Net Evolutionary Loss of Residue Polarity in Drosophilid Protein Cores Indicates Ongoing Optimization of Amino Acid Composition

Lev Y. Yampolsky[1,]*, Yuri I. Wolf[2], and Michael A. Bouzinier[3]

[1]Department of Biological Sciences, East Tennessee State University

[2]National Center for Biotechnology Information, NIH, Bethesda, Maryland

[3]InterSystems Corporation, Cambridge, Massachusetts

*Corresponding author: E-mail: yampolsk@etsu.edu.

Accepted: September 16, 2017

Abstract

Amino acid frequencies in proteins may not be at equilibrium. We consider two possible explanations for the nonzero net residue fluxes in drosophilid proteins. First, protein interiors may have a suboptimal residue composition and be under a selective pressure favoring stability, that is, leading to the loss of polar (and the gain of large) amino acids. One would then expect stronger net fluxes on the protein interior than at the exposed sites. Alternatively, if most of the polarity loss occurs at the exposed sites and the selective constraint on amino acid composition at such sites decreases over time, net loss of polarity may be neutral and caused by disproportionally high occurrence of polar residues at exposed, least constrained sites. We estimated net evolutionary fluxes of residue polarity and volume at sites with different solvent accessibility in conserved protein families from 12 species of *Drosophila*. Net loss of polarity, miniscule in magnitude, but consistent across all lineages, occurred at all sites except the most exposed ones, where net flux of polarity was close to zero or, in membrane proteins, even positive. At the intermediate solvent accessibility the net fluxes of polarity and volume were similar to neutral predictions, whereas much of the polarity loss not attributable to neutral expectations occurred at the buried sites. These observations are consistent with the hypothesis that residue composition in many proteins is structurally suboptimal and continues to evolve toward lower polarity in the protein interior, in particular in proteins with intracellular localization. The magnitude of polarity and volume changes was independent from the protein's evolutionary age, indicating that the approach to equilibrium has been slow or that no such single equilibrium exists.

Key words: solvent accessibility, residue polarity, residue volume, selection, stability, Drosophila.

Introduction

Amino acid composition of proteins evolves under strong selective constraints guarding thermodynamic stability of protein molecules (Pal et al. 2006; Bloom et al. 2007; Camps et al. 2007; Tokuriki et al. 2008; Wylie and Shakhnovich 2011; Serohijos and Shakhnovich 2014; Echave et al. 2016). Whether such constraints result mostly in stabilizing selection against missense mutations, or also in positive selection for further improving protein stability, is not known. Frequencies of amino acids in proteins are far from the evolutionary equilibrium (Zuckerkandl et al. 1971; Jordan et al. 2005; Misawa et al. 2008; Mannige et al. 2012). The slow approach to this hypothetical equilibrium manifests itself in net fluxes (losses or gains) of different amino acids when phylogenetically related proteins are compared. For example, a net loss of polar amino acids occurred during the last 70 Myr of protein evolution in drosophilids (Yampolsky and Bouzinier 2010). In contrast, on a larger scale, when deeper phylogenies covering major branches of prokaryotes and eukaryotes are considered, a net loss of hydrophobic amino acids is observed (Mannige 2014). Several hypotheses have been proposed to explain such net fluxes, most of them centered on the properties of the genetic code and mutational biases. Jordan et al. (2005) suggested that the fluxes were explained largely by amino acids' frequencies in genomes: common amino acids experienced net loss, whereas rare amino acids were gained. This

Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution 2017.

pattern was originally predicted by Zuckerkandl et al. (1971) within a much smaller data set. This may reflect the evolutionary history of the genetic code and therefore amino acid composition in ancient ancestral proteins, as rare amino acids whose frequencies are increasing, such as C, M, W appear to be late additions to the genetic code (Trifonov 2004). Alternatively, these amino acids may be rare due to the low frequency of their codons in random sequences, as the same three amino acids are among those coded by the lowest number of codons and therefore are innately rare in novel protein sequences recruited from noncoding regions. Finally, the net fluxes may reflect long-term selective pressure on amino acid substitutions. Mannige et al. (2012) observed that (gains to nonpolar C, M, and W notwithstanding) in the deep phylogenetic comparisons, a net loss of hydrophobicity is observed, interpreting this as evidence that ancient proteins were more hydrophobic than modern ones (Mannige 2014). This conclusion contrasts with the observation that on much smaller phylogenetic timescale of drosophilid divergence one can observe a net loss of polarity (Yampolsky and Bouzinier 2010), suggesting that these long-term selective preferences can be reversed over time.

The question of how much of the observed trends can be explained by current amino acid composition and mutational biases rather than long-term selective preferences is unclear. On the one hand, amino acid composition differences (including systematic differences in polarity) leading to clearly adaptive enhanced thermostability (Cambillau and Claverie 2000; Zeldovich et al. 2007) or tolerance to hypersaline conditions (Paul et al. 2008; Tadeo et al. 2009) have been very well documented. On the other hand, some of the shifts in amino acid composition may be partially explained by global changes in genomic CG content (Liu et al. 2010). Likewise, Misawa et al. (2008) suggested that many of the net fluxes observed in phylogenetic analysis could be explained by the asymmetry introduced by the CpG mutational bias. Indeed, because both amino acids whose codons contain CpG dinucleotides are polar, such a bias can play a role in the observed loss of polar amino acids, because codons that contain the CpG nucleotides are more likely to be lost than gained in organisms with DNA methylation at such sites.

If selective pressures that shape amino acid composition exist, they probably act differently on different parts of a protein's secondary and tertiary structure, most importantly at the interior and the exterior amino acid sites. Solvent accessibility is the major determinant of selective constraints on amino acid substitutions in general (Koshi and Goldstein 1997; Mirny and Shakhnovich 1997; Yang and Swanson 2002; Conant and Stadler 2009; Franzosa and Xia 2009; Echave et al. 2016), with interior sites being under much stronger stabilizing selection than the exposed ones. Membrane-bound proteins do not radically differ in this respect from water-soluble proteins, despite the different hydrophobicity of their environment (Franzosa et al. 2013), possibly with the exception of transmembrane domains of G protein-coupled receptors (Spielman and Wilke 2013). Two types of substitutions are particularly likely to be strongly selected against at the interior sites: those that increase polarity of the residue in the hydrophobic core and those that decrease its volume, creating an interior cavity, that is, reducing packing density (Echave et al. 2016). Introduction of too many polar residues at interior sites (Baldwin 2007; Garcia-Seisdedos et al. 2012; Koga et al. 2012) and creation of interior cavities (Kadonosono et al. 2003; Bueno et al. 2006; Maeno et al. 2015) are known to have a destabilizing effect on the protein structure and foldability. Conversely, at exposed sites, changes to small residues are well-tolerated and changes from nonpolar to polar amino acids can be energetically beneficial as they reduce nonburied hydrophobicity (Baldwin 2007). Therefore, that when a protein's amino acid composition is not at an equilibrium, net gains and losses of amino acids with different polarity and volume occur differently depending on solvent accessibility of the amino acid site. It should be also noted that selective preferences for amino acid composition are probably environment-specific and different in moderate versus thermophilic (Zeldovich et al. 2007), normosmic versus hypersaline (Paul et al. 2008), or aquatic versus terrestrial (Jobson and Qiu 2011) habitats.

Furthermore, in addition to the above effects of individual amino acids on the protein stability, there is a radical asymmetry between polar and nonpolar amino acids in the strength of pairwise interactions affecting protein structure. Although burying any individual hydrophobic amino acid into the interior of a protein can stabilize its structure, polar amino acids engage in stronger (i.e., higher deltaG) pairwise interactions (ionic and dipole interactions and hydrogen bond formation) than pairwise interactions (van der Waals and hydrophobic) that hold nonpolar residues together. Cysteine disulfide bridges are the only exception as they are the covalent bonds between two hydrophobic residues. This results in stronger epistatic interactions between selection forces operating on polar amino acids than those operating on nonpolar ones with the exception of cysteine. As a result, it is probably much easier to observe introduction of a single nonpolar amino acids into the protein core than of a single polar one. This is particularly true for protein interiors, as on the surface hydrogen bonds and dipole interactions occur between polar residues and water molecules, and thus the pairwise nature of polar interactions is of a lesser significance.

Earlier (Yampolsky and Bouzinier 2010), we reported a systematic net loss of polarity in drosophilid proteins that was more pronounced in proteins with intracellular localization than in membrane proteins. This observation suggested a possible role of structure-specific preferences toward particular amino acid substitutions. The observed flux was largely explained by gains of rare amino acids, including nonpolar amino acids M, I, C, F, and W and losses of frequent ones, including polar amino acids E, Q, and K, suggesting nonequilibrium. CpG codon-encoded R and S ranked low in either gains or losses. Generally, the loss–gain rank observed by Yampolsky and Bouzinier (2010) highly correlated with that reported by Jordan et al. (2005) analysis based on comparisons within bacteria, yeasts, and mammals with comparable phylogeny depths. The fact that the loss of polarity was more pronounced in soluble than in membrane-bound proteins suggested that the tertiary structure location of sites at which polarity-changing substitutions occurred might play a role in polarity-changing fluxes. If so, the "loss of common/gain of rare" explanation may not fully capture the nature of the net changes in amino acid composition, despite high occurrence of some polar amino acids (S, E) and low occurrence of some of the most nonpolar ones (M, C, W).

Two explanations for the existence of nonequilibrium amino acid composition can be suggested. Firstly, one might hypothesize that despite a long evolutionary history of selection for stability, protein interiors may still be far from an optimal amino acid composition in terms of polarity and size of residues (fig. 1*A*). In this case, the presence of net fluxes can be caused by substitutions at the interior sites, where, although polar amino acids are less common, polar to nonpolar substitutions are more likely to be beneficial than changes in the opposite direction (Koga et al. 2012). This hypothesis also predicts net gain of large amino acids at the interior sites, due to selection favoring smaller interior cavities (Kadonosono et al. 2003; Maeno et al. 2015). One would expect that evolutionarily older protein families may be closer to the equilibrium amino acid composition than younger ones and therefore show weaker net fluxes in residues' polarity and size.

Alternatively, the amino acid composition of the interior of proteins with respect to amino acids' polarity and volume may be close to an optimum in terms of the stability of proteins' tertiary structure (e.g., few polar and small residues in the interior). In this case, the observed fluxes should be stronger at the relatively weakly constrained surface of protein molecules, where polar residues are overrepresented. In this scenario, a nonzero net flux may be present only if such relaxation of stabilizing selection is an evolutionarily recent phenomenon, that is, if selective constraints have decreased over time, over either the lifespan of a protein family or overall geological time (fig. 1*B*). One may hypothesize that constraints may decrease over the lifespan of a protein because the foldability of the new protein is critical when a noncoding sequence, an alternative ORF, or a product of exon shuffling, are recruited to give birth to a new gene. Later, this constraint may become less important due to selection for stability that operates on cysteine bridges, or protein-specific chaperones, or translational fine-tuning that allows correct folding. Alternatively, constraints on amino acid composition may decrease over geological time regardless of the individual protein's age due to organisms acquiring a richer repertoire of chaperones, or strengthening intracellular homeostasis, for example, by evolving more stable pH or osmolality. Such hypothetically relaxed constraints on the originally polar amino acid-rich exterior will create a net flux toward polarity loss through neutral substitutions. This flux should largely occur at the surface sites (at least in water-soluble proteins) and may be stronger in evolutionarily younger proteins, although in this scenario the relationship with protein age may be complex.

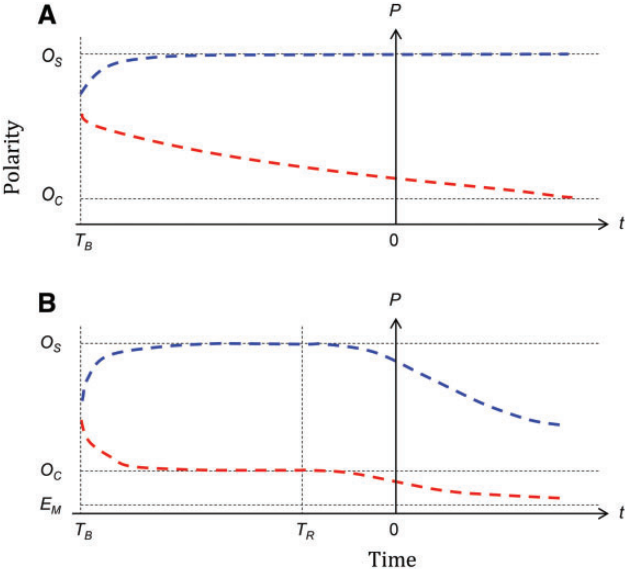

FIG. 1.—Schematic representation of the two hypotheses for the existence of net polarity fluxes. (*A*) Proteins' interior is under selection for lower average polarity; optimal composition has not been reached yet. (*B*) Recent relaxation of selective constraint on surface (and possibly core) sites; mutational equilibrium has not been reached yet. t, time; P, polarity; T_B, birth of a protein; T_R, relaxation of selective constraints; O_C, optimum polarity for the protein core; O_S, optimum polarity for the protein surface; E_M, polarity at mutational equilibrium. Red, protein core; blue, protein surface.

To summarize, both hypotheses imply nonequilibrium amino acid composition, the former hypothesis assuming largely selection-driven fluxes in the interior, whereas the latter one assuming largely neutral substitutions on the surface of the proteins. The two explanations are not mutually exclusive; it is entirely possible that some proteins, or even sites of the same proteins, are not at the equilibrium amino acid composition due to slow action of directional selection for optimality, whereas others show a nonzero flux due to recent relaxation of stabilizing selection.

In principle one more possible explanation exists. Any stability-protecting constraints may be too weak to generate noticeable selection pressure on individual amino acid substitutions and the observed fluxes are therefore the product of neutral mutations and drift. In this case, current amino acid composition, mutational spectra, and the properties of the genetic code should determine the fluxes' magnitude and sign. Given the highly nonrandom amino acid compositions in the core and exterior of proteins, this is unlikely for the overall protein sequences, but neutral expectations can provide a suitable null hypothesis for some portions of protein structure in which stability constraints are weak.

To test the predictions of these hypotheses, we calculated net change of polarity and volume of amino acid residues in >80,000 unambiguously reconstructed amino acid substitutions in 905 conserved protein families from 12 species of *Drosophila*. Using an extant protein sequence, we estimated relative solvent accessibility (RSA) at each site at which these substitutions occurred. To avoid a potential circularity introduced by the polarity of the ancestral residue that affects both the expected change in polarity and estimated RSA, we calculated solvent accessibility for ±2 residues surrounding each amino acid site at which these substitutions occurred (referred to as "context RSA") and used this measure for all analyses. The same analysis was done with 64 gene families for which experimental structural data are available. We then compared net changes of polarity and volume with those expected on the basis of the occurrence of amino acids at each RSA level, the structure of genetic code, and mutational biases. Finally, we estimated the evolutionary age of each protein family by finding the deepest clade that contained orthologs to each protein family and juxtaposed net fluxes observed in the same structural context in young versus old protein families.

Materials and Methods

Alignments and RSA Estimates

Protein alignments from 12 Drosophila genomes (Drosophila Comparative Genome Sequencing and Analysis Consortium 2007) were obtained from http://www.indiana.edu/~hahnlab/fly/DfamDB/drosophila_frb.html, last accessed October 27, 2017 (Hahn et al. 2007). Protein families with no indels longer than one amino acid were selected in order to be able to unambiguously match amino acid positions across homologs. Ancestral sequences at each node of the phylogenetic tree were reconstructed as previously described (Yampolsky and Bouzinier 2014). Solvent accessibility for each amino acid site of each of such 908 proteins was estimated by I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER, last accessed October 27, 2017; Yang et al. 2015) using the *D. melanogaster* sequence.

Using the estimate of RSA at the site of substitution would constitute circularity for both polarity and volume, resulting in a systematic bias toward loss of polarity and gain of volume at sites with high RSA. To reduce this bias, we used the average of RSA values for two amino acid sites on either side of the substitution site (not including the estimate for the site itself) as the measure of substitution site exposure. Such estimate is referred to as context RSA, as opposed to estimates for a specific site, referred to as site RSA. This averaging was done both without weights and with RSAs for the immediately adjacent sites given the weight of 2. The differences between results based on weighted and unweighted RSA averages are miniscule; the unweighted version is reported in the paper, whereas the main results for the weighted version of the estimate are available in the supplementary file S5, Supplementary Material online. Such averaging undoubtedly reduces our resolution in terms of site solvent accessibility, but represent the conservative approach with respect to possible biases. Correlations between context and site RSAs are shown on supplementary figure S2A, Supplementary Material online (supplementary file S5, Supplementary Material online).

Expected Change in Polarity and Volume

Grantham (1974) polarity units (AAIndex: GRAR740102) and Bigelow (1967) residue volume estimates (AAIndex: BIGC670101) were used as measures of residue's polarity and volume. The use of alternative estimates listed in the AAIndex (Kawashima et al. 2008) did not change the results in any noticeable manner. Expected change of polarity in substitutions from i-th amino acid was calculated as follows:

$$\pi_i = \frac{\sum_{k}^{Npath} \delta_{ik} b_{ik} f_{ik}}{\sum_{k}^{Npath} b_{ik} f_{ik}},$$

where N_{path} is the number of single nucleotide missense substitution paths from each codon encoding the i-th amino acid, δ_{ik} is the change of polarity as the result of a substitution through k-th path, b_{ik} is the mutational bias favoring this path and f_{ik} is the frequency of the codon capable to mutate through the k-th path. Codon frequencies in *Drosophila melanogaster* genome was used (http://www.kazusa.or.jp/codon/, http://www.indiana.edu/~hahnlab/fly/DfamDB/drosophila_frb.html; Nakamura et al. 2000). The bias coefficient was set to either 1 for all changes (no mutational biases), 2 for transitions (Keightley et al. 2009), or 20 for transitions at CpG sites. Amino acid substitutions requiring more than one nucleotide substitution were ignored.

Expected change of polarity by substitutions from all amino acids in all data, individual protein families, or at a subset of sites with a given solvent accessibility was calculated as the average π_i weighted by the frequency of i-th amino acid in the sequences:

$$E(dP_i) = \frac{\sum_{i}^{20} \pi_i A_i}{\sum_{i}^{20} A_i}$$

Expected change of residues' volume was calculated in the same manner. Data are available in supplementary file S2, Supplementary Material online.

Experimentally Determined RSA

We obtained estimates of solvent accessibility data for a subset of 64 *D. melanogaster* protein sequences for which a fully resolved experimental structure is available. To identify these sequences all PDB *Drosophila* sequences were blasted against the sequences in our data set and best hits with >98% identity and <2 substitutions were selected. Structures of these 64 proteins were downloaded from PDB and solvent accessibility calculated by DSSP (http://swift.cmbi.ru.nl/gv/dssp, last accessed October 27, 2017; Touw et al. 2015), with ASA values converted into RSA values by "empirical" normalization (Tien et al. 2013). These results are available in supplementary file S3, Supplementary Material online.

Gene Ontologies

Cellular localization and molecular function GO terms for each gene family were obtained from FlyBase (Gramates et al. 2017) with the FBgn ID of the *D. melanogaster* member of the family as the retrieval ID. When conflicting GO terms were reported for a given *D. melanogaster* gene (e.g., both "intracellular" and "membrane" cellular localization), or when there were two paralogous *D. melanogaster* genes present with different GO terms reported, such GO was termed "ambiguous". For the purpose of further analysis such cases were lumped together with those for which GO terms are unknown or unavailable.

Protein Age

A randomly chosen *Drosophila* sequence from each gene family was blasted against RefSeq database; the deepest clade with hits to at least 5 different species with evalue <1e-6 was identified. The common ancestor of such clade was used as the lower-bound estimate of the protein family age. Approximate ages of divergence of each clade are listed in supplementary table S1, Supplementary Material online. The effects of protein age and RSA on the magnitude of absolute changes in amino acid polarity and volume were analyzed by multivariate regression, with protein age log-transformed. Because there are no recent proteins among the intracellular proteins and few very ancient ones among the extracellular ones (see Results), the above analysis may be misleading due to confounding between protein age and cellular localization. Furthermore, because each protein family is characterized by a single age estimate and a single cellular localization assignment, using individual substitutions as independent observations inflates the number of degrees of freedom. To ameliorate both problems we calculated mean absolute polarity and volume change for each protein family, either for all sites or separately for 3 bins of context RSA values (buried, RSA < 0.3; intermediate, 0.3 < RSA < 0.5; and exposed, RSA > 0.5), and used these protein family means as independent observations.

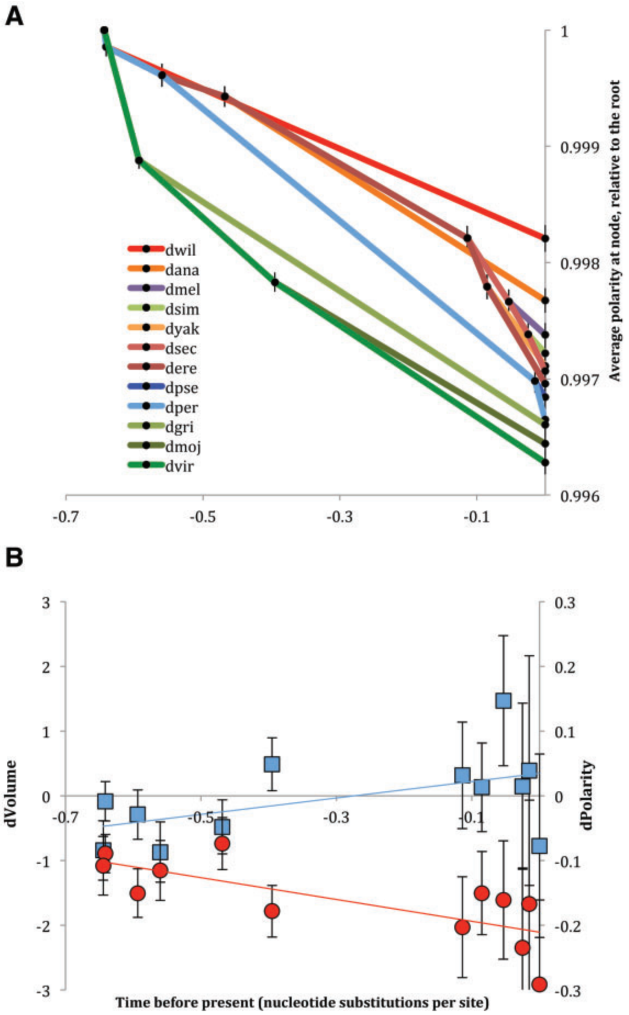

Fig. 2.—(*A*) Phylogenetic tree of 12 Drosophila species with vertical position of nodes reflecting average polarity of amino acids of reconstructed proteins at this node. Vertical bars are standard errors caused by variability among proteins and sites. Species names: dwil = *D. willingstoni*, dana = *D. ananassae*, dmel = *D. melanogaster*, dsim = *D. similans*, dyak = *D. yakuba*, dsec = *D. secchelia*, dere = *D. erecta*, dpse = *D. pseudoobscura*, dper = *D. persimilis*, dgri = *D. grimshawi*, dmoj = *D. mojavensis*, dvir = *D. virilis*. Regression coefficient of mean polarity of extant or reconstructed protein sequences over age in nucleotide substitution units = −0.00405; $P <$ 1E-8). (*B*) Mean change in residues' volume (Å^3, squares, left axis) and polarity (arbitrary units, circles, right axis). Vertical bars are approximate 95% CIs. Linear regressions for volume and polarity change, respectively: dVolume = 0.35 + 1.27×age ($P > 0.09$); dPolarity = −0.21 + 0.17×age ($P < 0.006$). Horizontal axis is the same on both (*A*) and (*B*) and represent time before present in nucleotide substitution units.

Linear regressions of mean fluxes over the age of substitutions (fig. 2) or protein (fig. 5) were calculated using averages for each independent variable category to avoid overinflation of the number of degrees of freedom. All statistical analysis was conducted using JMP (JMP Statistical manual, 2012)

Results

Average polarity decreased in all lineages of the 12 species of *Drosophila* studied; this decrease is small in magnitude (~0.3% relative to the ancestral condition over the course of ~70 Myr), but consistent among lineages and among phylogeny nodes of different age (fig. 2). The average change in Grantham (1974) polarity units per amino acid substitution across all sites and all lineages was −0.127 (SE = 0.0073). There was a slight trend toward an accelerated rate of net loss of polarity from early to recent ages (fig. 2*B*), which is particularly evident for substitutions that occurred in the *Drosophila melanogaster* clade (fig. 2*A*). The changes in average volume were less consistent: there was a hint of a net loss of large amino acids in early edges of the tree replaced by a slight net gain of large amino acids in more recent edges (fig. 2*B*).

This small but consistent loss of average polarity has been largely achieved through relatively small-step amino acid substitutions and not by more radical, but also much less frequent changes (supplementary fig. S1, Supplementary Material online). The majority of strong-contribution pairs of source (ancestral) and destination (derived) amino acids are located close to the diagonal of the source-by-destination matrix (sorted by polarity of source and destination amino acids). There were two notable exceptions. One was alanine, an intermediate polarity amino acid, which is, simultaneously, the second strongest polarity gain source amino acid, largely through relatively radical A->E changes, and the fifth strongest polarity-loss destination, largely through the E->A changes. The other exception involved the more radical polarity-losing and polarity-gaining Q->L and, respectively, L->Q changes. The most radical net polarity gain and polarity loss substitutions were not strong contributors to overall net polarity changes due to their rarity probably caused by both the structure of the genetic code and strong negative selection against such changes.

Supplementary figure S2A, Supplementary Material online, shows that the context RSA (see Materials and Methods) is significantly correlated with the site RSA, indicating that the context RSA is a suitable estimate of the site exposure. Thus, the results based on context RSA estimates are free of biases introduced by the specific amino acid occupying a given site, and they reflect solvent accessibility in the vicinity of this site reasonably well, at least allowing reliable differentiation between buried, intermediately exposed, and highly exposed regions.

Plotted against the context RSA estimates, mean net changes in polarity show that the loss of polar amino acids occurred at all sites; the net change is significantly <0 for all sites except the most exposed and the most buried ones (fig. 3*A*). At the sites with intermediate solvent accessibility (context RSA 0.3–0.4) the observed loss of polarity was similar to that predicted by the neutral expectation based on amino acid occurrences at such sites. Net changes in polarity in 65 proteins for which experimental structural data are available showed a slightly different pattern (fig. 3*B*). Just as for the predicted RSA, the strongest net polarity loss was observed at sites with RSA ~0.6. For other RSA values, polarity shows no change.

A net gain in amino acid volume occurred largely at both the most buried and the most exposed sites (fig. 2*C*), where the net gain was not significantly different from either 0 or from the neutral prediction. At intermediate context RSA values, where the neutral prediction calls for net volume loss, negative volume changes were indeed observed, although of a lower magnitude than predicted. No significant net change in residue volume was observed in substitutions in the proteins with known experimental structure (fig. 3*D*), except for the slight loss of volume at the most exposed class of sites for which substitutions were observed, where it is, however, not different from the neutral prediction.

Combined data on net loss or gain of polarity and volume can be misleading, because processes shaping these fluxes may be different in proteins with different cellular localization. The same data as on figure 3*A* and *C*, but separately for proteins with intracellular, membrane, and extracellular localization are presented on figure 4. These result show that most of the observed net polarity loss occurred at the interior sites of proteins with intracellular localization, which are presumably water-soluble proteins with polar exterior and nonpolar interior (fig. 4*A*). At sites with context RSA < 0.2, this loss occurred despite the neutral prediction of a net gain of mean polarity. At intermediate context RSA values the polarity loss was similar to the neutral prediction, whereas at the most exposed sites, it was not different from 0. There was also a slight and not statistically significant net gain of volume at the most buried sites in such proteins (fig. 4*E*).

In contrast, no net change of polarity or volume occurred at the buried sites of proteins with membrane (fig. 4*B* and *F*) or extracellular (fig. 4*C* and *G*) localization, with a net gain of polarity occurring at the most exposed sites in membrane proteins. Most nonzero losses of polarity in such proteins were similar to the neutral prediction, in particular in the extracellular proteins. Proteins with unknown cellular localization largely followed the patterns seen in membrane and extracellular proteins. The subset of proteins with experimentally determined structures could not be analyzed with respect to cellular localization because most of these proteins are annotated as intracellular (47 out of 65; cellular localization of additional ten proteins is unknown).

Estimates of the evolutionary age of analyzed protein families are summarized in supplementary table S1, Supplementary Material online. As expected, the vast majority of intracellular proteins are ancient, with cellular organisms or eukaryotes as the outmost clade with orthologs present. None originated earlier than the divergence of *Endopterygota* (holometabolic insects). In contrast and, again, as expected, few extracellular proteins originated prior to the appearance of multicellular animals and the most common class (ten families) appeared to be associated with the

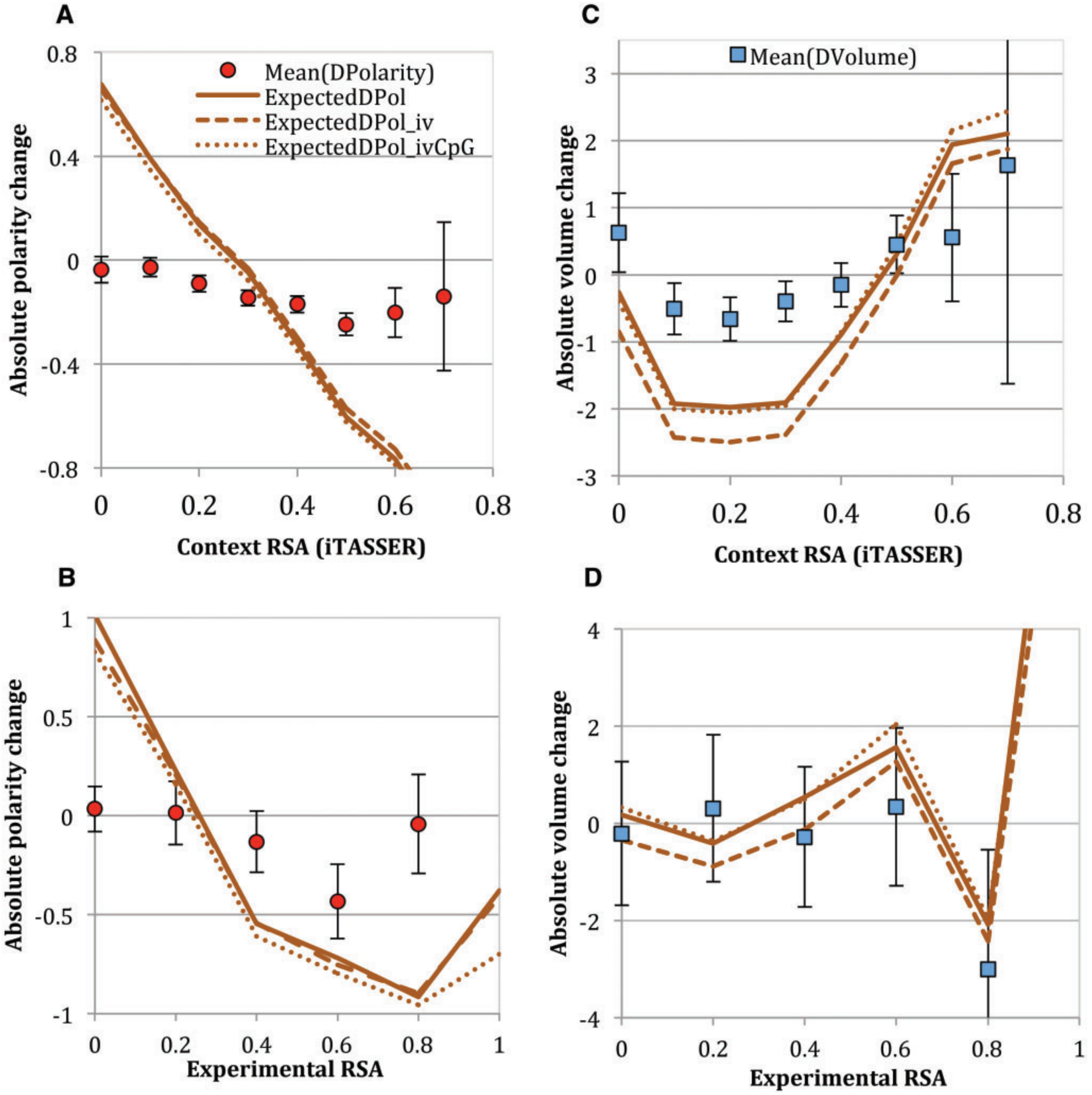

Fig. 3.—Mean net change in polarity (*A*, *B*; circles) and volume (*C*, *D*; squares) at sites with different relative solvent accessibility at neighboring sites (context RSA). Context RSA calculated either from I-TASSER-predicted (all proteins; *A*, *C*) or from experimentally determined solvent accessibility (64 proteins with structure data; *B*, *D*). Vertical bars are approximate 95% CIs. Neutral expectations: solid line, with no mutational biases; dashed line, with a 2-fold transition/transversion bias; dotted line, with 2-fold transition/transversion bias and 10-fold CpG bias.

divergence of neopteran insects. Proteins with membrane or unknown cellular localizations showed intermediate age distribution. This complicates the analysis, as the youngest intracellular proteins and the oldest extracellular proteins were represented by just a handful possibly highly biased families.

Contrary to the predictions, protein age did not show any effect on the strength of polarity loss either in the hydrophobic core or in the polar exterior of proteins (fig. 5 and table 1). There was a marginally significant ($P < 0.03$) age-by-RSA interaction effect for all proteins analyzed together, with the exposed sites showing, as expected, a slightly stronger polarity loss in younger proteins and figure 5. However, this effect does not survive multiple correction tests and is not replicated in the similar analysis conducted for each cellular localization class separately. Similarly, little evidence of the protein age effect is observed in the analysis using protein family means as independent observations (supplementary fig. S3 and table S2, Supplementary Material online). None of the 2-way ANOVA tests with age class and RSA class as factors showed either a significant effect of protein age or a significant protein age-by-RSA interaction. The only exception was the test for extracellular proteins, in which both effects were significant (supplementary fig. S2 and table S2, Supplementary Material online), largely due to unusually strong loss of polarity at the surface sites in the proteins that originated within the 500–700 Ma bracket, which includes early metazoan evolution.

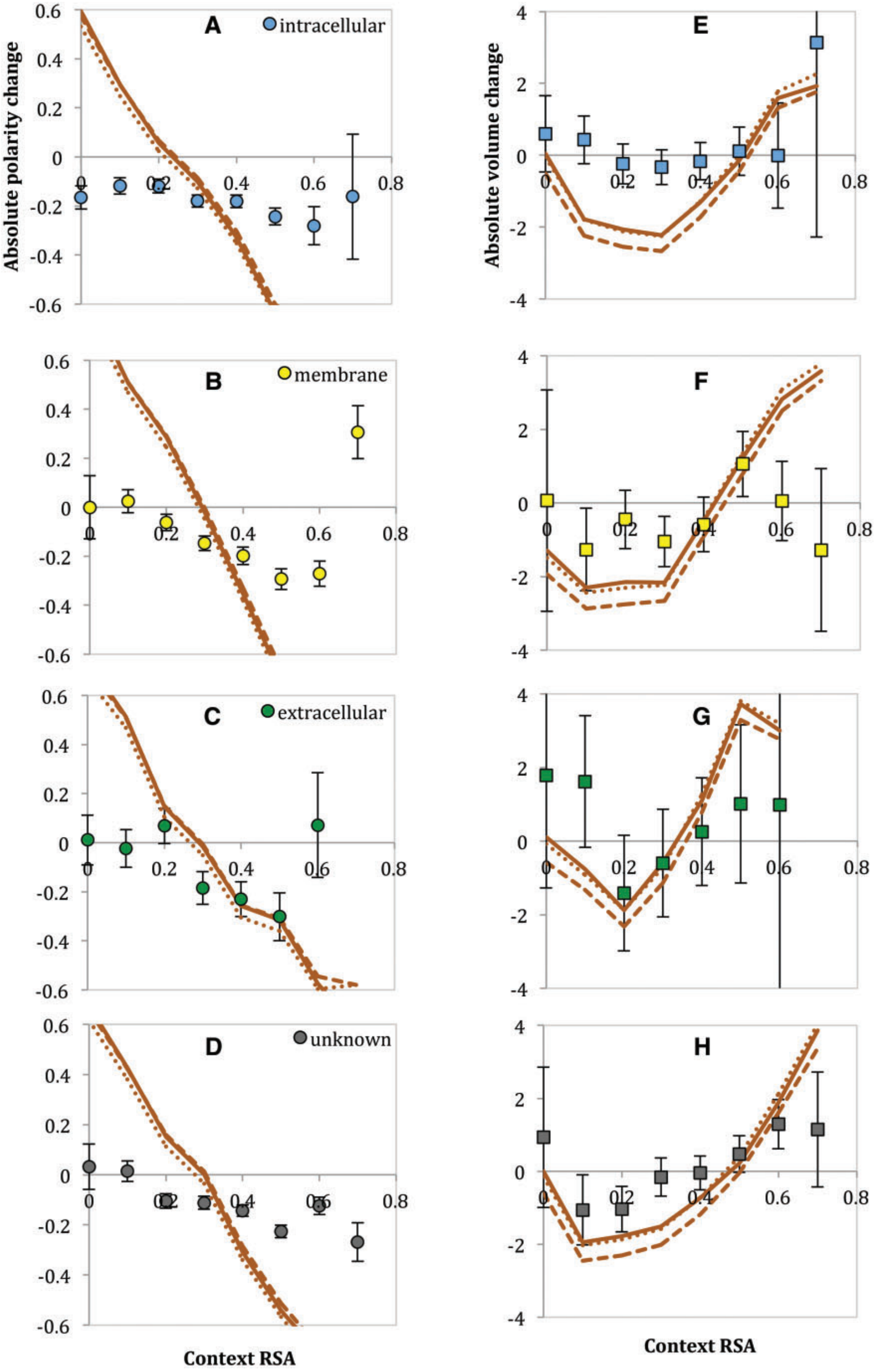

Fig. 4.—Mean net change in polarity (*A*–*D*) and volume (*E*–*H*) at sites with different relative solvent accessibility at neighboring sites (context RSA) in proteins with intracellular (*A* and *E*), membrane (*B* and *F*), extracellular (*C* and *G*), and unknown (*D* and *H*) cellular localization according to FlyBase GO annotations. Note different scales. Symbols, lines, and error bars as on figure 3.

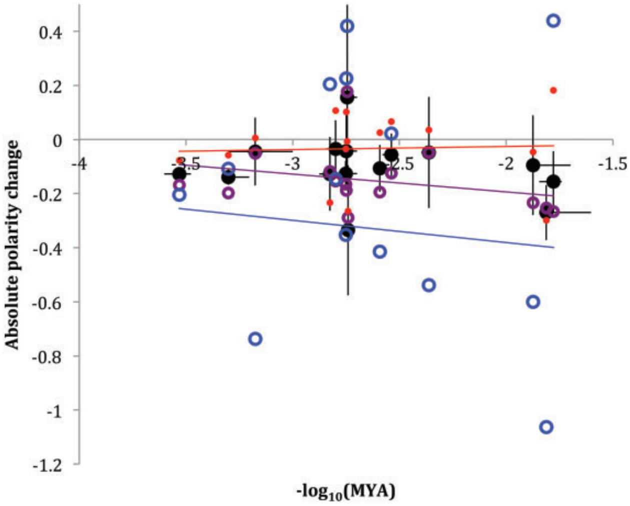

Fig. 5.—The relationship between protein family age (approximate Myr, log scale) and mean net change in polarity. Black dots: all substitutions; small red dots: protein interiors (I-TASSER-estimated context RSA 0–0.2); medium purple circles: intermediate sites (context RSA 0.3–0.5); large blue circles: exterior sites (context RSA > 0.5). Horizontal bars represent ranges of age estimate (supplementary table S1, Supplementary Material online). Vertical bars are approximate 95% CIs. Linear regression coefficients of mean dPolarity over log-transformed protein age: −0.042 ($P > 0.40$), 0.012 ($P > 0.88$), −0.066 ($P > 0.28$), and −0.082 ($P > 0.81$) for all, buried, intermediate, and surface sites, respectively. Error bar only shown for all sites data points for clarity.

Discussion

We analyzed net gains and losses of polarity and volume of amino acid residues along the phylogenetic tree of 12 species of *Drosophilidae* family which occurred at amino acid sites with different predicted solvent accessibility and in proteins of different cellular localization and evolutionary age. We observed a miniscule but consistent net evolutionary loss of average amino acid polarity across all drosophilid lineages. The net loss of polar residues at the most buried amino acid sites occurred in proteins with intracellular localization, despite polar amino acids being underrepresented at such sites, that is, despite the neutral prediction of polarity gain. Likewise, the interior sites of such proteins show some net gain of large amino acids (a weak gain, as the 95% confidence interval around mean net change includes 0), despite the neutral expectation of net loss of residue volume. No net loss of polarity is observed at buried sites of membrane and extracellular proteins. Proteins with extracellular localization indeed appeared to lose polar amino acids at sites with intermediate exposure, but such loss was not different form the neutral expectation. No net loss of polarity or volume is observed at the most exposed sites (RSA > 0.6); in fact, there is a nonzero gain of polar amino acids at such sites in proteins with membrane localization. To summarize, we observed patterns consistent with the hypothesis that current amino acid composition at buried and exposed protein sites is not just far from a dynamic equilibrium, but also far from the optimal composition in terms of proteins stability. Specifically, a typical intracellular drosophilid protein appears to have space for improvement in terms of having the interior mostly consisting of hydrophobic residues. The approach to such equilibrium is slow, possibly due to the pairwise interactions between polar residues still present in the core of proteins. There is no evidence that ancient proteins are closer to it than those originating more recently. The lack of net loss of polarity at the exposed sites suggests that there is no evidence favoring the hypothesis of recently relaxed selective constraint on the exterior of proteins, with the possible exception of extracellular proteins.

Several caveats need to be addressed to assess the generality of these findings. First of all, notwithstanding the support from the small number of proteins with experimentally resolved structure (fig. 3C), most of the results rely on homology-based prediction of solvent accessibility. The use of context RSA that does not take into the account the RSA estimate for the site of the substitution (see Materials and Methods) reduces but does not fully eliminate the bias toward predicting high RSA for sites with a polar ancestral amino acid. This potential bias, however, should lead to erroneously high estimates of polarity loss at exposed sites. This bias is therefore a possible explanation for the observed loss of polarity for the exterior sites of extracellular proteins, but it could only lead to the underestimate of loss of polarity at buried sites.

It should also be noted that due to the need of unambiguous ancestral state reconstruction, the analysis was limited to a conservative subset of sites in a conservative subset of proteins, as families with indels longer than one amino acid sites with ambiguous reconstructions were omitted. Furthermore, the analysis assumes that there were no radical changes in protein functionality (i.e., cellular localization) in the course of drosophilid evolution because GO annotations and experimental structure data are limited to *D. melanogaster*. We cannot exclude the possibility that if we could analyze a set of faster evolving proteins our conclusions would have been different. For example, we might have observed amino acid composition that is closer to an equilibrium and therefore has no net fluxes of any kind, or, alternatively, has a much greater relaxation of stabilizing selection at exposed sites, leading to neutral loss of polar amino acids at such sites. Thus, the results reported truly apply only to moderately conserved proteins.

On the other hand, the bias toward conserved proteins also probably means that the analyzed subset of proteins was enriched in ancient gene families and poor in gene families recently recruited from noncoding or functionally unrelated coding sequences. Thus, this subset of proteins has had a longer evolutionary time to approach the equilibrium amino acid composition, whereas the same analysis done in younger gene families might have revealed even stronger deviations from the structurally optimal amino acid composition.

With these caveats in mind, one conclusion is probably secure: amino acid composition of drosophilid proteins is

Table 1
Multivariate Regression of Absolute Changes in Polarity and Volume of Amino Acids on Protein Age and Context RSA on the Magnitude in All Proteins in the Data Set and for Each Cellular Localization Class Separately

	Response: dPolarity				Response: dVolume			
Source	**DF**	**SS**	**F**	**P**	**DF**	**SS**	**F**	**P**
All proteins								
ContextRSA	1	344	80.77	<0.0001	1	3,392	7.53	0.0061
Age	1	1.001	0.24	0.63	1	617.9	1.37	0.24
ContextRSA×age	1	20.299	4.77	0.029	1	339.9	0.75	0.39
Error	77,878	331,682.8			77,878	35,077,954		
Intracellular proteins								
ContextRSA	1	56.85	13.22	0.0003	1	132.41	0.31	0.58
Age	1	1.046	0.24	0.62	1	828.04	1.95	0.16
ContextRSA×age	1	0.066	0.015	0.90	1	34.25	0.08	0.78
Error	27,084	116,509.7			27,084	11,477,940		
Membrane proteins								
ContextRSA	1	141.04	35.84	<0.0001	1	708.04	1.47	0.22
Age	1	2.874	0.73	0.39	1	169.01	0.35	0.55
ContextRSA×age	1	0.019	0.005	0.94	1	256.32	0.53	0.47
Error	14,672	57,735.7			14,672	7,043,504.3		
Extracellular proteins								
ContextRSA	1	33.83	8.02	0.0046	1	7.89	0.02	0.9
Age	1	0.432	0.1	0.75	1	901.19	1.75	0.19
ContextRSA×age	1	1.891	0.45	0.50	1	925.99	1.80	0.18
Error	3,919	16,528.3			3,919	2,019,768.3		
Celular locallization unknown								
ContextRSA	1	116.99	26.74	<0.0001	1	5,329.75	11.81	0.0006
Age	1	8.869	2.03	0.15	1	2,192.76	4.86	0.03
ContextRSA×age	1	12.172	2.782	0.10	1	955.51	2.12	0.15
Error	32,191	140,826.7			32,191	14,526,737		

not at an equilibrium with respect to average polarity. Given the small magnitude of the net losses of polarity, one should not be surprised that this process has not yet achieved an equilibrium even in the oldest proteins. In fact, extrapolating the observed relative rate of change of polarity (~0.3% >70 Myr) for the lifespan of even the most ancient proteins translates in only ~15% of average polarity loss relative to the ancestral state and, of course, much less for more recently acquired proteins. Furthermore, one may hypothesize that for many proteins the field of selective pressures in terms of amino acid composition has been reset or at least altered much more recently, when ancestral insects colonized land, that is, ~400 Ma (Gaunt and Miles 2002), only ~6 times the span covered by the drosophilid phylogeny. Data from plant genomes (Jobson and Qiu 2011) indicate that the transition to land was accompanied by an increase of occurrence of some polar (in particular charged) and some nonpolar (in particular aromatic) amino acids. These changes are thought to be caused by selective pressure to develop proteins more resistant to desiccation and UV radiation (Jobson and Qiu 2011). We are unaware of such data for animals, but one might hypothesize that a similar selective preference has been operating since insects' transition to land. However, we do not observe patterns reported by Jobson and Qiu (2011). On the contrary, the charged amino acids lysine and glutamate are the two greatest net losers; among the aromatic amino acids tyrosine is a moderate loser, tryptophan is a slight loser, and phenylalanine and histidine are moderate gainers.

The differences between intracellular and extracellular proteins (fig. 4*A* and *C*) require special attention. Proteins with extracellular localization do show the pattern predicted by the "relaxed selective constraints on the surface" hypothesis: most of polarity loss in these proteins occurs at the moderately exposed sites and, moreover, this loss is seen only in those proteins that originated relatively soon after the origin of multicellular animals (supplementary fig. S2, Supplementary Material online). Proteins with extracellular localization are, expectedly, evolutionarily younger than those with intracellular or membrane localization, as regular excretion of proteins, crucial in multicellular organisms, was probably not as common in the unicellular ancestral forms. It has been long recognized that extracellular and intracellular proteins differ with respect of their amino acid composition (Nakashima and Nishikawa 1994), including, in particular, polarity of surface residues (Feng and Zhang 2001). Not surprisingly, the set of proteins analyzed here is not an exception (fig. 6): on the

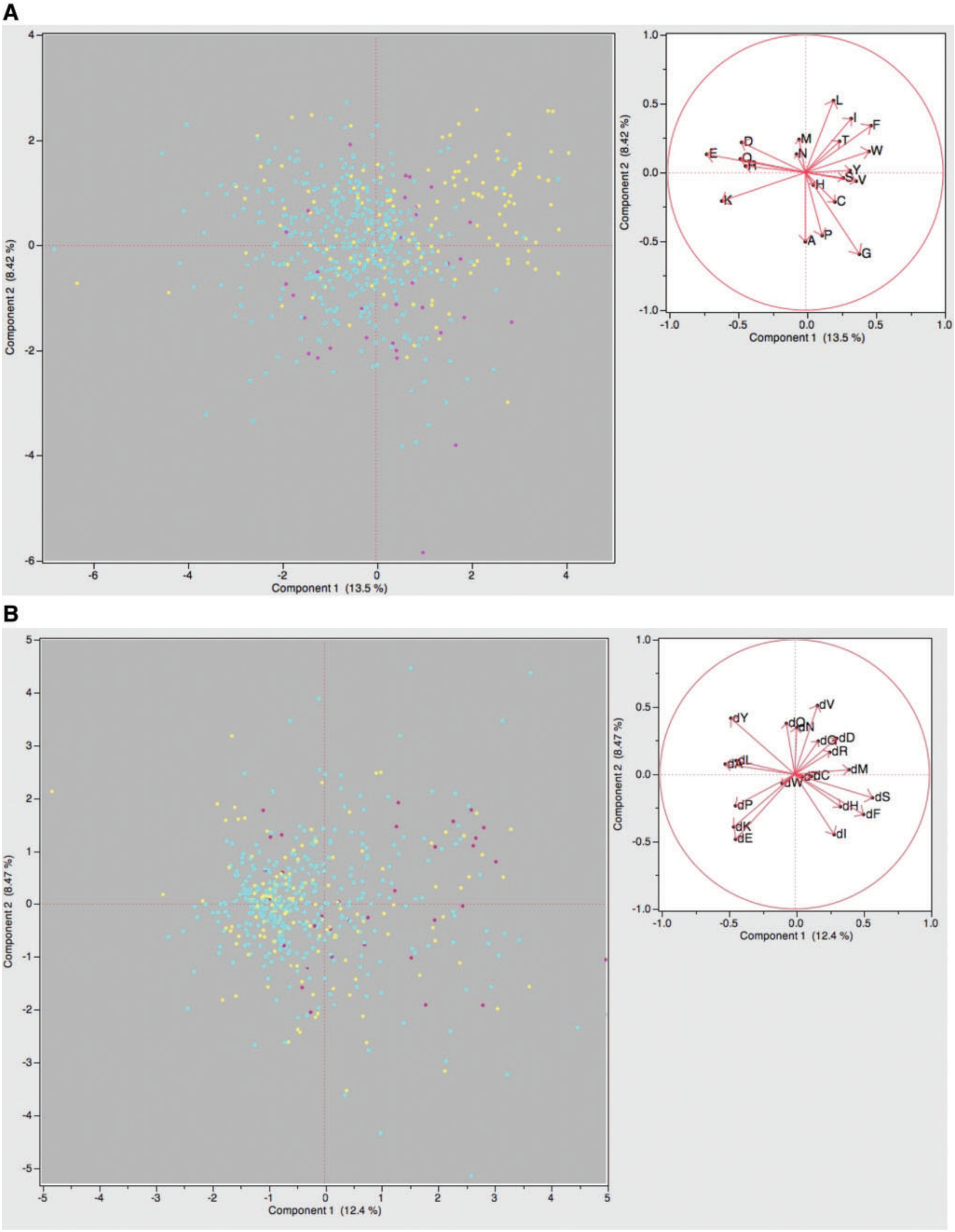

FIG. 6.—(*A*) Principal components analysis of 880 *Drosophila melanogaster* proteins in the analyzed data set by relative occurrence of 20 amino acids. Blue dots, intracellular proteins; purple dots, extracellular proteins; yellow dots, membrane proteins. Proteins with unknown cellular localization are included in the analysis, but not shown. (*B*) Principal components analysis of 905 *Drosophila* spp. proteins in the analyzed data set by net gains/losses of 20 amino acids.

plane of two first principal components the intracellular proteins tend to be centered around the overall center of gravity, extracellular proteins form a ring around them, whereas membrane proteins are shifted toward nonpolar amino acids. However, we are unaware of any data specifically indicating that the exteriors of extracellular proteins are under stronger selective constraint with respect to their amino acid composition. In principle, such different selective pressures might be caused by a greater occurrence of surface sites interacting with receptors or by a more hydrophobic environment such protein experience, relative to intracellular ones. Likewise, the low net loss of polarity observed at buried sites in extracellular as well as membrane-bound proteins (fig. 4*B* and *C*) may be an indication of different structure-imposed selective pressures that these proteins experience in the course of their evolution (Franzosa et al. 2013). It may be revealing that while the intracellular proteins are centered around the same center as the intracellular ones on the PCA by amino acid occurrences (fig. 6*A*), they are shifted away from the modal intracellular proteins on the PCA by net gains and losses of amino acids (fig. 6*B*).

How much of the observed patterns of net fluxes shown on figures 3 and 4 can be explained by local amino acid composition and mutation biases? Predictions based on RSA-specific amino acid occurrences, genetic code properties, and mutational biases do, of course, change from buried to exposed sites; different mutational biases result in fairly similar predictions. However, the only nonzero net loss of polarity that closely follows the neutral predictions is observed at the sites with intermediate context RSAs, mostly in intracellular proteins and proteins with unknown localization (fig. 4*A* and *D*). It should be noted that the incorporation of mutational biases does not radically change the neutral predictions, indicating that mutational argument (Misawa et al. 2008) is hardly applicable to drosophilids. Although the transition–transversion bias has been well documented in *Drosophila*, CpG bias is not observed (Keightley et al. 2009). However, one should keep in mind that context-specific mutation rates may differ among drosophilid lineages (Chachick and Tanay 2012) and that the absence of CpG bias was detected in a relatively short laboratory experiment (Keightley et al. 2009) and may not reflect long-term mutability of CpG sites due to higher levels of germline DNA methylation in conditions different from standard laboratory conditions.

The observed net changes in amino acids residues' polarity and, to some extent, residues' volume, are consistent with the hypothesis that the extant drosophilid proteins with intracellular localization appear to still have nonequilibrium and higher-than-optimal occurrence of polar residues in their interiors and that the net fluxes in amino acid composition are therefore caused by continuing selection for protein stability. In other words, proteins currently in existence are far from perfect in terms of their amino acid composition effects on structural stability, confirming a recent theoretical analysis (Hormoz 2013). Indeed, recent successes in protein engineering of proteins that are more stable, more temperature resistant and more reliably folding (Wijma et al. 2013; Xiong et al. 2014; Denard et al. 2015) than their naturally occurring counterparts is a perfect illustration of this fact.

What are the causes of the slow rate of approach to an optimal amino acid composition? We can formulate several hypotheses. Firstly, there simply may have not been enough evolutionary time for the equilibrium to be approached. In this case we would expect stronger recent fluxes in younger proteins, which we did not observe. Alternatively, such an optimal amino acid composition may be a moving target due to geological scale environmental changes that affect thermodynamics of protein folding, such as temperature (Zeldovich et al. 2007), or shifts in the lineage-specific environment, such as transition to land (Jobson and Qiu 2011). Likewise, continuing selection for protein stability may be endlessly counterbalanced by directional selection for rare substitutions that create new functionality. Such substitutions often compromise stability (Tokuriki et al. 2008), possibly regardless of the residue polarity change. All these explanations are consistent with the lack of the correlation between flux magnitude and protein age. Finally, some portion of the observed polarity loss observed at amino acid sites with intermediate solvent accessibility is indistinguishable from that expected based on random drift, suggesting previously existing constraint on hydrophobicity of such sites that no longer exists.

Conclusions

Average amino acid polarity has consistently decreased as the result of evolutionary substitutions occurring in 12 *Drosophila* lineages. This net loss of polarity mostly occurred at core sites of intracellular proteins and shows no correlation with the evolutionarily age of proteins. Net gains and losses of polarity and volume in membrane and extracellular proteins are either close to 0 or similar to neutral predictions based on amino acid composition and properties of the genetic code, except for the most exposed sites. These observations are consistent with the idea that the optimal amino acid composition of many proteins has not yet been reached (fig. 1*A*). The lack of correlation between the polarity flux and the evolutionary age of protein families suggests that such optimal composition either does not exist or may never be reached.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We are grateful to Matthew Hahn for the Drosophila protein alignments, to Yang Zhang for assistance with I-TASSER, and to Fyodor Kondrashov, King I. Jordan, Galina Yampolskaya, Claus Wilke, and several anonymous reviewers for useful discussion, comments, and criticisms. L.Y.Y. has been supported by NSF 1136706 and Y.I.W. by intramural funds of the US Department of Health and Human Services (to the National Library of Medicine).

Literature Cited

Baldwin RL. 2007. Energetics of protein folding. J Mol Biol. 371(2):283–301.

Bigelow CC. 1967. On the average hydrophobicity of proteins and the relation between it and protein structure. J Theor Biol. 16(2):187–211.

Bloom JD, Raval A, Wilke CO. 2007. Thermodynamics of neutral protein evolution. Genetics 175(1):255–266.

Bueno M, Campos LA, Estrada J, Sancho J. 2006. Energetics of aliphatic deletions in protein cores. Protein Sci. 15(8):1858–1872.

Cambillau C, Claverie JM. 2000. Structural and genomic correlates of hyperthermostability. J Biol Chem. 275(42):32383–32386.

Camps M, Herman A, Loh E, Loeb LA. 2007. Genetic constraints on protein evolution. Crit Rev Biochem Mol Biol. 42(5):313–326.

Chachick R, Tanay A. 2012. Inferring divergence of context-dependent substitution rates in Drosophila genomes with applications to comparative genomics. Mol Biol Evol. 29(7):1769–1780.

Conant GC, Stadler PF. 2009. Solvent exposure imparts similar selective pressures across a range of yeast proteins. Mol Biol Evol. 26(5):1155–1161.

Denard CA, Ren H, Zhao H. 2015. Improving and repurposing biocatalysts via directed evolution. Curr Opin Chem Biol. 25:55–64.

Drosophila Comparative Genome Sequencing and Analysis Consortium. 2007. Evolution of genes and genomes in the context of the Drosophila phylogeny. Nature 450:203–218.

Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. Nat Rev Genet. 17(2):109–121.

Feng ZP, Zhang CT. 2001. Prediction of the subcellular location of prokaryotic proteins based on the hydrophobicity index of amino acids. Int J Biol Macromol. 28:255–261.

Franzosa EA, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. Mol Biol Evol. 26(10):2387–2395.

Franzosa EA, Xue R, Xia Y. 2013. Quantitative residue-level structure-evolution relationships in the yeast membrane proteome. Genome Biol Evol. 5:734–744.

Garcia-Seisdedos H, Ibarra-Molero B, Sanchez-Ruiz JM. 2012. How many ionizable groups can sit on a protein hydrophobic core? Proteins Struct Funct Bioinformatics 80:1–7.

Gaunt MW, Miles MA. 2002. An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. Mol Biol Evol. 19:748–761.

Grantham R. 1974. T Amino acid difference formula to help explain protein evolution. Science 185:862–864.

Gramates LS, et al. 2017. FlyBase at 25: looking to the future. Nucleic Acids Res. 45(D1):D663–D671.

Hahn MW, Han MV, Han S-G. 2007. Gene family evolution across 12 Drosophila genomes. PLoS Genet. 3(11):e197.

Hormoz S. 2013. Amino acid composition of proteins reduces deleterious impact of mutations. Sci Rep. 3:2919.

JMP Statistics Manual. 2012. Gary, NC: SAS Institute Inc.

Jobson RW, Qiu YL. 2011. Amino acid compositional shifts during streptophyte transitions to terrestrial habitats. J Mol Evol. 72(2):204–214.

Jordan IK, et al. 2005. A universal trend of amino acid gain and loss in protein evolution. Nature 433(7026):633–638.

Kadonosono T, Chatani E, Hayashi R, Moriyama H, Ueki T. 2003. Minimization of cavity size ensures protein stability and folding: structures of Phe46-replaced bovine pancreatic RNase A. Biochemistry 42(36):10651–10658.

Kawashima S, et al. 2008. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 36:D202–D205.

Keightley PD, et al. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. Genome Res. 19:1195–1201.

Koga N, et al. 2012. Principles for designing ideal protein structures. Nature 491(7423):222–227.

Koshi J, Goldstein R. 1997. Mutation matrices and physical-chemical properties: correlations and implications. Proteins Struct Funct Genet. 27:336–344.

Liu X, et al. 2010. Genome wide exploration of the origin and evolution of amino acids. BMC Evol Biol. 10:77.

Maeno A, et al. 2015. Cavity as a source of conformational fluctuation and high-energy state: high-pressure NMR study of a cavity-enlarged mutant of T4 lysozyme. Biophys J. 108(1):133–145.

Mannige RV, Brooks CL, Shakhnovich EI. 2012. A universal trend among proteomes indicates an oily last common ancestor. PLoS Comput Biol. 8(12):e1002839.

Mannige RV. 2014. Origination of the protein fold repertoire from oily pluripotent peptides. Proteomes 2(2):154–168.

Mirny LA, Shakhnovich EI. 1997. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol. 291:177–196.

Misawa K, Kamatani N, Kikuno RF. 2008. The universal trend of amino acid gain–loss is caused by CpG hypermutability. J Mol Evol. 67(4):334–342.

Nakamura Y, Gojobori T, Ikemura T. 2000. Codon usage tabulated from the international DNA sequence databases: status for the year 2000. Nucleic Acids Res. 28(1):292.

Nakashima H, Nishikawa K. 1994. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol. 238(1):54–61.

Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7(5):337–348.

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(4):R70.

Serohijos AW, Shakhnovich EI. 2014. Contribution of selection for protein folding stability in shaping the patterns of polymorphisms in coding regions. Mol Biol Evol. 31(1):165–176.

Spielman SJ, Wilke CO. 2013. Membrane environment imposes unique selection pressures on transmembrane domains of G protein-coupled receptors. J Mol Evol. 76(3):172–182.

Tadeo X, et al. 2009. Structural basis for the aminoacid composition of proteins from halophilic Archea. PLoS Biol. 7(12):e1000257.

Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. 2013. Maximum allowed solvent accessibilites of residues in proteins. PLoS One 8(11):e80635.

Tokuriki N, Stricher F, Serrano L, Tawfik DS. 2008. How protein stability and new functions trade off. PLoS Comput Biol. 4(2):e1000002.

Touw WG, et al. 2015. A series of PDB related databases for everyday needs. Nucleic Acids Res. 43(D1):D364–D368.

Trifonov EN. 2004. The triplet code from first principles. J Biomol Struct Dyn. 22(1):1–11.

Wijma HJ, Floor RJ, Janssen DB. 2013. Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. Curr Opin Struct Biol. 23(4):588–594.

Wylie CS, Shakhnovich EI. 2011. A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc Natl Acad Sci U S A. 108(24):9916–9921.

Xiong P, et al. 2014. Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. Nat Commun. 5:5330.

Yampolsky LY, Bouzinier MA. 2010. Evolutionary patterns of amino acid substitutions in 12 *Drosophila* genomes. BMC Genomics 11 Suppl 4:S10.

Yampolsky LY, Bouzinier MA. 2014. Faster evolving Drosophila paralogs lose expression rate and ubiquity and accumulate more nonsynonymous SNPs. Biol Direct. 9:2.

Yang J, et al. 2015. The I-TASSER Suite: protein structure and function prediction. Nat Methods 12(1):7–8.

Yang ZH, Swanson WJ. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol. 19(1):49–57.

Zeldovich KB, Berezovsky IN, Shakhnovich EI. 2007. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol. 3(1):e5.

Zuckerkandl E, Derancourt J, Vogel H. 1971. Mutational trends and random processes in the evolution of informational macromolecules. J Mol Biol. 59(3):473–490.

Associate editor: Tanja Stadler