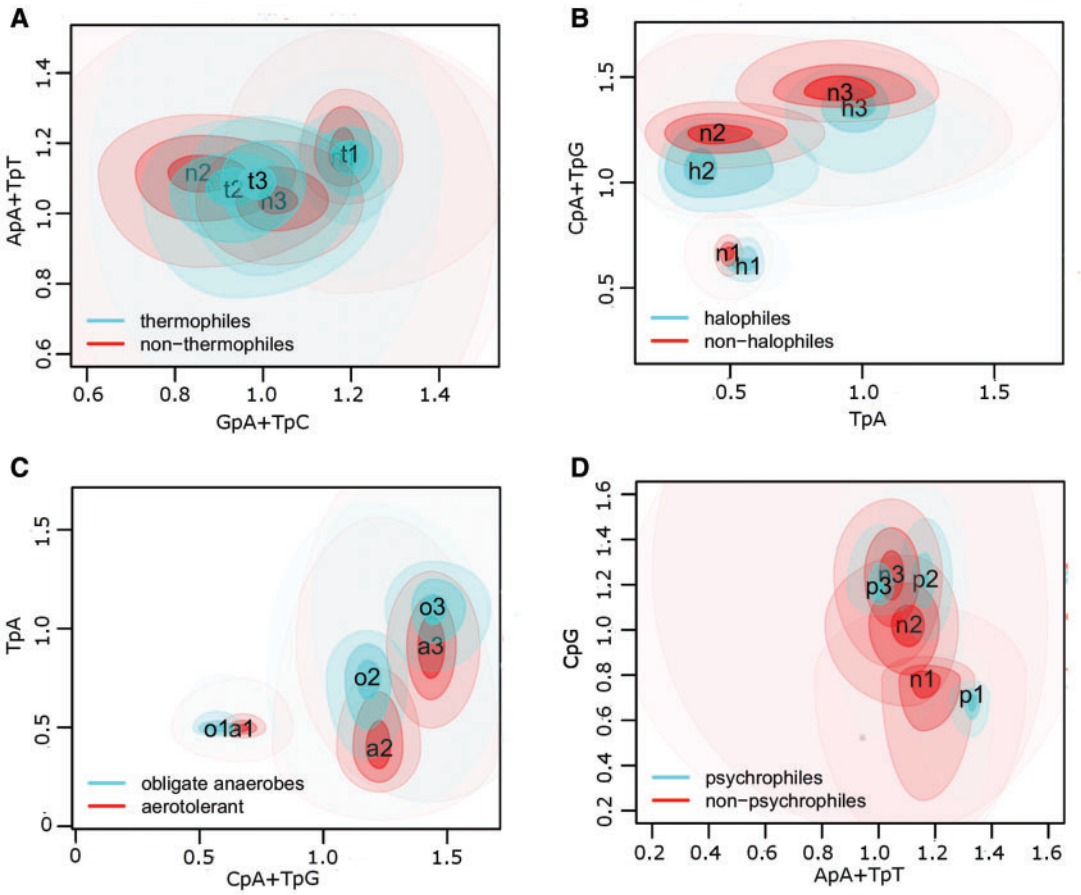

Fig. 7.—Distributions of selected dinucleotide frequencies at 1st, 2nd, and 3rd codon positions of protein-coding genes. (*A*–*D*) Ellipses show nine-number summaries of distributions, with borders indicating (in the increasing intensity of coloration) the minimum–maximum, 1st–7th octile, 2nd–6th octile, and 3rd–5th octile. Dinucleotide frequencies are normalized to the expected frequency given the G + C content. Plotted separately for thermophiles (*A*), halophiles (*B*), aerotolerant organisms (*C*), and psychrophiles (*D*). Letters in center of ellipse denote the environmental preference (t, thermophile; h, halophile; a, aerotolerant; p, psychrophile), and the number indicates the 1st, 2nd, or 3rd codon position this repeats.

and Hickey 2000; Moura et al. 2013), which might be adaptive in various environments (Rocha and Feil 2010). However, the G + C-associated trend does not extend to many amino acids (Lightfield et al. 2011) nor does it explain some prominent environmental AAC signatures (Zeldovich et al. 2007; Paul et al. 2008).

In addition to the G + C, the genomic dinucleotide usage is biased and varies between organisms sufficiently that it can be used to classify them from genome fragments; moreover, this variability was suggested to stem from changes in both intergenic and protein-coding regions (Karlin 1998). Here, we find that the compositional properties of the two parts of the genome are tightly coupled, and we quantitate this relationship. We find that DNA word frequencies in intergenic DNA have a surprising power to predict amino acid usage in prokaryotic and eukaryotic proteomes, up to the point where very little unexplained variability remains. A corollary is that the previously proposed adaptations of proteomes to environmental challenges (Greaves and Warwicker 2007; Graziano and Merlino 2014; Vidovic et al. 2014) may need to be reinterpreted, while taking into account the evolutionary forces shaping genomic DNA oligonucleotide frequencies. Consistently, after factoring out the influences of underlying DNA composition and of phylogenetic inertia from the AAC, it becomes considerably less predictive of the environmental preferences.

The key question then is whether the observed AAC changes are purely a secondary effect of the directional mutation pressures and/or adaptation of the DNA (or RNA) through oligonucleotide frequency shifts, while not necessarily being adaptive at the protein level. Most of the analyses above seem to suggest that, to a first approximation, this may hold true, providing a caveat to assigning adaptive significance to environment-related enrichments of certain amino acids

based on changes in AAC alone. However, our reading frame-specific analyses indicate that compositional shifts are in some cases sensitive to reading frame, a finding that is not expected under a nucleotide-level-selection-only model. Furthermore, the AAC is weakly but still significantly associated with environmental preferences even after factoring out the genomic oligonucleotide composition. Thus, there is an apparent evolutionary signature of amino acid-level selection specific to different environments, but it may be faint and easily overwhelmed by the signal of background nucleotide composition, which is very strongly reflected in the proteome. This might be explained by an amino acid-level selection that operates only on a smaller number of structurally important sites rather than on the general protein composition, thus having a quantitatively lesser (but still important) contribution to the compound AAC signal. This contribution may be larger, smaller, or absent depending on the amino acid and the environment. In certain instances, it may even have the opposite sign of the observed AAC difference (see examples for thermophiles and aerotolerant microbes in fig. 3*A* and *B*).

In addition to the above, it is also plausible that the selective forces known to operate on nucleotide composition of genomes (Hershberg and Petrov 2010; Hildebrand et al. 2010) are, at least in part, driven by their downstream effects on proteome composition (as determined by the genetic code). In other words, our overall data would be consistent with amino acid-level selection on proteomes if compositional adaptations at the nucleotide and amino acid levels were tightly coupled—as we observe—with compositional shifts at the former triggering adaptive benefits at the latter. Although we cannot currently resolve to what extent selection acts on nucleotide versus AAC levels, our results provide an important quantitative baseline for further assessment and suggest that claims of adaptive amino acid usage should be interpreted with caution if they are based solely on AAC compositional shifts.

Our principal finding that genomic word usage tightly constrains the spectrum of compositional variability between proteomes has further implications. An obvious consequence is that AAC variation has, in effect, less degrees of freedom than expected—it can vary only along a lower-dimensional manifold within the amino acid frequency space, largely determined by the genome-wide oligonucleotide frequencies and the genetic code. Also, given that di/trinucleotides cross codon borders, these constraints would likely affect also di-amino acid frequencies, imposing further limits on the structural landscape that can be explored by natural proteins. Accounting for the underlying DNA composition may thus have implications for development of protein evolutionary models, for remote homology search, or for protein structure prediction.

Furthermore, given that many amino acid changes in a protein sequence are governed simultaneously by a global factor—the DNA composition—the coevolution of pairs of amino acids, often interpreted as their functional linkage and/or spatial proximity in the protein structure, may sometimes not reflect either of those things. In other words, a certain baseline level of correlated changes of sites in protein alignments is to be expected due to the overarching effect of DNA composition on amino acid frequencies, and not due to epistasis. Consistently, when predicting protein structures from pairs of coevolving sites, a global method drawing on partial correlations was vastly superior to approaches that consider individual pairs separately (Marks et al. 2011; Hopf et al. 2012).

In practical terms, dinucleotide frequencies can be more precisely measured from a short stretch of DNA than the amino acid frequencies can be measured from the DNA's translation, and moreover, a gene finding step to determine the correct reading frame is not necessary. This opens up new possibilities for use of single short reads from environmental sequencing—without further assembly—to deduce the phenotypic traits of the various microbes in a metagenome (Willner et al. 2009). This could be particularly important for the multitude of rare taxa that contribute relatively few reads to the total sequencing output, but appear to make up the largest part of the species' tally in human microbiomes (Dethlefsen et al. 2008).

Supplementary Material

Supplementary table S1 and figures S1–S12 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Mediterranean Institute for Life Sciences and Fondation Nelia et Amadeo Barletta (to M.B. and A.K.), by MRC core funding and an Imperial College Junior Research Fellowship (to T.W.), by FP7 FET grant ICT-2013-612944 MAESTRA (to F.S. and M.B.), by FP7 REGPOT grant InnoMol (to F.S.), by the Croatian Science Foundation grant HRZZ-9623 (to M.B.), and the Croatian Ministry of Science and Sport grant 098-0000000-3168 (to F.S.).

Literature Cited

Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. 2008. Support vector machines and kernels for computational biology. PLoS Comput Biol. 4:e1000173.

Berka RM et al. 2011. Comparative genomic analysis of the thermophilic biomass-degrading fungi Myceliophthora thermophila and Thielavia terrestris. Nat Biotech. 29:922–927.

Blockeel H, Raedt LD, Ramon J. 1998. Top-down induction of clustering trees. In: Shavlik JW, editor. Proceedings of the Fifteenth International Conference on Machine Learning. ICML '98; Madison, Wisconsin. San Francisco (CA): Morgan Kaufmann Publishers Inc. p. 55–63. Available from: http://dl.acm.org/citation.cfm?id=645527.657456.

Bohlin J, Brynildsrud O, Vesth T, Skjerve E, Ussery DW. 2013. Amino acid usage is asymmetrically biased in AT- and GC-rich microbial genomes. PLoS One 8:e69878.

Breiman L. 2001. Random forests. Mach Learn. 45:5–32.

Chang C-C, Lin C-J. 2011. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol. 2:27.
DeLong ER, DeLong DM, Clarke-Pearson DL. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–845.
Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B. 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol Biol Evol. 16:1391–1399.
Dethlefsen L, Huse S, Sogin ML, Relman DA. 2008. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS Biol. 6:e280.
Freeland JC, Gale EF. 1947. The amino-acid composition of certain bacteria and yeasts. Biochem J. 41:135–138.
Garcia-Vallve S, Guzman E, Montero MA, Romeu A. 2003. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Res. 31:187–189.
Graziano G, Merlino A. 2014. Molecular bases of protein halotolerance. Biochim Biophys Acta. 1844:850–858.
Greaves RB, Warwicker J. 2007. Mechanisms for stabilisation and the maintenance of solubility in proteins from thermophiles. BMC Struct Biol. 7:18–40.
Gu X, Hewett-Emmett D, Li WH. 1998. Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. Genetica 102–103:383–391.
Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. 6:e1001115.
Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet. 6:e1001107.
Hopf TA, et al. 2012. Three-dimensional structures of membrane proteins from genomic sequencing. Cell 149:1607–1621.
Hsu CW, Chang CC, Lin CJ. 2010. A practical guide to support vector classification. Available from: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.
Hurst LD, Merchant AR. 2001. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc Biol Sci. 268:493–497.
Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 2:13–34.
Karlin S. 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. Curr Opin Microbiol. 1:598–610.
Krisko A, Copic T, Gabaldón T, Lehner B, Supek F. 2014. Inferring gene function from evolutionary change in signatures of translation efficiency. Genome Biol. 15:R44.
Lambros RJ, Mortimer JR, Forsdyke DR. 2003. Optimum growth temperature and the base composition of open reading frames in prokaryotes. Extremophiles 7:443–450.
Lightfield J, Fram NR, Ely B. 2011. Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. PLoS One 6:e17677.
Marks DS, et al. 2011. Protein 3D structure computed from evolutionary sequence variation. PLoS One 6:e28766.
Molina N, van Nimwegen E. 2008. Universal patterns of purifying selection at noncoding positions in bacteria. Genome Res. 18:148–160.
Moura A, Savageau MA, Alves R. 2013. Relative amino acid composition signatures of organisms and environments. PLoS One 8:e77319.
Muto A, Osawa S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci U S A. 84:166–169.
Nekrutenko A, Li W-H. 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res. 10:1986–1995.
Paul S, Bag SK, Das S, Harvill ET, Dutta C. 2008. Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. Genome Biol. 9:R70.
Pohl M, Theißen G, Schuster S. 2012. GC content dependency of open reading frame prediction via stop codon frequencies. Gene 511:441–446.
Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. 2003. Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Res. 13:145–158.
Rocha EPC, Feil EJ. 2010. Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria? PLoS Genet. 6:e1001104.
Schietgat L, et al. 2010. Predicting gene function using hierarchical multilabel decision tree ensembles. BMC Bioinformatics 11:2–15.
Singer GAC, Hickey DA. 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol. 17:1581–1588.
Škunca N, et al. 2013. Phyletic profiling with cliques of orthologs is enhanced by signatures of paralogy relationships. PLoS Comput Biol. 9:e1002852.
Slavkov I, Gjorgjioski V, Struyf J, Džeroski S. 2010. Finding explained groups of time-course gene expression profiles with predictive clustering trees. Mol Biosyst. 6:729–740.
Smole Z, et al. 2011. Proteome sequence features carry signatures of the environmental niche of prokaryotes. BMC Evol Biol. 11:26–35.
Stokes JL, Gunness M. 1946. The amino acid composition of microorganisms. J Bacteriol. 52:195–207.
Sueoka N. 1988. Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci U S A. 85:2653–2657.
Supek F, Skunca N, Repar J, Vlahovicek K, Smuc T. 2010. Translational selection is ubiquitous in prokaryotes. PLoS Genet. 6:e1001004.
Tekaia F, Yeramian E. 2006. Evolution of proteomes: fundamental signatures and global trends in amino acid compositions. BMC Genomics 7:307.
Vidovic A, Supek F, Nikolic A, Krisko A. 2014. Signatures of conformational stability and oxidation resistance in proteomes of pathogenic bacteria. Cell Rep. 7:1393–1400.
Willner D, Thurber RV, Rohwer F. 2009. Metagenomic signatures of 86 microbial and viral metagenomes. Environ Microbiol. 11:1752–1766.
Zeldovich KB, Berezovsky IN, Shakhnovich EI. 2007. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol. 3:e5.

Associate editor: Ruth Hershberg