Protein Insertions and Deletions Enabled by Neutral Roaming in Sequence Space

Ágnes Tóth-Petróczy and Dan S. Tawfik*
Department of Biological Chemistry, Weizmann Institute of Science, Rehovot, Israel
*Corresponding author: E-mail: dan.tawfik@weizmann.ac.il.
Associate editor: Jianzhi Zhang

Abstract

Backbone modifications via insertions and deletions (InDels) may exert dramatic effects, for better (mediating new functions) and for worse (causing loss of structure and/or function). However, contrary to point mutations (substitutions), our knowledge of the evolution and structural–functional effects of InDels is limited and so is our capability to engineer them. We sought to assess how deleterious InDels are relative to point mutations and understand the mechanisms that mediate their acceptance. Analysis of the evolution of InDels in orthologous protein phylogenies indicated that their rate of purging is 9- to 100-fold higher than for point mutations. In yeast, for example, the substitutions-to-InDels ratio is approximately 14-fold higher in protein coding than in noncoding regions. The incorporation of InDels relative to substitutions is not only slow but also nonlinear. On average, ≥50 substitutions accumulate before the appearance of the first InDel. We also found enriched substitutions in sequential and spatial proximity to InDels, suggesting that certain substitutions are correlated with InDels. As indicated by the lag in InDels accumulation, some of these correlated substitutions may have occurred first, as apparently neutral mutations, and later enabled the accumulation of InDels that would be otherwise purged. Thus, compensatory substitutions may follow InDels in an “adaptive walk” as traditionally assumed, but might also accumulate first, by “neutral roaming.” The dynamics of InDels accumulation also depends on their genomic frequencies—InDels in flies are 4-fold more frequent than in yeast and tend to be compensated rather than enabled.

Key words: indels, correlated mutation, covariance, compensatory mutation, enabling mutation.

Introduction

Short insertions and deletions (InDels, in the order of several nucleotides) are frequent events, occurring in some genomes at a frequency that is comparable to that of point mutations (Denver et al. 2004; Lynch et al. 2008). However, although the effects of substitutions (nucleotide and amino acid point mutations) were extensively studied, our understanding of InDels is limited. Similarly, although protein engineering by substitutions is a matter of routine, the engineering of backbone modifications comprises a challenge (Hu et al. 2007; Murphy et al. 2009; Afriat-Jurnou et al. 2012). InDels can severely perturb the structural integrity of proteins and may therefore be highly deleterious. At the same time, InDels often comprise the key step in the acquisition of new functions (Tawfik 2006; Akiva et al. 2008; Neuenfeldt et al. 2008; Britten 2010; Cooley et al. 2010; Hashimoto and Panchenko 2010). For example, a deletion in an active-site loop, combined with an adjacent, correlated substitution, led to a >10^6 change in enzyme selectivity (Afriat-Jurnou et al. 2012). Such dramatic shifts are not observed with substitutions. Purifying selection (neutral evolution), and adaptive evolution, may therefore act more strongly on InDels than on substitutions, but the relative magnitudes remain unknown. Similarly, compensatory mechanisms are likely to be critical to InDels acceptance, but the nature of these mechanisms is scarcely addressed (Leushkin et al. 2012).

Here, we have examined the relationship between sequence divergence by amino acid substitutions and InDels. We investigated how InDels evolve in natural proteins and obtained new insights regarding the mechanisms of acceptance of InDels. Specifically, we explored the following questions: What is the magnitude of purifying selection acting on InDels? Are InDels less abundant than substitutions because they occur less frequently than point mutations or because they are more intensely purged (and are therefore also harder to engineer)? Is there a correlation between InDels and certain substitutions? By which mechanism do InDels become tolerated: Are InDels compensated by certain substitutions after their incorporation, that is, via adaptive walks (Leushkin et al. 2012) as generally assumed for deleterious mutations? Or do, perhaps, substitutions accumulate first as neutral and thereby allow InDels to accumulate at a later stage (a mechanism dubbed here “neutral roaming”)?

To explore the above issues, we studied orthologous proteins along two different phylogenies—yeast and fly, and examined the emergence of substitutions and InDels within their proteins. The relative frequencies of occurrence of InDels in these organisms are substantially different. In flies, where compensatory substitutions have been identified in correlation with InDels (Leushkin et al. 2012), point mutations occur only four times more frequently than InDels. In yeast, however, InDels are more rare (approximately 1/16 relative to point mutations).
Results

The Substitutions to InDels Ratio

We assembled a database of common orthologs of 22 fungi species that experienced approximately 500 My of evolution and aligned their sequences. After filtering of low-quality sequences and annotation mistakes, our data set comprised 1,310 orthologous families (supplementary data set S1, Supplementary Material online). Gapped positions were identified by pairwise alignments. InDels were defined as consecutive gapped regions within domains only (i.e., the N- and C-termini length variations were excluded). The most frequent InDels (33%) were of a single amino acid residue, and the median length was of two residues (supplementary fig. S1, Supplementary Material online). We then compared the level of sequence divergence with respect to InDels versus substitutions. Comparison of both proteomes and of individual families indicated that the substitutions/InDels ratio (S/I hereafter) is not constant—at low divergence, InDels accumulated at a slower rate than at high divergence (fig. 1a and b and supplementary fig. S2, Supplementary Material online).

Handling InDels comprises a challenge for alignment algorithms (Golubchik et al. 2007). To assess the bias introduced by the alignment algorithm, we tested four algorithms that use different gap-handling strategies (see Materials and Methods). The nonlinear, lag pattern was observed throughout although its magnitude varied (fig. 1c). We subsequently chose MUSCLE that is the most conservative with gap assignments and exhibited the weakest signal of nonlinearity (fig. 1c). The results presented here therefore represent the most conservative estimate of correlation between substitutions and InDels.

Correlated Emergence of Substitutions and InDels

The lag phase in the accumulation of InDels (fig. 1a and b) suggests that certain substitutions accumulate before the acceptance of InDels. A similar lag behavior was found between the accumulation of surface and core mutations, suggesting a surface–core correlation, and a mechanism of apparently neutral surface substitutions enabling the accommodation of core substitutions that would be otherwise deleterious (Toth-Petroczy and Tawfik 2011).

To quantify the nonlinear relationships, we used a second-order polynomial function: \( y(x) = ax^2 + bx \) that correlates the number of substitutions (x) and InDels (y) in evolving proteins (fig. 1a; \( R^2_{\text{secondorder}} = 0.96 \)). We defined the “lag intensity” as the absolute value of \( \frac{b}{a} \); the lower the value the more intense the lag is. An alternative measure is “lag length,” defined as the intercept with the x axis according to a linear fit. By the null hypothesis of independent accumulation, the number of substitutions at zero InDels would be zero. However, the lag lengths observed are higher than zero (fig. 1a; lag length = 8% sequence divergence, or on average, ~35 amino acid substitutions per domain).

In addition to the comparison of proteomes (fig. 1a), we compared individual protein families. We grouped fungi orthologous pairs by sequence divergence and calculated the average number of InDels for each bin (fig. 1b). This representation confirms the nonlinear relationship between InDels and substitutions accumulation. Despite the huge variability between proteins, the mean values in each bin are best fitted by the second order function \( R^2_{\text{secondorder}} = 0.99 \), and the linear fit \( R^2_{\text{linear}} = 0.89 \) gives a lag length of 8% sequence divergence by substitutions. Further, grouping proteins by their evolutionary rates to slow, average, and fast categories (Toth-Petroczy and Tawfik 2011) showed that, as expected, slowly evolving proteins show the highest substitution/InDels ratio, as well as the strongest lag (supplementary fig. S3a, Supplementary Material online).

Finally, we examined the S/I ratios in clusters of orthologous groups (COGs) that include sequences from all kingdoms of life. The lag trend is clearly observed (fig. 1d). Despite considerable variability, >90% of COGs show a lag, with the average lag length being 56 amino acid substitutions (fig. 1d inset).

Substitutions and InDels in Noncoding Regions Are Not Correlated

Is the observed correlation because of the genetic mechanisms that incorporate substitutions together with InDels (Yang et al. 2008; De and Babu 2010; Hicks et al. 2010), or it is the outcome of purifying selection acting on InDels within protein domains? To address this question, we analyzed a data set of the flanking noncoding 5′-regions of the 1,310 orthologous yeast proteins (Kellis et al. 2003). Although these regions are partly under selection, they were chosen because, unlike most other noncoding genome parts, they could be reliably aligned. Unlike coding regions, 5′-UTR regions exhibit a linear relationship between nucleotide substitutions and InDels (fig. 2). A comparison of a subset of four closely related yeast species, where limited divergence enables even more reliable alignments, gave the same result (supplementary fig. S3b, Supplementary Material online). This suggests that even under selection, the appearance of InDels in noncoding regions is independent of the accumulation of nucleotide substitutions.

The measure of correlations between InDels and substitutions can also be affected by saturation or multiple hits at the same site. Saturation might result in underestimation of the divergence rate for substitutions or for InDel. In the case of InDel, there is no available model for reversible or multiple InDel events. However, we could estimate from the occurrence of reverting InDels in independent lineages (i.e., InDels that occurred at the same site) that multiple hits are rare (5% in fungi phylogenies, and 1% in the fly phylogenies described later). In addition, the fraction of multiple hits is proportional to the overall fraction of InDels. Thus, correction for multiple InDel hits would only increase the lag intensity. Correcting for saturation of substitutions would decrease the lag intensity. However, the comparison of coding regions to 5′-UTRs (fig. 2) indicates that the nonlinearity observed in coding regions is unlikely to be the outcome of saturation of sites resulting in underestimated divergence rates (Gojobori...
1983), as saturation occurs much faster for base than for amino acid exchanges.

Thus, irrespectively to the protein sets compared, and even when the most conserved alignment algorithm is applied, the accumulation of InDels shows a nonlinear relationship with sequence divergence as measured by substitutions. This nonlinearity seems to be the outcome of purifying selection.

**InDels in Ordered versus Disordered Regions**

InDels appear in loops more frequently than in secondary structure elements (de la Chaux et al. 2007). However, the same trend holds for substitutions (Tokuriki et al. 2008). We, therefore, examined the relative tolerance of InDels vs. substitutions in secondary structure elements and in loops (regions with no secondary structure). We analyzed all yeast orthologous families where the *Saccharomyces cerevisiae* ortholog has a known structure (*N* = 170). The number of substitutions in loops and in secondary structures largely follows linear relationship with a slope of 1.25. On average, approximately 2/3 of these proteins’ residues comprises secondary structure elements. The 1.25 slope, therefore, reflects a 2-fold higher tolerance of substitutions in loops then in structured regions (supplementary fig. 4, Supplementary Material online). For InDels, however, the relationship is nonlinear, and loops are approximately 14 times more tolerable to InDels than secondary structured regions (fig. 3). In accordance, InDels are more tolerated in disordered regions than in ordered ones; the S/I is on average 3-fold lower for disordered regions, and the relative accumulation rate of InDels and substitutions is nearly linear (fig. 3 and supplementary fig. S5, Supplementary Material online).

Because of higher occurrence of polymerase slippage, InDels occur at higher frequency in repeat regions (McDonald et al. 2011), and disordered regions are enriched in repeat sequences (Simon and Hancock 2009).
occurrence of InDels in disordered regions due to the presence of repetitive sequences? Our data set contains relatively few tandem repeats (0.04% of all InDels, defined as ≥4 residues duplications). Thus, enrichment of InDels in repeat regions is not a significant phenomenon within conserved orthologs.

Enabling Substitutions in InDel Neighboring and Contacting Residues

As demonstrated earlier, in ordered protein domains, on average, 50 substitutions accumulate before the appearance of the first InDel. However, only few of these substitutions might actually be correlated with the appearance of a particular InDel. In general, strong enrichment in substitutions in contacting residues appear to underline drifting orthologs (Callahan et al. 2011; Toth-Petroczy and Tawfik 2011). We therefore examined the rate of substitutions in positions that are in sequential and spatial neighborhood to InDels. We compared them to the rate of substitutions in the remaining parts of the protein, thus deriving a relative measure of local/global rates.

Approximately 70% of the InDel flanking positions (sequential neighbors) and contacting positions (spatial neighbors) evolve faster than the rest of the protein (fig. 4a). We also calculated the sequence identity of regions that flank InDels relative to the entire protein and found increased divergence (fig. 4b). The enriched divergence drastically declines with increasing window size—the strongest signal is at window size of five residues of sequential neighbors, and the signal largely disappears at a window of 20 residues. Approximately 20% of flanking positions show the opposite behavior: The InDel regions evolve slower than the rest of the protein. However, in these cases, the regions where the InDels occurred show much higher divergence than the rest of the protein (supplementary fig. S6b, c, Supplementary Material online). For noncoding regions, no increased divergence is seen in regions that are flanking InDels (mean local identity – global identity ~0; supplementary fig. S6b, c, Supplementary Material online). The signals of enriched divergence, therefore, suggest that correlated mutations occur primarily in proximity, or even in direct contact, with the "future" InDel. Thus, as indicated by other data (fig. 2), the increased local divergence around InDels is due to selection rather than due to the genetic mechanisms that incorporate InDels (Hicks et al. 2010; Borneman et al. 2011).

The increased divergence rates of indel flanking regions (fig. 4) may also relate to the fact that InDels accumulate in regions, which are generally more permissive to mutations, and specifically to loop regions (fig. 3). To support the hypothesis of InDel-correlated substitutions, we compared close orthologs with and without InDels. The alignments follow approximately 500 My of drift yet only include 22 sequences from the currently available genomes. Consequently, the predicted ancestral nodes and their descendent proteins are separated by numerous sequence exchanges. Nonetheless, we assessed the number of InDel-specific substitutions via comparison of reference and outgroup sequences in a manner similar to that of Leushkin et al. (2012). The alignments were polarized using a reference sequence where the InDel did not occur and an outgroup (fig. 5a). As found for fly orthologs (Leushkin et al. 2012), InDel-carrying sequences show an increase in the frequency of substitutions relative to their reference sequences. In both the fly and fungi sets, the increase is most pronounced at window size of 5 residues and
disappears by 20 residues (fig. 5b and c, N = 534 common fungi and fly orthologs). The shift is, however, far more pronounced for the fly phylogenies (fig. 5c), probably due to the InDel rates in fly genomes being approximately 4-fold higher than in fungi (see later). Furthermore, the positions, where

InDel-specific substitutions occur show, on average, slower evolutionary rates than the reference-specific ones (supplementary fig. S8, Supplementary Material online). Namely, the InDel-specific substitutions appear at more conserved sites, which supports their correlated appearance with the InDels.

**Fig. 4.** Increased substitution rates in sequential and spatial proximity to InDels. (a) The evolutionary rates of positions in the vicinity of InDels (local rate) were compared with the global rate (mean rate for the entire protein). The neighboring residues of InDels both in sequence (10 residue window, black, mean = 1.24) and in space (<4 Å distance, red, mean = 1.33) show increased evolutionary rates (local rate/global rate > 1; t-test P values < 10^-5 for both sets). (b) Similarly, InDel neighboring positions exhibit increased sequence divergence (local identity − global identity < 0). The difference in the sequence identities disappears with increasing window size from 5 to 20 residues (N_\text{InDels} = 150,824, means are −0.17, −0.11, and −0.06, respectively; t-test P values < 10^-5 for all data sets). Sequence identity was calculated based on pairwise comparison of sequences, as the number of substitutions divided by the aligned length of the sequences (i.e., excluding gaps). (c) Increased sequence divergence of InDel flanking positions in fly orthologs at window sizes of 5, 10, 20, 30, and 50 residues (N_\text{InDels} = 167,362, means are −0.31, −0.26, −0.20, and −0.17, respectively, t-test P values < 10^-5).

**Fig. 5.** The correlated appearance of InDels and substitutions. (a) InDel-specific substitutions were identified by virtue of appearing in one sequence (an InDel carrying ortholog, or InDel-specific) and in neither the reference nor the out-group orthologs. The fraction of InDel-specific and reference-specific substitutions were determined for all positions. (b, c) The distribution of the differences of the occurrences of InDel-specific versus reference-specific substitutions for the fungi (b) and fly (c) phylogenies. In both sets, InDel-specific changes are enriched in the InDels’ vicinity, and this enrichment diminishes with increasing window size (from 5 to 20 residues; t-test P values are 1.8 × 10^-5, 0.015, 0.12, respectively; P values are < 10^-5 for the flies for all window sizes).
InDels Incorporation in Fly versus Fungi Proteins

Organisms significantly differ in the ratio of appearance of point mutations versus InDels. Caenorhabditis elegans, for example, comprises an extreme case with a higher InDel rate than substitution rate (Denver et al. 2004). However, even for S. cerevisiae and Drosophila melanogaster the S/I ratios differ from 16 to 4, respectively (Lynch et al. 2008). Do these different genomic rates affect the S/I ratio in proteins? Does the higher rate affect the mechanism by which InDels accumulate?

To examine these questions, we constructed a data set of 12 different Drosophila species that share in total 10,119 orthologs (supplementary data set S1 and fig. S9, Supplementary Material online). We found that fly proteins accumulated up to 3-fold more InDels relative to substitutions compared with yeast (supplementary fig. S7a, Supplementary Material online). However, in the fly phylogeny that corresponds to approximately 50 My of evolution, approximately 80% of the proteome is assigned as shared orthologs, whereas in the fungi phylogeny (~500 My), only approximately 15% are assigned as orthologs. Consequently, the fungi orthologs represent highly conserved and structurally ordered proteins, whereas the fly orthologs include more disordered proteins (supplementary fig. S7b, Supplementary Material online). To eliminate this bias (fig. 3 and supplementary fig. S5, Supplementary Material online), we analyzed orthologs that are common to both phylogenies (N = 534 proteins). For this data set, the S/I ratio in the fly phylogeny was 2-fold lower than for yeast (fig. 6). Thus, although InDels occur approximately four times more frequently in fly than in yeast genomes, under selection, only 2-fold higher rate of InDels is observed. Thus, purifying selection eliminated most of the higher proportion of InDels that occurred in fly genomes.

The lag in InDels appearance seems less pronounced in the fly relative to the fungi phylogeny (fig. 6), but the signal for InDels and co-occurring substitutions in their flanking regions is stronger (fig. 5c vs. 5b). Although other factors such as population sizes and difference in the rate of sequencing errors in flies and fungi genomes may apply, it appears that the higher rate of InDels in fly proteins resulted in a larger fraction of InDels being compensated after their appearance, as previously argued (Leushkin et al. 2012), than being enabled by neutral mutations, as seen in the yeast phylogenies.

Discussion

What are the origins of the difference in the rate of accumulation of substitutions versus InDels? Two main reasons might apply: 1) differences in the intrinsic mutation rates of these two types of genomic modifications (i.e., regardless of selection) and 2) InDels are far more deleterious than point mutations and purifying selection acts more strongly on them. Our analysis of noncoding regions indicates an S/I (substitutions, including both synonymous and nonsynonymous exchanges, to InDels ratio) of approximately 8 (table 1). The mutation rates measured for S. cerevisiae indicate an S/I of approximately 16 (0.33 × 10⁻⁹ base substitution events per germ-line cell division versus 0.02 × 10⁻⁹ for short InDels) (Lynch et al. 2008). The difference between S/I of 16 and 8 may originate from differences in laboratory versus naturally occurring strains. Underestimation of the substitution rate due to reversible exchanges (saturation) is also a possibility, although we measured a similar ratio for all 22 species and for 4 closely diverged ones (supplementary fig. S3b, Supplementary Material online). In coding regions, approximately 2/3 of the InDels are not accepted due to frame shifts, whereas the frequency of nonsense substitutions is approximately 0.05. So we would expect substitutions to be approximately 8 × 3, that is, approximately 24 times more frequent than InDels in coding regions. However, the observed S/I rate in coding regions (including synonymous substitutions) is, on average, approximately 220 (table 1). This implies that regardless of frame shifts, purifying selection acts much more strongly on InDels. Further, the degree of purifying selection acting on InDels and the concomitant level of correlation between InDels and substitutions vary depending on the degree of structural packing and cooperativity. Weak selection and no correlation are observed in DNA promoter regions that comprise linear motifs. In proteins, weak correlation is observed in disordered protein regions, whereas the highest level of purifying selection and correlation is seen in secondary structural elements that comprise the highest degree of structural order. Within ordered regions, selection acts at least 9-fold more intensely on InDels and up to 100-fold more intensely in secondary structure elements (fig. 3). Consequently, InDels are preferentially located in regions with no assigned structure (de la Chaux et al. 2007).

The substitutions-to-InDels ratio is not only high due to the intensity of selection but is also not constant—it decreases on the level of sequence divergence (fig. 1 and supplementary fig. S2, Supplementary Material online). An opposite tendency was shown for nucleotide exchanges and InDels while comparing full genome alignments of closely related mammalian species (Chen et al. 2009). The
Table 1. Substitutions-to-InDel Ratios in Coding and Noncoding Regions of Different Genomes.

<table>
<thead>
<tr>
<th>Genomes</th>
<th>Subset</th>
<th>Region</th>
<th>S/I Ratio</th>
<th>Lag Length&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Lag Intensity&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Values Derived from Figure:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;sup&gt;I = 1&lt;/sup&gt; I = 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yeast</td>
<td>Four &lt;em&gt;sensu stricto&lt;/em&gt; species (N = 4,192)</td>
<td>Complete proteins</td>
<td>21</td>
<td>7</td>
<td>23</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Coding regions&lt;sup&gt;d&lt;/sup&gt;</td>
<td>219&lt;sup&gt;10&lt;/sup&gt;</td>
<td>75</td>
<td>70</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Noncoding regions&lt;sup&gt;d&lt;/sup&gt;</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;em&gt;Saccharomyces cerevisiae&lt;/em&gt;</td>
<td>Spontaneous mutation rate&lt;sup&gt;e&lt;/sup&gt;</td>
<td>16.5</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Fungi</td>
<td>Orthologs of 22 fungi species (N = 1,310)</td>
<td>Complete proteins</td>
<td>50</td>
<td>25</td>
<td>56</td>
<td>158</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Secondary structure regions&lt;sup&gt;f&lt;/sup&gt;</td>
<td>110</td>
<td>35</td>
<td>20</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Loop regions&lt;sup&gt;f&lt;/sup&gt;</td>
<td>54</td>
<td>28</td>
<td>13</td>
<td>182</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ordered regions&lt;sup&gt;g&lt;/sup&gt;</td>
<td>56</td>
<td>25</td>
<td>55</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disordered regions&lt;sup&gt;g&lt;/sup&gt;</td>
<td>18</td>
<td>11</td>
<td>3.4</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>Common fungi–fly orthologs (N = 534)</td>
<td></td>
<td>47</td>
<td>24</td>
<td>49</td>
<td>167</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sliding window&lt;sup&gt;h&lt;/sup&gt;</td>
<td>102 ± 50</td>
<td>28 ± 9</td>
<td>55 ± 50</td>
<td>42 ± 117</td>
</tr>
<tr>
<td></td>
<td>Drosophila melanogaster</td>
<td></td>
<td>4</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Orthologous protein families (N = 600)</td>
<td></td>
<td>102 ± 50</td>
<td>28 ± 9</td>
<td>55 ± 50</td>
<td>42 ± 117</td>
</tr>
</tbody>
</table>

<sup>a</sup>Because of the nonlinear dependency, the average S/I (the S/I ratio) was derived for various levels of divergence, namely, for the first and tenth InDel (I = 1 and I = 10, respectively).

<sup>b</sup>"Lag length" relates to the intercept of the x axis of the linear fit (—b/a) in InDel-to-substitution graphs (e.g., fig. 1).

<sup>c</sup>"Lag intensity" is the b/a value derived from the nonlinear fit ax<sup>2</sup> + bx in InDel-to-substitution graphs (e.g., fig. 1).

<sup>d</sup>These S/I values are measured for nucleotide substitutions (rather than amino acids, as in the "Complete proteins" data set) and InDels based on DNA alignments.

<sup>e</sup>The spontaneous mutation rates as observed in laboratory strains (Lynch et al. 2008).

<sup>f</sup>InDels and substitutions were determined in regions that adopt helical or sheet structures ("secondary structure regions"), or in loop and/or disordered regions, based on the crystal structures of the <em>S. cerevisiae</em> orthologs (N = 170).

<sup>g</sup>Disordered regions were predicted using IUPred (see Materials and Methods).

<sup>h</sup>The average and the sd. relates to the 600 fits for the individual orthologous protein families.
discrepancy might reflect the different tendencies in noncoding regions (McDonald et al. 2011) (fig. 2), and especially in repeat regions, and may also relate to ambiguities associated with aligning full genomes. In general, alignment algorithms handle gaps very poorly even in coding regions (Golubchik et al. 2007). Indeed, the choice of the alignment algorithm strongly influences the S/I ratio. We tested four different algorithms that use different strategies for gap handling. The highest frequency of InDels is assigned by PRANK, which is using phylogenetic information for the identification of InDels, and avoids multiple penalization of InDel events (Loytynoja and Goldman 2008). Nonetheless, the correlated nature of InDels and substitutions was corroborated by all algorithms tested (fig. 1c).

Accumulating evidence indicates that point mutations and InDels are interdependent (De and Babu 2010; Hicks et al. 2010; Leushkin et al. 2012). At the genomic level, base substitutions and InDels co-occur and positively covary both within and between species (Hardison et al. 2003; Longman-Jacobsen et al. 2003). Because of the low fidelity of nonreplicative DNA repair polymerases that introduce break-repair-induced single-nucleotide mutations in the vicinity of genomic structural alterations (Rattray et al. 2001; Rattray and Strathern 2003; Hicks et al. 2010), certain regions are prone to be simultaneously altered by both mutation types (Tian et al. 2008; De and Babu 2010). Is the covariance of InDels and substitutions observed here, and by others (Leushkin et al. 2012), driven by the genetic mechanism that incorporates InDels, or is it primarily the outcome of selection? We observed increased substitution rates in the sequential vicinity of InDels in protein coding regions and no increase in the proximity of InDels within noncoding regions (up to 60 nucleotides upstream and downstream). The effect of DNA repair may therefore be limited to long InDels and not to the short InDels analyzed here. Taken together with the increased substitution rates in the spatial vicinity of InDels (that cannot relate to the DNA repair machinery), our data are consistent with the InDel–substitution correlation being the outcome of selection.

The nonlinear accumulation of InDels (fig. 1), and the covariance of substitutions and InDels (figs. 4 and 5; see also Leushkin et al. 2012) indicates that the occurrence of InDels is highly context dependent or epistatic—InDels seem to occur in combination with certain substitutions. The primary assumption is that these correlated substitutions compensate for the deleterious effects of InDels and are therefore fixed by positive selection, that is, via adaptive walks (Leushkin et al. 2012). Given the level of purifying selection acting on InDels, particularly in ordered regions, let alone in secondary structure elements, many InDels, unless enabled by certain point mutations, may result in complete nonfunctionalization. The lag behavior also suggests that correlated substitutions may occur before, presumably as neutral, and later enable the accumulation of InDels. A similar phenomenon seems to underline the accumulation of highly deleterious core substitutions that may be enabled by initially neutral substitutions on the surface (Toth-Petroczy and Tawfik 2011).

Overall, “compensatory mutations” comprise the ruling paradigm, whereas the notion of “enabling mutations,” neutral mutations coming first and thereby enabling deleterious changes, is relatively unrecognized (ISI search [topics] for “compensatory mutation” versus “enabling mutation” is 751:2 hits and Google fight 3200:114 hits). Both in the fly and fungi phylogenies, a high fraction of InDel-specific substitutions might have occurred as compensatory. However, the lag behavior suggests a measurable fraction of enabling mutations that appeared before the InDel events. Leushkin et al. (2012) proposes that InDels trigger adaptive changes, namely adaptive amino acid substitutions occur after an InDel event, to compensate for its fitness effect. We surmise that both pre- and postcompensation contribute to the acceptance of InDels (fig. 7). Thus, InDels, and other types of highly deleterious mutations, such as point mutations in core residues (Toth-Petroczy and Tawfik 2011), can get incorporated via neutral roaming. Indeed, several studies identified substitutions that accumulated as neutral (permissive mutations) and had subsequently enabled highly deleterious point mutations to accumulate (Ortlund et al. 2007; Lunzer et al. 2010). InDels are, in general, far less studied than point mutations. Nonetheless, they are subjected to the same epistatic constraints. Namely, InDels that accumulated within one sequence context are incompatible with others. In yeast, for example, the coinducer Gal3 diverged by duplication from the enzyme Gal1. Divergence involved a deletion of two amino acids within Gal1’s active site, rendering Gal3 enzymatically inactive yet functional as coinducer. The same deletion, however, causes complete loss of function in the contemporary Gal1, both as an enzyme and as coinducer. Similarly, the respective insertion causes loss of function in Gal3 (Hittinger and Carroll 2007). Thus, in the course of divergence, certain substitutions appeared that enabled this deletion.

FIG. 7. Schematic representation of the two mechanisms of divergence: Adaptive walking and neutral roaming. The x and y axes represent genotypes, and the z axis represents phenotype, that is, fitness. The plot describes two lineages, marked with a start and end point, along which a correlated substitution and an InDel occur. The red lineage describes an “adaptive walk”—a deleterious InDel occurs first, followed by a compensatory substitution. The blue lineage describes “neutral roaming”—a nearly neutral substitution occurs first, and thereby enables the acceptance of the deleterious InDel that follows. Evidence for correlated substitutions and InDels is presented here (fig. 5) and elsewhere (Leushkin et al. 2012), but correlation in itself does not indicate the mechanism. However, the lag phase in the accumulation of InDels relative to substitutions (figs. 1a and 2) lends support to the role of neutral roaming in accumulation of InDels.
So far, we compared only orthologs that largely bear the same function, and the sequence changes relate mostly to genetic drift. However, the role of InDels in the functional adaptation of proteins is well recognized (Grishin 2001; Tawfik 2006; Guburchik et al. 2007; Akiva et al. 2008; Jochens et al. 2009; Hashimoto and Panchenko 2010; Afriat-Jurnou et al. 2012; Saab-Rincon et al. 2012), and examples of beneficial InDels are known (Podlaha and Zhang 2003; Podlaha et al. 2005). We found, however, that paralogs exhibit only a slightly higher InDel rate, and a slightly lower S/I ratios, relative to the corresponding orthologs (supplementary fig. S2, Supplementary Material online). The weak signal may be due to the fact that adaptation is also manifested in the increased substitution rate, and along a given phylogeny, most InDels relate to drift and only few actually drive neo-functionalization.

Finally, the newly identified features described here also suggest ways of facilitating the engineering of InDels, particularly of short InDels within active-site loops. Our own experience has been, for example, that a deletion within an active-site loop became advantageous only when a point mutation in a residue upstream to this deletion was incorporated (Afriat-Jurnou et al. 2012). In other loop grafting experiments, the introduction of variability at hinge residues that connect the loops with the scaffold facilitated the acceptance of loop InDels (Ochoa-Leyva et al. 2009, 2011). We surmise that the loops with the scaffold facilitated the acceptance of loop InDels (Ochoa-Leyva et al. 2009, 2011). We surmise that enabling point mutations, both in sequential and spatial vicinity, comprise the key to InDels incorporation. Methods for predicting such enabling mutations may promote our understanding of natural evolution and facilitate the laboratory engineering of proteins.

**Material and Methods**

**Data Sets**

Fungi sequences and orthology annotations were obtained from the Fungal Orthogroups Repository (Wapinski et al. 2007) and Flybase (Tweedie et al. 2009) (species are listed in supplementary fig. S9, Supplementary Material online). Because of many gene annotation errors, the sequences were filtered. Orthologs that diverged more than 70%, or deviated in length by more than 50%, were excluded. Overall, we used 1,310 fungi and 10,119 fly proteins with orthologs in all species within their phylogeny. The final list of proteins can be found in supplementary data set S1, Supplementary Material online. Orthologous proteins groups (COGs) were downloaded from eggNOGv2 database (Muller et al. 2009). Only COGs that include >100 sequences and shared >40% sequence identity were included to minimize alignment errors (N = 600). Saccharomyces cerevisiae proteins with known structures were retrieved from PDB.

**Sequence Analysis**

We tested four different alignment methods: MUSCLE (Edgar 2004), Dialign (Subramanian et al. 2008), ClustalW (Larkin et al. 2007), and PRANK (Loytynoja and Goldman 2008) (fig. 1c). Gapped positions were counted, and InDels were defined as consecutive gapped regions, and only InDels within domains were included. InDel analysis and statistical calculations were performed using Perl scripts. Secondary structure assignments were done by dssp (Kabsch and Sander 1983) based on high-resolution X-ray structures (<2.5 Å). Missing regions within structures were assigned as “loops,” because the lack of density usually relates to high flexibility. InDels within secondary structure elements were defined as such if positions flanking both sides of the InDel were annotated as helix or sheet (i.e., InDels at the ends of helixes and sheets were assigned as loop). InDel-contacting residues were defined based on spatial proximity (<4 Å) to the InDel flanking residues (one residue upstream and downstream) in the S. cerevisiae structures. We categorized domains as structured and unstructured based on disorder prediction by IUPred (Dosztanyi et al. 2005). Disordered domains were defined as ≥30 residues long regions, with less than three residues consecutively predicted to be ordered, and average disorder score >0.5. Ordered domains were defined as regions with average disorder scores <0.5 and <3 consecutive disordered residues.

**Trees and InDel-Specific Lineage Analysis**

Maximum likelihood phylogenetic trees were created by PhyML (Guindon and Gascuel 2003) based on the concatenated alignments of all orthologs with known structures (supplementary fig. S9, Supplementary Material online). Evolutionary rates were calculated by Rate4Site (Pupko et al. 2002) using JTT matrix.

InDel-specific substitutions were determined by polarizing the alignments of three species from the phylogeny: ((InDel,Reference),Outgroup) (fig. 5a). For the fungi data set, the following triplets were considered: ((Scer,Spar),Smik), ((Spar,Smik)Sbay), ((Cgla,Scas),Sklu), ((Sklu,Kwal),Agos), ((Klac,Agos),Ylip), ((Ctro,Caib),Dhan), ((Cpar, Lelo), Dhan), ((Dhan,Cgii),Ylip), ((Anid,Ncra),Sjap), and ((Soct,Spom),Sjap). For the fly phylogeny, the following triplets were considered: ((Dsim, Dse),Dmel), ((Dyak,Dere),Dana), ((Dpsd,Dper),Dwil), and ((Dmoj, Dvir),Dgri). Within each triplet, substitutions were counted that occurred only in the InDel containing sequence (InDel-specific divergence) but not in the reference and in the outgroup sequence. Similarly, substitutions were counted that occurred only in the reference sequence (reference-specific substitutions). The sequence divergence was calculated in the vicinity of the InDel (at window size of 5, 10, and 20 residues, local divergence) and for the whole sequence (global divergence). Relative divergence is defined as the difference between the local and global divergence (figs. 4 and 5). The difference between the relative InDel-specific and relative reference-specific divergence is plotted in figure 5 (for the common orthologs, N = 534).

Figures and statistical analysis were made by Matlab.

**Supplementary Material**

Supplementary figures S1–S9 and data set S1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).
Acknowledgments

This work was supported by the Israel Science foundation. D.S.T. is the Nella and Leon Benoziyo Professor of Biochemistry.

References


