Positive Selection Differs between Protein Secondary Structure Elements in *Drosophila*

Kate E. Ridout, Christopher J. Dixon, and Dmitry A. Filatov*

Department of Plant Sciences, University of Oxford, Oxford, United Kingdom

*Corresponding author: E-mail: dmitry.filatov@plants.ox.ac.uk.

Accepted: 22 February 2010

Abstract

Different protein secondary structure elements have different physicochemical properties and roles in the protein, which may determine their evolutionary flexibility. However, it is not clear to what extent protein structure affects the way Darwinian selection acts at the amino acid level. Using phylogeny-based likelihood tests for positive selection, we have examined the relationship between protein secondary structure and selection across six species of Drosophila. We find that amino acids that form disordered regions, such as random coils, are far more likely to be under positive selection than expected from their proportion in the proteins, and residues in helices and β-structures are subject to less positive selection than predicted. In addition, it appears that sites undergoing positive selection are more likely than expected to occur close to one another in the protein sequence. Finally, on a genome-wide scale, we have determined that positively selected sites are found more frequently toward the gene ends. Our results demonstrate that protein structures with a greater degree of organization and strong hydrophobicity, represented here as helices and β-structures, are less tolerant to molecular adaptation than disordered, hydrophilic regions, across a diverse set of proteins.

Key words: positive selection, protein secondary structure, dN/dS ratio, *Drosophila*.

Introduction

Factors affecting the rates of evolution in protein-coding regions have long been studied by evolutionary biologists. Rates of evolution vary not only between proteins but also between different sites within a single protein, and many factors have been proposed to account for this variation, such as distance from functional sites (Dean et al. 2002), base composition (Bernardi 2005), codon usage (Bulmer 1991; Bernardi 2005; Holloway et al. 2008; Yang and Nielsen 2008), and degree of solvent exposure (Hughes and Nei 1988; Benach et al. 2000; Bishop et al. 2000; Dean et al. 2002; Lin et al. 2007). Functional residues are often the most conserved regions of the protein (Benach et al. 2000; Dean et al. 2002; O'Farrell et al. 2008), and solvent-exposed residues are the most changeable. Regions of the amino acid chain that are buried in the protein do not evolve freely (Lin et al. 2007), whereas disordered regions of the protein tend to evolve more rapidly (Brown et al. 2002). However, the action of positive selection in the protein tends to be more complex. In functional regions, for example, those involved in protein–protein interactions, certain residues may be highly conserved, or the region might comprise a patch of residues, in which the surrounding physiochemical properties rather than the exact residues are critical (Binkowski and Joachimiak 2008; Bouvier et al. 2009).

Protein secondary structure, the physical arrangement of the amino acid chain produced mainly by the amino acid sequence, is another factor that may contribute to varying rates of evolution at different amino acid positions. The amino acid order directly affects protein folding, and therefore tertiary structure and function, and is highly conserved between homologous proteins. It is known that different secondary structures have different physical and chemical properties and roles in the protein. Although this would suggest that protein secondary structure may be involved in determining rates of evolution, this question has not fully been explored, and existing investigations have been on a small scale (Benach et al. 2000; Dean et al. 2002; Hanada et al. 2006; Petersen et al. 2007), where results were specific to a particular protein domain or family. However, it is known that the type of protein secondary structure (i.e., α-helix, β-sheet, or coil) affects base composition, amino acid frequency, and even substitution rates in mammals (Chiusano et al. 1999). There is therefore good reason to suspect that

protein secondary structure plays a role in determining site-specific rates of evolution. To investigate this possibility, a large-scale genomic study is required, using source organisms with well sequenced, mapped and annotated genomes.

The publication of complete genomes from 12 closely related species of fruit fly (Drosophila 12 Genomes Consortium 2007) provides a valuable comparative resource in which to study the action of natural selection. Similarly, the wealth of knowledge about these organisms facilitates the biological interpretation of any observed trends. Using this data set, Larracuente et al. (2008) investigated and reviewed the many factors that can affect the variation in rates of evolution between different proteins in Drosophila. These included gene expression, essentiality, intron number, intron and protein lengths, protein–protein interactions, recombination, and translational selection. These factors were shown to act by either increasing the rate of adaptive evolution or by imposing evolutionary constraints. Because selection was calculated for whole proteins, secondary structure was not included and has remained largely unexplored.

Secondary structures are traditionally separated into two types—ordered regions and aperiodic/unstructured regions. The ordered regions form two main structures, helices and β-structures, whereas the aperiodic regions can be divided into random coils—natively unstructured stretches of the amino acid chain—and turns (or loops), which are amino acid chain reversals, usually containing one or more hydrogen bonds (Shepherd et al. 1999; Marcelino and Gierasch 2008).

The arrangement of an amino acid chain into a secondary structure is based on both the residues in that chain and the surrounding environment. Although particular amino acids are more frequent in different structures, these correlations are weaker than previously thought, and neighboring residues (in sequence or in space) are important in determining secondary structure (Beck et al. 2008).

The likelihood of positive selection to alter an amino acid at a given site may depend on several factors: the physical and chemical nature of the amino acid (will the replacement interact favorably with the surrounding residues and environment without damaging protein function?), the functional importance of the site (how critical is it that the exact residue or a physiochemically similar residue is maintained?), the surrounding environment (does the residue or comprising structure require a specific range of hydropathy?), the physical properties of the structure (degree of order), and the folding properties of the structure. These restrictions on the occurrence of positive selection are complex and not all of these can be analyzed with the data available.

It has been found that the most variable regions of a protein are on the solvent accessible surfaces (Lin et al. 2007) and are therefore likely to include a high proportion of hydrophilic residues. The weak correlation between secondary structure and the frequencies of different amino acids, which each have different hydropathies based on their side chain charge, means that the four secondary structure categories have different likelihoods of containing hydrophobic or hydrophilic residues (Chou and Fasman 1974). In particular, β-turns often contain hydrophilic residues (Marcelino and Gierasch 2008) and are thought to sit on the outer (solvent exposed) surfaces of the protein where they might play a role in protein folding and protein–protein interactions (Shepherd et al. 1999; Marcelino and Gierasch 2008). We might therefore expect to see an increase in both positive selection and purifying selection, as both conservation and adaptation of these residues is important. β-strands (a common type of β-structure) often contain the most hydrophobic residues, and these hydrophobic interactions are the predominant factor that stabilizes β-sheets (Chou and Fasman 1974; Koehl and Levitt 1999), which are therefore often buried in the protein core. β-strands may therefore contain less positively selected sites than the other structures. Helices are amphipathic overall (Chou and Fasman 1974), and may therefore occur anywhere in the protein, with one side of a helix often being hydrophobic and the other side hydrophilic, although, like β-strands, helices can form hydrophobic bundles in the protein core. The numbers of positively selected sites is therefore likely to be greater in helices than β-strands but less than in β-turns. Helices and β-strands are the most rigidly structured types of secondary structure and should therefore contain fewer positively selected sites than β-turns and coils because a greater proportion of potential mutations would be disruptive to the secondary structure. Indeed, several amino acids are known to break the structure of helices and β-strands in their native state (Chou and Fasman 1974; Beck et al. 2008). The other type of β-structure examined here, the β-bridge, is not expected to differ significantly from β-strands. Finally, random coils (unstructured regions) are by definition free of the structural interactions necessary for other secondary structures; they are therefore less likely to have constraints on hydropathy, position in the protein, or amino acid composition. Differences in rates of selection between secondary structures may have profound effects on protein evolution and therefore on phenotypic change. Understanding the degree to which secondary structure determines the amount of positive selection will help to explain the general patterns of evolution and uncover a previously neglected level at which natural selection may act between the amino acid and the protein levels.

Here, we infer positive selection in a phylogenetic framework (using the ratio of nonsynonymous to synonymous substitutions, dN/dS; Hughes and Nei 1988) across six species of Drosophila, using a data set of c. 8,500 genes published by Larracuente et al. (2008). We also analyze

the distribution of secondary structures and selected sites along the length of a gene and investigate how it affects the degree of positive selection. Finally, we examine the levels of hydropathy for sites and structures undergoing positive selection to build an overall picture of how evolution is influenced by protein secondary structure. We demonstrate that within this diverse range of proteins, residue changes characterized as being positively selected are distributed unevenly among protein secondary structures.

Materials and Methods

Data Acquisition

Aligned nucleotide sequence data were obtained from the published genomes of 12 species of Drosophila (Clark et al. 2007). Following the methods used by Larracuente et al. (2008), genes that exist as single-copy orthologs in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, and *D. ananassae* were selected for analysis. Saturation in silent site divergence outside the *melanogaster* species group precludes the use of all 12 genomes (Larracuente et al. 2008). Drosophila sex chromosomes evolve at different rates to autosomes, with lower levels of polymorphism and faster divergence (Begun et al. 2007) and were therefore excluded. Masked nucleotide alignments (i.e., alignments from which uncertain sections have been removed) from the six species in the *D. melanogaster* group were downloaded from the FlyBase FTP site (ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/alignments/melanogaster_group.guide_tree.longest.cds.masked.tar.gz). Following Larracuente et al. (2008), all sites in the aligned sequences with gaps or ambiguous sites in more than one of the six sequences were removed. In addition, we also reanalyzed the same data after exclusion of all sites with gaps present in any of the six species. This has not affected the conclusions of the paper. Any genes whose length varied between the two data sets were then excluded, leaving a total of 8,492 genes for our analyses. Because different alignments can produce different outcomes in phylogenetic analyses (Wong et al. 2008), we realigned all the genes with ClustalW (Thompson et al. 1994), DIALIGN-TX (Subramanian et al. 2008), and MUSCLE (Edgar 2004) using the default options and used the resulting alignments in addition to those obtained from Larracuente et al. (2008). As the results presented below are robust to the choice of alignment software, we used the alignments obtained from Larracuente et al. (2008).

Determination of Secondary Structure

All protein structure sequences (145,944 at the time of writing) from the RCSB Protein Data Bank (PDB) were downloaded and aligned against the 8,492 Drosophila genes using the National Center for Biotechnology Information's Blast (BlastX) with an expectation *E* value cutoff of 10^{-6}. For every match, the top hit from each Blast run was taken. In total, 1,092,117 experimentally determined structure residues aligned to portions of 3,884 genes.

In addition to the experimentally determined structural data, we used computational methods to predict secondary structures for our data set. *Drosophila melanogaster* has the best-characterized genome of any of the 12 Drosophila species; we therefore chose this model organism for the secondary structure prediction. Because the other sequences were aligned to the *D. melanogaster* genome, any section of the alignment where the sequence for *D. melanogaster* was unavailable would be unreliable and was excluded from further analyses.

PSIPRED (Jones 1999; Bryson et al. 2005) was used to predict secondary structures. PSIPRED uses neural networking and searches for homologous proteins with known structures to determine the most likely structure at each residue position. The homology information is collected using PSI-Blast and is combined with individual properties of the amino acids for creating or breaking different secondary structures and the likely structure lengths. Local sequence information is incorporated using a sliding window approach. Many of the most reliable secondary structure prediction methods available use neural networking in combination with Blast or PSI-Blast searches (Montgomerie et al. 2006). Results obtained during testing using the CASP3 project (Critical Assessment of Techniques for Protein Structure Prediction experiment) demonstrated that the PSIPRED method was the most accurate at that time, achieving a score of nearly 80%, the highest of all programs tested (Moult et al. 1997). Since these tests, PSIPRED has continued to be used for further developing structure prediction (Zhang et al. 2008) and remains a leading secondary structure prediction program (Birzele and Kramer 2006).

PSIPRED reports the probabilities for each site of falling into each of the three structural categories, based on the DSSP structure definitions (Kabsch and Sander 1983): helix, which contains both the α-helix (DSSP code H) and the 3_{10} helix (DSSP code G); strand, which contains β-sheets (DSSP code E) and isolated β-bridge residues (DSSP code B); and finally coil (all remaining DSSP codes including β-turns). We used the probabilities of each of these states rather than the single most likely structure in order to incorporate the uncertainty of the structure prediction method.

PSIPRED classifies hydrogen-bonded turns and natively unstructured regions together as "coils." In order to tease apart these two structural classes, we used the probabilities given by PSIPRED in conjunction with the predictions made by the neural networking program BTPRED (Shepherd et al. 1999). BTPRED takes the secondary structure predictions produced by PSIPRED and can predict whether or not a residue is in a hydrogen-bonded β-turn with an accuracy of over 70% (Kaur and Raghava 2002), although it has

a tendency to overpredict β-turn residues (Shepherd et al. 1999). BTPRED predicts whether a site is more likely to be a β-turn or a coil and provides a "reliability index"—the amount by which the predicted structure is more likely than the alternative, in tenths. The probability assigned by PSIPRED to the "coil" class was therefore divided between β-turn and natively unstructured, according to the probability derived from BTPRED. In the few cases where BTPRED's chosen prediction was actually the less likely of the two (reliability index = "*"), both were considered equally likely. The probability was therefore used as a conditional probability of BTPRED's prediction being true, given that the structure was considered a coil by PSIPRED.

Inference of Positive Selection

The program codeml from the phylogenetic analysis package PAML 4.0 (Yang 2007) was used to infer sites that have experienced positive selection, based on the ratio of nonsynonymous nucleotide changes per nonsynonymous site to synonymous changes per synonymous site (dN/dS = ω) at each codon. Synonymous changes are assumed to be functionally neutral (Kimura 1968). The program assumes a certain number of classes to which sites are assigned depending on the calculated value of ω. We used the default parameters and two pairs of nested models: M1a/M2a and M7/M8. In each case, the more general model differs from the other only in allowing an additional class of sites with $\omega > 1$, that is, sites under positive selection. Thus, a likelihood ratio test (LRT) between such nested models is explicitly testing whether the gene is under positive selection. Model M1a (Yang et al. 2000) has only two classes—one where ω is between 0 and 1 (negative selection) and one where $\omega = 1$ (neutral evolution)—whereas model M7 (Yang et al. 2000) has 10 classes with the value of ω for each following a β distribution between 0 and 1. The models M2a and M8 are similar to M1a and M7 but both include an extra class of codons with $\omega > 1$ to accommodate positively selected sites (Yang et al. 2000). We used the robust Bayes empirical Bayes procedure (Wong et al. 2004; Yang et al. 2005) implemented in PAML to detect individual sites under positive selection in genes identified by the LRT. PAML gives a probability of each site belonging to each class, and the probability for the class where $\omega > 1$ is therefore the probability that the site is under positive selection, which we will call P_s. Sites in genes with significant LRT, $\omega > 1$, and $P_s \geq 0.9$ are considered to be positively selected. In the experimentally determined structure data set, we also used lower threshold, $P_s \geq 0.5$, to increase the number of sites available for analysis. This might have increased the number of false positives in the data, therefore, wherever possible, the $P_s \geq 0.9$ threshold was also used.

The greater complexity of model M8 is likely to better fit the situation in nature but explicitly including a class where $\omega = 1$ in M2a can allow sites evolving under weak positive selection or neutral evolution to fall into this class instead of the class under positive selection. This conservative approach is particularly appropriate for analysis with few taxa (Anisimova et al. 2002). By using both models to search for the same underlying trends, we hope to avoid any specific effects of individual models and thus provide stronger support for any results found (Anisimova et al. 2002).

Because different genes may follow different gene trees, each gene was analyzed using the most appropriate tree topology for that gene. The tree that provided the best result for each gene is listed at ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/paml. Over the 8,492 genes, three trees were used, differing only in the placement of two species, *D. erecta* and *D. yakuba*, for which there is known discordance between gene trees and species trees (Pollard et al. 2006).

It was suggested by Lindsay et al. (2008) that codon models used to estimate ω might be affected by sequence composition. More recently, Yap et al. (2009) demonstrated that this is indeed the case for such models, for example, the Goldman and Yang method (GY) used by PAML. The GY model uses continuous time Markov processes to model substitutions (in order to estimate ω), the rates of which are specified by an instantaneous rate matrix, the parameters being based on rates of codon change (in this case in the gene). This matrix is then weighted by the frequency of the codon being changed to rather than the frequency of the nucleotide being changed to. Thus, if sequence composition varied between secondary structures, the rate assumptions made by the codon model would be violated, making them unsuitable. Lindsay et al. (2008) suggested that models which weight substitutions by nucleotide frequencies, such as the MG model (Muse and Gaut 1994), are more robust to nucleotide composition than the GY model.

The models used for the PAML analysis, M1a/M2a and M7/M8, all use the GY method. It is therefore possible that ω might vary between structures based on their sequence composition. To gain a better understanding of any effects of this bias in ω on our data, the following simulations were run: Sequences were simulated with PyCogent (Knight et al. 2007), under the MG codon substitution method. The rate parameters for the substitution matrix (i.e., transition/transversion rates and divergences between species) were taken from the concatenation of all 8,492 genes used in our analyses. One large gene was simulated for each of the four structures, where the nucleotide frequencies used to simulate each gene were taken from the overall proportion of a given nucleotide in a particular structure in the real data set (e.g., the proportion of thymine nucleotides in all helix structures in the 8,492 genes). Two sets of simulations were run; one with ω equal to 1 in all structures and another with the average ω from the real data, 0.26. Using the

distributions of gene lengths and structure lengths found in the 8,492 genes, 1,000 genes for $\omega = 1$ and 1,000 for $\omega = 0.26$ were simulated. In the simulation, different secondary structure elements will evolve equivalently, so long as nucleotide composition varying between the structures has no effect on the result. Therefore, if the GY method is unbiased in this instance, there should be no difference in the proportion of positively selected sites between the four structure classes (when the simulated data was analyzed by codeml, a program within PAML, to search for positive selection). Though this analysis gives us a better understanding of whether protein secondary structure over the entire data set varies enough in general sequence composition to confound our results, it is not definitive. Due to nucleotide and codon composition potentially varying between secondary structure elements (Chiusano et al. 1999) and the different structural compositions of genes, it is possible that this effect may still confound the results.

We also investigated the degree of codon bias in different structures because certain secondary structures may use rare codons preferentially, in order to slow translation down, and thereby aid protein folding (Komar 2008). An excess of rare codons in any of the secondary structures could lead to a reduction in the synonymous substitution rate, decreasing dS, which could artificially increase ω. To test whether variation in codon bias across the structures could affect synonymous substitution rate and hence estimates of ω, we compared the effective number of codons (Wright 1990) in the four secondary structures.

Amino acid content may vary between structures; it is therefore possible that differing rates of selection in amino acids might lead to the difference between the secondary structures. To examine the link between amino acid content and positive selection in a structure, the amino acid and the predicted structure at each selected position from the *D. melanogaster* lineage was recorded. The secondary structure each amino acid belongs to was also recorded at all sites. The fraction of selected sites was calculated for each amino acid by dividing the number of selected sites of an amino acid by the total number of sites of that amino acid (regardless of structure). The expected number of sites under selection for a particular amino acid in a structure was then estimated by multiplying the fraction of selected sites for each amino acid by the observed number of that amino acid in each structure. This number was compared with the observed numbers of selected sites for each amino acid in all four structures.

Hydropathy in Selected and Nonselected Sites

Because amino acids with different hydropathies can favor different secondary structures (Chou and Fasman 1974), a better understanding of how the likelihood of selection varies with secondary structure might be gained by looking at the changes in hydropathy resulting from changes between the current amino acid and its ancestral state at selected sites. Changes in hydropathy, measured with the hydropathy index of Kyte and Doolittle (1982), at selected sites and nonselected sites were calculated for each of the four structures in the predicted structure data set using the amino acids corresponding with the ancestral nucleotide sequence reconstructed by PAML (marginal reconstruction) from the M8 analysis. The mean hydropathy was also calculated from the current amino acids for each of the four structures in all sites and at selected sites. Variation in hydropathy along the length of a single protein could lead to a bias in the amino acids and hence relative proportions of the secondary structures found at different positions in a protein. To test for this possibility, each gene was divided into 20 equal segments and the mean hydropathy of the amino acids calculated for each.

The distance of a residue in a protein from the periphery and the core of a protein has an effect on the likelihood of positive selection (Lin et al. 2007). In addition, the likelihood of a secondary structure to be solvent exposed and therefore in the exposed peripheral residues of the protein varies due to the intrinsic amino acid content of each secondary structure (Chou and Fasman 1974). To explore this link, we used experimentally determined structures from the PDB to produce an independent estimate of how often different secondary structures are present on the exposed surfaces of proteins. All structures reported for *D. melanogaster* were examined, with duplicates (proteins that displayed over 95% sequence similarity) excluded. In total, 160 proteins were available. The solvent-exposed areas of each structure from this random data set were calculated. Secondary structures were taken directly from the PDB, and solvent accessibility was calculated using maximal speed molecular surface (Sanner et al. 1996).

Spacing of Selected Sites

Distances between selected sites were recorded along each gene. To test whether any clustering of selected sites was due to the different proportions of positively selected sites in different secondary structures, rather than directional selection, we ran the following simulation: Data were simulated using the known proportions of selected sites in different secondary structures and the observed length distributions of secondary structures in the data set as a whole. The length distributions of the different structures were recorded from the 8,492 Drosophila genes by taking the most likely of the four structures (from the combined PSIPRED and BTPRED structure predictions) to be the absolute structure at each residue. For the simulation, lengths of structures were chosen randomly from the observed distribution, without replacement. Each site was given the appropriate structure-specific probability of being under positive selection. The intervals from one selected site to the next were

recorded. Simulations were performed in Java and repeated 5,000 times in order to provide confidence limits on the observed frequencies.

We also tested whether the clustering of sites under positive selection is due to changes on the same or different branches of the phylogeny. The proportion of parallel changes that we might expect to observe on a single branch of the phylogeny depends on the branch lengths in the tree because the probability of a second substitution occurring on the same branch is equal to the branch's length as a proportion of the entire tree length. The overall proportion of parallel substitutions is therefore $\Sigma(L_i^2)$, where L_i is the length (as proportion of the whole tree) of the ith branch. These lengths were approximated using the PAML ancestral state reconstructions to count the occurrences of an amino acid at a selected site changing along each branch of the tree. Pairs of adjacent selected sites along a given gene were categorized by the number of amino acids between them, with distances above an arbitrary boundary of 30 amino acids being considered large and those below 30 amino acids considered small. The proportion of pairs of adjacent selected sites which experienced substitutions on the same branch of the tree was compared against the expectation, for both large and small distances. It is possible that both the observed and expected results are slightly underestimated, as multiple changes in the same position on the same branch cannot be detected. However, both results are calculated using the same method, and therefore, we do not expect this to introduce a bias.

Selection in the Ends of the Genes

To investigate whether the proportion of sites under selection is influenced by the position within the gene, every gene was arbitrarily split into 20 equal segments, and the number of selected and nonselected sites were counted in each computationally predicted structure and in each segment. Using the M8 data, the ω values of every residue were plotted for each of the five gene segments to see if the distribution of ω varies with position in the gene. A skew of ω values to be closer to 1 at the ends of genes would indicate a relaxation of purifying selection in these regions. In addition, if the ends of the gene experience a relaxation of purifying selection then in the context of the whole gene these regions would be more likely to be picked up as positively selected sites. To examine this possibility, each gene was manually divided into two parts. One contained the first 15% of the gene, concatenated with the last 5%, as these regions encapsulate the gene segments with the sharpest increase in the proportion of positively selected sites. The second part comprised the remainder (the central part) of the gene. PAML model M2a was run twice for every gene on the two parts separately to determine whether there was a difference in strength of positive selection, number of positively selected sites, strength of purifying selection, and number of sites under purifying selection between the gene ends and the gene center.

Indels within Structures

Pascarella and Argos (1992) demonstrated that insertions and deletions (indels) were enriched in reverse turn and coil structures. Indels occurring in one or more species in our data were removed; despite this, the remaining indels and the areas that surround previous indel sites may represent areas of increased alignment ambiguity. This could potentially lead to a perceived increase in substitution rate and therefore a high false inference of positive selection in these regions, predominantly at turn and coil sites, where we would expect the most indels. To test whether regions surrounding indels bias the proportion of selected sites found in each structure, we examined the distribution of selected sites around each indel within a predicted structure. Each gene was examined individually, for each section of a structure along the length, the distance from every site, and every selected site to the nearest indel was recorded. As all sequences were aligned against *D. melanogaster*, the gaps in this species of the raw data (as downloaded, before any gaps were removed) was used.

Availability of Programs

All the data processing and manipulation was automated using Perl, Python, and Java programs, which are available on request.

Results

Selection in Secondary Structures

Among the 3,884 genes for which experimentally determined structure data was available, a total of 1,092,117 residues were included. Selected sites identified using model M8 at $P_s \geq 0.5$ from genes with a significant M7/M8 LRT were not randomly distributed among the four experimentally determined structures (χ^2 test: $P < 0.00001$), with strands and β-turns containing fewer residues undergoing positive selection than would be expected by chance (0.53 × expectation and × 0.57, respectively). Coil regions contained more positively selected residues than expected by chance (× 1.83) and helix regions slightly less (× 0.95). Similar results were obtained using the M1a/M2a LRT (fig. 1*A*). The results did not qualitatively differ using M8 model with $P_s \geq 0.9$ threshold for positively selected sites: strands, β-turns, and helix regions contained fewer selected residues than expected (× 0.24, × 0.43, and × 0.75, respectively) and coils more (× 4.19).

The data set of computationally predicted secondary structures comprised 8,492 genes, with a total of 4,125,829 aligned residues (table 1). Similarly to the experimentally determined structures, the distribution of selected

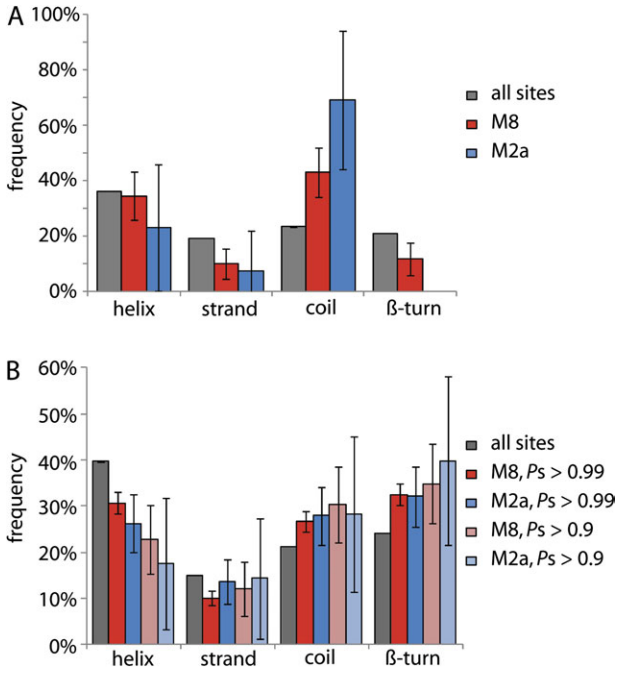

Fig. 1.—Proportions of all sites (gray bars) and positively selected sites (colored bars) according to the M1a/M2a and M7/M8 LRT in different secondary structures determined experimentally (*A*) and predicted computationally (*B*). A threshold probability of $P_s \geq 0.5$ was used in (*A*) and two thresholds ($P_s \geq 0.9$ or ≥ 0.99) were used in (*B*).

sites was not random (χ^2 test: $P < 0.00001$ where $P_s \geq 0.9$ and $P = 0.000176$ where $P_s \geq 0.99$, using the M8 model in genes with a significant M7/M8 LRT). Strands were less likely to contain positively selected sites than expected (× 0.67); however, β-turns contained more positively selected sites than expected (× 1.35). It was also observed that coil regions contained more selected sites than expected (× 1.26) and helix structures slightly less (× 0.77) (fig. 1*B*). Again, the results were not qualitatively different using genes identified as being positively selected by the M1a/M2a LRT to determine selected sites. For both M8 and M2a models, the results were similar at the threshold values $P_s \geq 0.9$ and $P_s \geq 0.99$. When examining only sites with $P_s \geq 0.99$, β-turns (× 1.45) and coils (× 1.43) again have more selected sites than expected, whereas helices (× 0.57) and strands (× 0.80) are both underrepresented (fig. 1*B*).

Yap et al. (2009) demonstrated that estimates of ω are affected by nucleotide composition. If composition varies between the four structures, the assumptions made by the codon models used in this study would be violated, confounding the results. To test whether nucleotide composition heterogeneity could lead to the observed differences in positive selection between secondary structures, we analyzed simulated data sets generated in such a way that the only difference between the regions with different structure was their nucleotide composition. There should therefore be no significant difference in the number of positively selected sites between the four structure classes, unless the difference is due to nucleotide composition. Indeed, no such difference was detected (table 2), confirming that nucleotide composition differences between the regions encoding different protein secondary structures are unlikely to cause the observed difference in the number of positively selected sites.

If the data set contained an excess of rare codons or strong codon bias, particularly in one structure over the others, this could lead to decreased values of dS (the number of synonymous mutations at synonymous sites) when compared with dN (nonsynonymous mutations at nonsynonymous sites). This decrease in dS relative to dN could cause the artificial inflation of ω (dN/dS) and hence the false inference of positive selection. To investigate this possibility, we examined codon bias in the four structures. There was a difference in the effective number of codons (Wright 1990) used in the different structures. The two ordered structures, helices and strands, had values of 50.47 and 50.73, respectively, whereas the aperiodic regions showed weaker codon bias (β-turns: 52.48; coils: 52.56). This is the opposite to what we would expect if stronger codon bias was inflating the signal of positive selection in β-turns and coils.

Observed and expected (see Materials and Methods) rates of selection for each amino acid in every structure were compared with test whether biased positive selection of

Table 1
Positively Selected Sites in Secondary Structures

	Predicted			Experimentally Determined			
	Total Sites	$P_s \geq 0.9$	$P_s \geq 0.99$	Total Sites	$P_s \geq 0.5$	ENC	Mean ω
Helix	1,635,453.8	437.072 (0.21%)	27.315 (0.013%)	398,083	41 (0.095%)	50.47	0.135
Strand	621,585.2	143.728 (0.21%)	14.466 (0.021%)	208,798	12 (0.053%)	50.73	0.129
Coil	873,646.1	380.503 (0.54%)	36.444 (0.052%)	227,527	51 (0.175%)	52.56	0.160
β-Turn	995,143.8	462.893 (0.32%)	41.911 (0.029%)	257,769	14 (0.054%)	52.48	0.158
Total	4,125,829.0	1424.196 (0.29%)	120.136 (0.024%)	1,092,117	118 (0.098%)	51.53	0.146

Note.—Summary statistics for each of the four secondary structures, including total number of sites in each data set, along with the number of sites under selection (at both $Ps \geq 0.9$ and $Ps \geq 0.99$ for predicted structures but only $Ps \geq 0.5$ for experimentally determined structures using model M8; percentages are expressed as a proportion of all sites in genes with a significant LRT), the effective number of codons (ENC), and the mean value of ω.

Table 2
Overall Nucleotide Content of the Four Structures Taken from the 8,492 Genes and the Proportions of Selected Sites in the Four Structures Taken from the Simulated Genes

	Amino Acids				Simulated Genes		
	A	C	T	G	Total Sites	Selected	95% CI
Helix	0.247	0.261	0.218	0.274	213252	186	0.00075–0.00100
Sheet	0.220	0.253	0.275	0.251	91437	86	0.00074–0.00114
Turn	0.256	0.302	0.164	0.278	169129	185	0.00094–0.00125
Coil	0.262	0.279	0.184	0.276	45333	45	0.00070–0.00128

NOTE.—Data were collected using pyCogent, gaps were excluded. CI represents the confidence interval of the proportion of selected sites per structure.

specific amino acids could be responsible for the trend that proportions of positively selected sites vary between structures. The expected proportion of selected sites in each structure was calculated, assuming that neither structure nor amino acid content had any effect on the number of positively selected sites in each structure. These proportions were compared with the expected number of selected sites if amino acid content alone had an effect on the proportion of selected sites in each structure (data not shown). The analysis revealed that if amino acid content were the cause of the distribution of positively selected sites in secondary structure, we would expect fewer selected turn and coil sites than if the distribution was random, slightly less selected helix sites and a greater number of selected sheet sites. These results are very different from the proportions of selected sites found in the structures, where turns and coils contain more selected sites than expected and sheets less. Helices contain fewer sites under selection than expected at random, although it is significantly less than predicted by the rate of selection in the amino acids. Thus, it appears unlikely that the difference between the proportions of selected sites in secondary structures is due to different frequencies of amino acids.

Analysis of Hydropathy in Selected and Unselected Sites

Changes in amino acid hydropathy from the ancestral state to the derived state were measured at selected and unselected sites (table 3). Overall, structures show decreasing hydropathy at both selected ($P_s \geq 0.9$, M7/M8) and not selected sites. This indicates that protein hydropathy is not at equilibrium in these six species of Drosophila. In β-turns, hydropathy at positively selected sites is more conserved than at all other sites, unlike the other three structures, which show the opposite trend; hydropathy is more conserved at sites that are not undergoing selection. β-Turns are expected to extend into the solvent and therefore might be expected to have different hydropathy characteristics. In terms of overall composition, strand residues were found to be highly hydrophobic on average, β-turns and coils were strongly hydrophilic, and helices were weakly hydrophilic.

The degree of solvent exposure was calculated for the set of all experimentally determined structures deriving from *D. melanogaster* (regardless of whether we possess a corresponding sequence alignment for six Drosophila species). The corresponding experimentally determined secondary structure was then taken to determine if any structure was more likely to be solvent exposed or accessible. β-Turns are the most likely structure to be in the solvent-exposed regions (fig. 2). Coils are the next most solvent accessible, followed by helices and finally strands.

Spacing of Selected Sites

Under a purely random distribution, the distance from one selected site to the next would be expected to follow a geometric distribution because the probability of each

Table 3
Changes in Hydropathy from the Ancestral Amino Acid State to the Derived State

	Selected	Not Selected
Helix	−0.001193	−0.0006577
Strand	−0.001358	−0.0005468
Coil	−0.001322	−0.0012434
β-Turn	−0.000380	−0.0014548
All	−0.000994	−0.0009550

NOTE.—Mean changes in hydropathy from the PAML reconstructed ancestral state per amino acid substitution for each secondary structure, at selected ($Ps \geq 0.9$) and at not selected ($Ps < 0.9$) sites using the model M8.

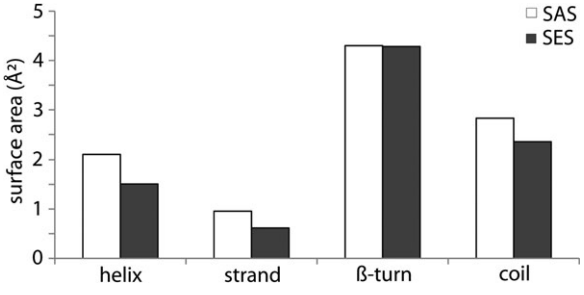

FIG. 2.—Mean solvent exposed (SES, black bars) and solvent accessible (SAS, white bars) areas, expressed in square ångströms, in each of the four secondary structures.

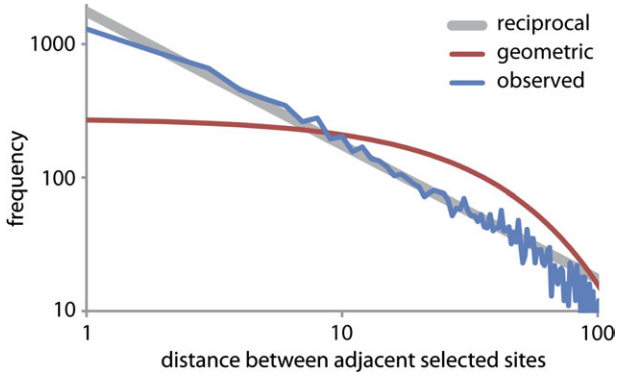

Fig. 3.—The size distribution of intervals between adjacent selected sites on a log–log scale, with the geometric curve expected given no clustering of sites, and a power law fitted to the curve at lower gap sizes.

subsequent site being under positive selection will be equal. We observe a large departure from this expectation, with the likelihood of being under positive selection decreasing with increasing distance from other selected sites. After one selected site, the next selected site was more likely to be encountered within the following 30 amino acids than expected by chance (fig. 3). This result was significant ($P < 0.0001$) in genes with a significant LRT under all examined combinations of models and threshold values (M1a/M2a: $P_s \geq 0.9$ and M7/M8: $P_s \geq 0.9$, $P_s \geq 0.99$).

To test whether this effect was due to the different rates of selection in different secondary structures, data were simulated using the known proportions of selected sites in different secondary structures and the observed length distributions of secondary structures in the data set as a whole. These simulations showed only a slight deviation from the expected geometric distribution, equivalent to a small increase of 15% in the frequencies of positively selected sites 1 residue apart, compared with the 5-fold increase observed in our data. Thus, the observed clustering of positively selected sites cannot be explained by different rates of selection in different secondary structures.

There are at least two possible explanations for clustering of sites under positive selection. One possibility is that a given gene region may be particularly prone to positive selection forming evolutionary "hotspots." On the other hand, selection-driven change at one site may cause an increase in selection at nearby sites, such as compensatory mutations. We can distinguish between these two types of process by observing where on the species phylogeny amino acid changes at selected sites occur. Compensatory mutations should cause adjacent selected site to evolve in concert, on the same branch of the tree. If, on the other hand, selective hotspots are responsible for the pattern, the amino acid changes should be distributed randomly across the phylogeny. We found that the proportion of amino acid changes at adjacent selected sites occurring on the same branch of the tree for smaller intervals (selected residues <30 amino acids apart) and for larger intervals (≥30 residues apart) were significantly greater than the expected values (table 4). Thus, sites under selection within an individual gene are more likely to occur on the same branch of the gene tree than different branches. This result was stronger where sites were closer together (<30 amino acids) and where a more stringent threshold of positive selection was used.

Selection in the Ends of the Genes

When dividing each gene into sections of equal length, secondary structures were found to vary in frequency along the length of a gene (fig. 4), with the beginning of a gene and, to a lesser extent, the end, showing a significant decrease in strands ($P < 0.0001$). In addition, there is a significant increase of positively selected sites at both ends of the gene (fig. 5). There was not sufficient data to determine whether this variation at the gene ends was due to the increased number of selected residues at β-turn and coil sites and the increase in β-turns and coils at the ends of genes. However, this is unlikely to account for the entirety of the variation, as the increase in β-turn and coil residues is far smaller than the increase in selected sites at the ends of genes, suggesting that at the ends of the gene there is an additional change to the selective pressures. This result does not differ qualitatively between the two PAML model comparisons nor between different threshold values. Exclusion of sites with alignment gaps reveals the same pattern: the number of selected sites is still increased at the N and C termini of the genes (supplementary fig. S1, Supplementary Material online).

The distribution of ω along the length of a gene is shown in figure 6. Mean ω is inflated in the first 15% and the last 5% of a gene. Values of ω closer to 1 may be explained by

Table 4
Proportion of Adjacent Amino Acid Changes Occurring on the Same Branch of the Phylogeny

Number of Observations	>30 A.A.	≤30 A.A.	>30 A.A.	≤30 A.A.	Expectation
M2a, $P_S \geq 0.9$	80	468	52.63% (41.96–63.30%)*	76.22% (71.69–80.75%)*	13.6%
M8, $P_S \geq 0.9$	488	949	47.42% (43.32–51.53%)*	60.56% (57.31–63.81%)*	14.6%
M8, $P_S \geq 0.99$	17	352	58.62% (34.52–82.72%)*	83.61% (78.86–88.36%)*	13.8%

*Note.—99.15% confidence interval (equivalent to 95% CI but after Bonferroni correction for multiple testing). Distances between adjacent selected sites are significantly different from the expected values in all counts.

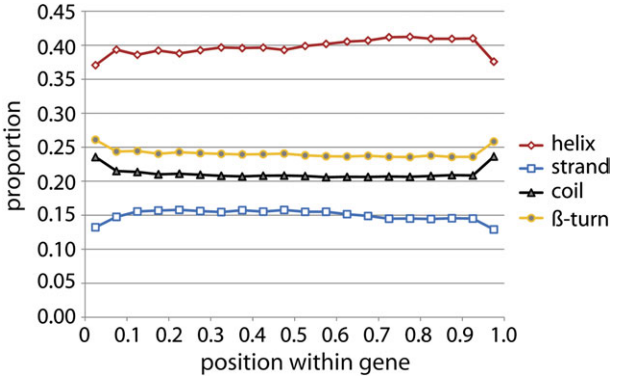

Fig. 4.—Variation in the frequencies of different secondary structures along the length of genes when divided into 20 equal segments at all sites.

either of two phenomena: a relaxation of purifying selection or an increase of positively selected sites. Due to the relative scarcity of positively selected sites compared with the number of sites under purifying selection, the distribution of ω, where $\omega < 1$ in each of the four structures provides an indication of the strength of purifying selection. No significant differences in average ω (where $\omega < 1$) were determined between the four structures (data not shown). It is therefore unlikely that the different proportions of selected sites found between secondary structures are due to relaxed purifying selection that has been mistaken for positive selection.

Partitioning the results obtained from running model M2a on all genes into the ends of the genes (the first 15% and the last 5%) and the middle (the remainder) revealed that the distribution of $\omega > 1$ (positive selection) is not significantly different between the middle and the ends of genes ($P = 0.65$, unpaired *t*-test). However, the frequency of positively selected sites at the ends of genes was significantly greater than in the middle ($P < 0.0001$) (supplementary fig. S2a, Supplementary Material online). The number of sites under purifying

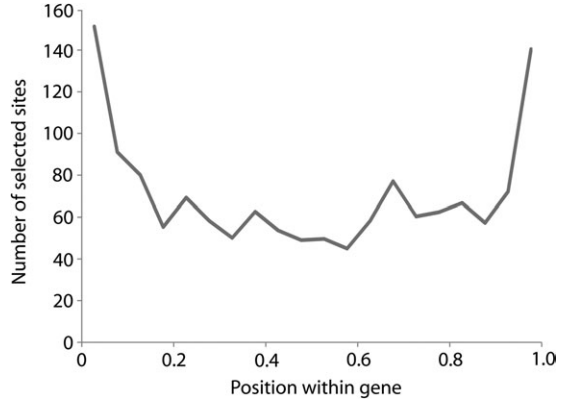

Fig. 5.—Variation in the number of sites under selection along the length of a gene, when divided into 20 equal segments. Zero marks the start (N-terminus) and 1 the end (C-terminus).

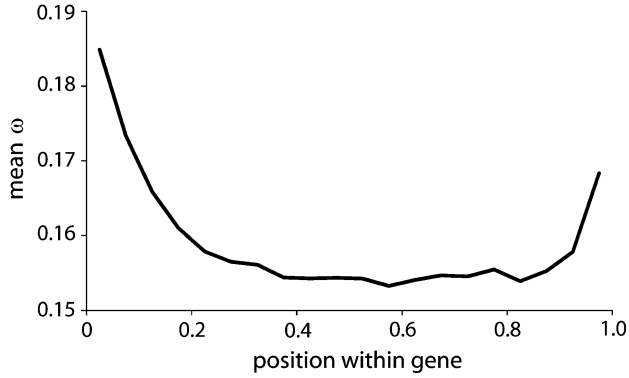

Fig. 6.—Graph of mean ω against position in the gene. Data were binned into 20 equal segments along the gene.

selection was significantly lower ($P < 0.0001$) at the ends of the genes than in the middle, as more sites were under positive selection (supplementary fig. S2b, Supplementary Material online). In addition, the distribution of $\omega < 1$ (purifying selection) was significantly skewed toward 1 and therefore weaker at the ends of the genes ($P < 0.0001$).

Indels within Structures

Indels may potentially affect the number of positively selected sites identified in a region, as they introduce some uncertainly into alignments. Examining the distribution of indels in different secondary structures reveals that β-turn structures contain fewer indels than we would expect to see, whereas all other structures contain more indels than the expected value (if indels were equally distributed between structures—data not shown). If the abundant of indels in β-turns were causing the increase in the proportion of selected residues in these regions, we would expect to see the opposite result; thus, indels are unlikely to be the cause of the unequal distribution of positively selected sites between secondary structures.

Discussion

Previous studies of positive selection in secondary structure have examined single genes or domain families (Mondragon-Palomino et al. 2002; Kosiol et al. 2008). The results of these analyses each tell us something about the evolution of a specific protein or protein family, but though thorough studies exist to explore many factors affecting the rate of evolution (Larracuente et al. 2008), no such studies have yet been conducted to examine the relationship between positive selection and secondary structure on the genomic scale. One recent study examined the correlation between single nucleotide polymorphism and secondary structure (Liu et al. 2008). Solvent-exposed regions were the least conserved, whereas helices and strands were under stronger purifying selection, although the effects of

positive selection were not analyzed. Those studies that have discussed the structures in which selection occurs (Alvarez-Valin et al. 2000) have not had the power to determine differences in selection between secondary structures. This is particularly important where the variability of amino acid residues is used as a proxy to determine sites of functional importance. For example, Thomas et al. (2003) specifically use conserved regions of coding sequence to infer functionality. However, our results show that different structures (particularly strands) are also likely to produce regions where the amino acids are strongly constrained. In this case, it would be useful to examine the structural composition of the region to determine if this is the case. It would appear from previous studies that regions of functional importance which must adapt quickly (e.g., virus-binding regions) contain more positively selected sites (Kosiol et al. 2008). This demonstration of how positive selection can be spatially limited along a gene demonstrates the importance of understanding why selection varies along a gene.

We demonstrate here that secondary structure has a significant effect on the rate of adaptive evolution in proteins. It appears that of the four predicted secondary structures, β-turns and coils are the most likely to experience positive selection and more periodic strands and helices the least. On the other hand, in the data set with experimentally determined structures, β-turns contained less positively selected sites than expected. This might be due to the difficulty to predict β-turns, however, neural networking methods such as PSIPRED are the most reliable methods of structure prediction currently available (Kaur and Raghava 2002). Alternatively, it might be due to the difficulty in determining the structure of disordered protein regions. Disordered regions do not have a definite 3D structure and are therefore difficult to crystallize. Thus, experimentally determined structures may not be a random sample of the Drosophila genome. As unstructured regions contain more instances of positive selection, particularly in hydrophilic areas likely to be on the outer surface of the protein, the β-turns and hydrophilic regions of structures (and thus positively selected sites) might be under-represented in the experimentally determined data set. Therefore, the β-turns that remain in the experimentally determined structures are likely to be internal to the protein and therefore behave in a similar fashion to structured regions.

Changes in hydropathy calculated from the ancestral state of an amino acid to the descendent state at both the $P_s \geq 0.9$ and $P_s \geq 0.99$ threshold levels (M7/M8) revealed that hydropathy is not at equilibrium in the 8,492 genes examined in the six species of Drosophila. The decreasing hydropathy at sites that were not identified as evolving under positive selection may suggest that additional factors not examined here play a role in shaping the amino acid sequences of proteins. Hydropathy at positively selected sites in coil, helix, and strand regions is less conserved than at all other sites, however, the opposite is found in β-turns. This suggests that β-turns might have different hydropathy characteristics to the other three structures examined here. We have also demonstrated for the first time that secondary structures are not evenly distributed along the length of the gene, there being more β-turns and coils toward the ends. Positively selected sites are also more likely to be located at the ends of the gene (fig. 5). However, the increase in β-turns and coils at the ends of genes is not sufficient to fully explain the increase of positively selected sites at the ends of genes.

Distributions of ω values for positively selected residues ($\omega > 1$) is not significantly different between the central part and the ends of the gene (supplementary fig. S2a, Supplementary Material online), although there are significantly more sites under positive selection at the ends of the genes than in the middle. When looking at codons with $\omega < 1$, it was noted that the distribution of ω was closer to 1 at the ends of the genes and there were fewer sites under purifying selection, indicating an overall relaxation in purifying selection at terminal parts of genes, relative to the middle (supplementary fig. S2b, Supplementary Material online). This reduction in purifying selection, coupled with more variable sites and less structure (more coils and β-turns), suggests that amino acids at the ends of genes are less constrained than in the middle, and there is therefore more opportunity for mutations to be positively selected.

When observing the variation of selected sites across the length of a gene, the distances between adjacent selected sites deviated from the expected distribution, with a significant excess of sites at shorter distances. It would be reasonable to assume that this clustering of selected sites is because mutations would either be compensatory or in a region of decreased conservation. Purifying selection may tolerate mutations constrained by protein structure only after certain neighboring mutations have occurred. The fact that amino acid changes at neighboring selected sites were more likely to be on the same branch of the reconstructed tree suggests that such mutations are not independent and possibly reflect compensatory evolution. A similar tendency for selection to act on nearby sites along the same branch in a phylogeny has been noted previously for mammals (Bazykin et al. 2004). It is interesting that parallel changes are detectable over such long timescales as the rat–mouse divergence or speciation within the *D. melanogaster* group. It would be interesting to study how quickly these parallel changes can occur by carrying out similar comparisons for more closely related taxa. Aris-Brosou (2005) presented the extended complexity hypothesis, discussing the nature of proteins within complex interaction networks to be more conserved by evolution. It may be possible that the observed clustering of selected sites is related to this hypothesis, which might suggest regions of conservation where interactions occur.

Varying strengths in codon bias between the structures could lead to the observed signal of more positive selection in β-turns and coils, compared with other structures. However, although codon bias does differ between the secondary structures, the difference is in the opposite direction to that which would be expected if stronger codon bias were the cause. We have also examined the possibility that a relaxation of purifying selection in certain structures might have been mistaken for positive selection. However, by examining the frequency distribution of ω (where $\omega < 1$) for each structure, we have revealed no difference between the four structures (data not shown).

In a recent study of positive selection in *Escherichia coli*, Petersen et al. (2007) pointed out that positive selection was more often found on the outside of the protein and in proteins on the outer surface of the cell. For example, external loops thought to be responsible for phage binding contain many more positively selected sites than the internal β-barrel region (composed of strands). This is a strong demonstration that regions of proteins that are in contact with external forces are a more likely target for positive selection. Our initial expectation was that the more structured regions (strands and helices) would contain fewer positively selected sites (and polymorphic sites) because they are governed by more strict rules about which residues are physicochemically acceptable than unstructured regions. For example, proline, glycine, and valine are known to break helices in their native state (O'Neil and DeGrado 1990; Beck et al. 2008). Therefore, mutations toward these amino acids might not be favorable in helical regions. In addition, the more structured regions—strands in particular—are more likely to contain hydrophobic residues (Chou and Fasman 1974; Koehl and Levitt 1999), which is consistent with our results. They are therefore less likely to be in the solvent-exposed regions of the protein and more likely to be important for protein stability (Dudgeon et al. 2008) and the prevention of protein aggregation due to hydrophobic interactions. It has also been suggested that internal residues are more important for maintaining the folding of a protein (Creighton and Darby 1989; Alvarez-Valin et al. 2000) and that external regions have lower structural constraints, again suggesting that external regions should be more susceptible to positive selection and are more robust to both synonymous and nonsynonymous polymorphism. In contrast, coil and β-turn regions are more likely to be on the outside of a protein as they do not have the same structurally induced physiochemical constraints (e.g., necessary hydrophobicity). Thus, these unstructured regions (β-turns in particular) are often hydrophilic (Marcelino and Gierasch 2008). Helices are known to be amphipathic and can contain both hydrophobic and hydrophilic residues (Chou and Fasman 1974; Koehl and Levitt 1999). Ferrada and Wagner (2008) discuss the correlation between protein robustness and evolution, they suggest that the more "designable" a protein is (the number of sequence variations that can fold into the correct structure) the greater its ability to evolve. Therefore, proteins that contain structures with more amino acid flexibility (turns and coils) might be expected to have a faster rate of evolution.

Unfortunately, the prevalence or absence of residues in different structures alone is not enough to predict protein secondary structure, and it has recently been contested that the intrinsic tendencies of amino acids for specific conformational preferences is not as strong as previously assumed (Beck et al. 2008). Our own investigations have determined that in our data set with predicted structures, β-turns are the most likely to occur in the solvent-exposed regions of the protein and are the most hydrophilic and contain the greatest number of positively selected sites. Strands occurred on the external solvent-exposed regions of the protein the least out of all the structures, were the most hydrophobic, and contained the lowest proportion of positively selected sites. Helices contained slightly more positively selected sites than strands, were slightly more hydrophilic, and slightly more likely to occur on the periphery of the protein. Finally, coils were slightly less hydrophilic than β-turns and were slightly less likely to occur on the outside of the protein. From these results, a pattern begins to emerge where the most structured regions form the complex highly folded, hydrophobic, conserved protein core that experiences more purifying and less positive selection, compared with coils and β-turns.

These results are the first of their kind to demonstrate on a genomic scale that the probability of a residue being under positive selection is dependent on the structure to which the residue belongs. We also determine that other factors, such as position along the gene, hydropathy, and distance from the closest selected site have an effect on selection. It will be important for future studies to understand exactly why selection varies along the length of the gene and to what extent all the results found in this study affect the likelihood of a site to experience positive selection. Knowing how secondary structures are selected will help to disentangle the reasons behind positive selection in a region of a protein and therefore aid the discovery of positively selected sites that may be functionally important.

Supplementary Material

Supplementary figures S1–S2 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

The authors would like to thank Amanda Larracuente for providing PAML results data, Adrian Shepherd for making BTPRED available, Charlotte Dean and Andrew Dalby for discussions and use of computing facilities, Maxim Kapralov and Graham Muir for their helpful discussions and Laurence

Hurst, Yuri Wulf, Dmitri Petrov, Shamil Sunyaev, and Gavin Huttley for suggestions that helped to improve the manuscript. This work was supported by a PhD fellowship to K.E.R. from the Biotechnology and Biological Sciences Research Council, UK and a grant to D.A.F. from the Natural Environment Research Council, UK.

Literature Cited

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol Biol Evol. 19:950–958.

Alvarez-Valin F, Tort JF, Bernardi G. 2000. Nonrandom spatial distribution of synonymous substitutions in the GP63 gene from Leishmania. Genetics. 155:1683–1692.

Aris-Brosou S. 2005. Determinants of adaptive evolution at the molecular level: the extended hypothesis. Mol Biol Evol. 22: 200–209.

Bazykin GA, Kondrashov FA, Ogurtsov AY, Sunyaev S, Kondrashov AS. 2004. Positive selection at sites of multiple amino acid replacements since rat–mouse divergence. Nature. 429:558–562.

Beck DAC, Alonso DOV, Inoyama D, Daggett V. 2008. The intrinsic conformational propensities of the 20 naturally occurring amino acids and reflection of these propensities in proteins. Proc Natl Acad Sci U S A. 105:12259–12264.

Begun DJ, et al. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. PLoS Biol. 5:e310.

Benach J, Atrian S, Fibla J, Gonzàlez-Duarte R, Ladenstein R. 2000. Structure-function relationships in Drosophila melanogaster alcohol dehydrogenase allozymes ADH^S, ADH^F and ADH^{UF}, and distantly related forms. Eur J Biochem. 267:3613–3622.

Bernardi G. 2005. Structural and evolutionary genomics—natural selection in genome evolution. Amsterdam (the Netherlands): Elsevier Science

Binkowski TA, Joachimiak A. 2008. Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites. BMC Struct Biol. 8:45.

Birzele F, Kramer S. 2006. A new representation for protein secondary structure prediction based on frequent patterns. Bioinformatics. 22:2628–2634.

Bishop JG, Dean AM, Mitchell-Olds T. 2000. Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. Proc Natl Acad Sci U S A. 97:5322–5327.

Bouvier B, Grünberg R, Nilges M, Cazals F. 2009. Shelling the Voronoi interface of protein-protein complexes reveals patterns of residue conservation, dynamics, and composition. Proteins. 76:677–692.

Brown CJ, et al. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. J Mol Evol. 55:104–110.

Bryson K, et al. 2005. Protein structure prediction servers at University College London. Nucleic Acids Res. 33:W36–W38.

Bulmer M. 1991. The selection-mutation drift theory of synonymous codon usage. Genetics. 129:897–907.

Chiusano ML, et al. 1999. Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure. Gene. 238:23–31.

Chou PY, Fasman GD. 1974. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry. 13:211–222.

Clark AG, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 450:203–218.

Creighton TE, Darby NJ. 1989. Functional evolutionary divergence of proteolytic enzymes and their inhibitors. Trends Biochem Sci. 14:319–324.

Dean AM, Neuhauser C, Grenier E, Golding GB. 2002. The pattern of amino acid replacements in α/β-barrels. Mol Biol Evol. 19:1846–1864.

Dudgeon K, Famm K, Christ D. 2008. Sequence determinants of protein aggregation in human VH domains. Protein Eng Des Sel. 22:217–220.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Ferrada E, Wagner A. 2008. Protein robustness promotes evolutionary innovations on large evolutionary time scales. Proc R Soc Lond B Biol Sci. 275:1595–1602.

Hanada K, Gojobori T, Li W-H. 2006. Radical amino acid change versus positive selection in the evolution of viral envelope proteins. Gene. 385:83–88.

Holloway AK, Begun DJ, Siepel A, Pollard KS. 2008. Accelerated sequence divergence of conserved genomic elements in Drosophila melanogaster. Genome Res. 18:1592–1601.

Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature. 335:167–170.

Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 292:195–202.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22:2577–2637.

Kaur H, Raghava GPS. 2002. An evaluation of β-turn prediction methods. Bioinformatics. 18:1508–1514.

Kimura M. 1968. Evolutionary rate at the molecular level. Nature. 217:624–626.

Knight R, et al. 2007. PyCogent: a toolkit for making sense of sequence. Genome Biol. 8:R171.

Koehl P, Levitt M. 1999. Structure-based conformational preferences of amino acids. Proc Natl Acad Sci U S A. 96:12524–12529.

Komar AA. 2008. A pause for thought along the co-translational folding pathway. Trends Biochem Sci. 34:16–24.

Kosiol C, et al. 2008. Patterns of positive selection in six mammalian genomes. PLoS Genet. 4:e1000144.

Kyte J, Doolittle R. 1982. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 157:105–132.

Larracuente AM, et al. 2008. Evolution of protein-coding genes in Drosophila. Trends Genet. 24:114–123.

Lin YS, Hsu WL, Hwang JK, Li WH. 2007. Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. Mol Biol Evol. 24:1005–1011.

Lindsay H, Yap VB, Ying H, Huttley GA. 2008. Pitfalls of the most commonly used models of context dependent substitution. Biol Direct. 3:52.

Liu J, Zhang Y, Lei X, Zhang Z. 2008. Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. Genome Biol. 9:R69.

Marcelino AMC, Gierasch LM. 2008. Roles of β-turns in protein folding: from peptide models to protein engineering. Biopolymers. 89:380–391.

Mondragon-Palomino M, Meyers BC, Michelmore RW, Gaut BS. 2002. Patterns of positive selection in the complete NBS-LRR gene family of Arabidopsis thaliana. Genome Res. 12:1305–1315.

Montgomerie S, Sundararaj S, Gallin WJ, Wishart DS. 2006. Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinformatics. 7:301.

Moult J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT. 1997. Critical assessment of methods of protein structure prediction (CASP) round II. Proteins. 1:2–6.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast gene. Mol Biol Evol. 11:715–724.

O'Farrell HC, Xu Z, Culver GM, Rife JP. 2008. Sequence and structural evolution of the KsgA/Dim1 methyltransferase family. BMC Res Notes. 1:108.

O'Neil KT, DeGrado WF. 1990. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. Science. 250:646–651.

Pascarella S, Argos P. 1992. Analysis of insertions/deletions in protein structures. J Mol Biol. 224:461–471.

Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R. 2007. Genes under positive selection in Escherichia coli. Genome Res. 17:1336–1343.

Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genet. 2:e173.

Sanner MF, Spehner J-C, Olson AJ. 1996. Reduced surface: an efficient way to compute molecular surfaces. Biopolymers. 38:305–320.

Shepherd AJ, Gorse D, Thornton JM. 1999. Prediction of the location and type of beta-turns in proteins using neural networks. Protein Sci. 8:1045–1055.

Subramanian AR, Kaufmann M, Morgenstern B. 2008. DIALIGN-TX: improvement of the segment-based approach for multiple sequence alignment by combining greedy and progressive alignment strategies. Algorithms Mol Biol. 3:6.

Thomas JW, et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. Nature. 424:788–793.

Thompson JD, Higgins GD, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Wong KM, Suchard MA, Huelsenbeck JP. 2008. Alignment uncertainty and genomic analysis. Science. 319:473–476.

Wong WSW, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics. 168:1041–1051.

Wright F. 1990. The 'effective number of codons' used in a gene. Gene. 87:23–29.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol. 25:568–579.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155:431–449.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22:1107–1118.

Yap VB, Lindsay H, Easteal S, Huttley G. 2009. Estimates of the effect of natural selection on protein coding content. Mol Biol Evol. doi: 10.1093/molbev/msp232

Zhang H, et al. 2008. Sequence based residue depth prediction using evolutionary information and predicted secondary structure. BMC Bioinformatics. 9:388.

Associate editor: Laurence Hurst