doi:10.1016/j.jmb.2004.05.045 *J. Mol. Biol.* (2004) **341**, 617–631

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

# Empirical Analysis of Protein Insertions and Deletions Determining Parameters for the Correct Placement of Gaps in Protein Sequence Alignments

**Mike S. S. Chang[1] and Steven A. Benner[2]***

[1]*Foundation for Applied Molecular Evolution, 1115 N.W. 4th Street, Gainesville FL 32601, USA*

[2]*Department of Chemistry University of Florida Gainesville, FL 32611-7200 USA*

**Corresponding author*

To understand how protein segments are inserted and deleted during divergent evolution, a set of pairwise alignments contained exactly one gap, and therefore arising from the first insertion–deletion (indel) event in the time separating the homologs, was examined. The alignments showed that "structure breaking" amino acids (PGDNS) were preferred within and flanking gapped regions, as are two residues with hydrophilic side-chains (QE) that frequently occur at the surface of protein folds. Conversely, hydrophobic residues (FMILYVW) occur infrequently within and flanking the gapped region. These preferences are modestly different in protein pairs separated by an episode of adaptive evolution, than in pairs diverging under strong functional constraints. Surprisingly, regions near an indel have not evolved more rapidly than the sequence pair overall, showing no evidence that an indel event must be compensated by local amino acid replacement. The gap-lengths are best approximated by a Zipfian distribution, with the probability of a gap of length $L$ decreasing as a function of $L^{-1.8}$. These features are largely independent of the length of the gap and the extent of divergence (measured by both silent and non-silent sequence changes) separating the two proteins. Surprisingly, amino acid repeats were discovered in more than a third of the polypeptide segments in and around the gap. These correspond to repeats in the DNA sequence. This suggests that a signature of the mechanism by which indels occur in the DNA sequence remains in the encoded protein sequences. These data suggest specific tools to score gap placement in an alignment. They also suggest tools that distinguish true indels from gaps created by mistaken gene finding, including under-predicted and over-predicted introns. By providing mechanisms to identify errors, the tools will enhance the value of genome sequence databases in support of integrated paleogenomics strategies used to extract functional information in a post-genomic environment.



*Keywords:* insertions; deletions; protein evolution; alignments; gaps

## Introduction

Algorithms have been available for 30 years that generate the optimal alignment of two protein sequences given a metric that scores matches and mismatches.[1,2] These are supported by scoring matrices extracted from empirical analyses of amino acid replacements during divergent evolution under functional constraints.[3,4] These scoring matrices are used by many alignment algorithms such as CLUSTAL W and MultAlign.[5–7] Scoring matrices have also been used with improved heuristics to generate multiple sequence alignments that hasten the process and improve accuracy, such as SAGA, COFFEE and T-Coffee.[8–10]

A significantly difficult task in creating alignments is the placement of gaps in pairwise alignments and in the multiple sequence alignments that are derived from them. Most alignment algorithms implement a cost for introducing a gap in

Abbreviations used: MSA, multiple sequence alignments; PAM, point accepted mutation; OGD, one gap database; indels, insertions and deletions.

E-mail address of the corresponding author: benner@chem.ufl.edu

an alignment, and add an incremental value to this cost each time the gap is extended by a single site.[11] This strategy fits conveniently within dynamic programming tools, and continues to be widely used, even though it has been known for a decade to be an imprecise description of gaps in real aligned protein sequences divergently evolving under functional constraints.[12]

To circumvent these inadequacies, programs such as CLUSTAL W adjust gap scoring to reflect simple rules in structural biology.[5] For example, gap penalties are reduced in glycine-rich regions, where many gaps have been empirically observed.[5,12] In addition, gaps are preferentially placed within segments of peptide sequence that have five or more hydrophilic amino acid residues. These are likely to be associated with loops or coil regions, which are believed to accept gaps easily in proteins diverging under functional constraints.[5,13]

Even with the most sophisticated tools, users of multiple sequence alignments (MSA) are inclined to adjust gaps that are provided by automatic tools. Such shuffling indicates a failure of the tool to produce an MSA that reflects biochemical intuition about where gaps in the MSA should appear.

The problem of placing gaps becomes more urgent in a post-genomic research environment. Tools that find genes in DNA sequences frequently over or under-predict introns, start sites, and stop sites. These create gaps in alignments with homologous proteins. In principle, a tool that distinguishes between gaps that arise from insertion and deletion events (indels) in the history of the protein family, and gaps that arise from faulty gene finding, would facilitate detection of mistakes within the context of homologs.

Ultimately, indels must be placed on individual branches of an evolutionary tree. This requires a stepwise process that begins with contemporary sequences, reconstructs ancestral states, makes decisions about whether to place a gap or an insert in the ancestral state, and then assigns insertion and deletion events explicitly to branches between nodes in a phylogenetic tree.[14] Such an approach has proven valuable to understand compensatory co-variation in protein sequences[15] and to correlate the genomic record with the geological and paleontological records of earlier life.[14,16]

In 1993, Benner *et al*. generated an empirical study of all gaps within the then-available database, with the goal of obtaining an empirical model for indels.[12] This pre-genomic analysis relied on a small database, meaning that many questions about the empirical behavior of gaps could not be addressed systematically. Further, the earlier work could not exploit recently developed tools for analysing divergent function and dating sequence divergences (T. Li *et al.*, unpublished results).[14,16]

The explosive growth of genetic databases allows us to re-examine indels in greater detail. Our goal is to provide a post-genomic empirical model of indel events just as the first indel is appearing in the time separating two sequences. This will provide a better understanding of what types of indels can be accepted by proteins divergently evolving under functional constraints, unconfused by multiple indel events in the same region of the sequence. This, in turn, should provide better tools for placing gaps in alignments and indels on trees, better tools for scoring indels, and better methods to distinguish gaps arising truly from indels and those arising from gene-finding errors.

## Results

### Analysis by evolutionary metrics

Table 1 illustrates separation of the one gap database (OGD) into bins based on PAM/$f_2$ evolutionary metrics (as described in Materials and Methods). "Typical" proteins are represented by three PAM/$f_2$ windows: $f_2 > 0.95$ with PAM $< 10$; $0.80 <= f_2 =< 0.95$ with $10 <=$ PAM $=< 100$; $f_2 < 0.80$ with PAM $> 100$. These windows contain 1006, 906 and nine protein pairs, respectively. These proteins pairs represent 39% of our OGD, and are populated by proteins that have recently diverged as well as those which have diverged long ago.

"Conserved" proteins divergently evolving under strong functional constraints are associated with low $f_2$ values and low PAM. A total of 2506 sequence pairs, from bins ($0.80 <= f_2 =< 0.95$ with PAM $< 10$) and ($f_2 < 0.80$ with $10 <=$ PAM $=< 100$) were found that fall into this conserved class, indicating that they have diverged under relatively strong purifying selection pressure. An additional 278 protein pairs appear to be in the "very conserved" class ($f_2 < 0.80$ with PAM $< 10$). Conserved proteins are the most abundant in our dataset, representing an overwhelming 56% of the OGD.

Adaptively diverging proteins appear to be the rarest, representing only 5% of the OGD. No pairs of proteins displayed extremely adaptive behavior, which would be categorized by $f_2 > 0.95$ with PAM $> 100$ or $0.80 < f_2 < 0.95$ with PAM $> 100$. However, $f_2$ and PAM values for 247 pairs of proteins do suggest that an episode of positive selection separates these pairs ($f_2 > 0.95$ while $10 <$ PAM $< 100$).

In the six $f_2$/PAM bins containing a significant number ($>100$) of pairs, the typical protein sequence pair aligned between 100 and 400 amino acid residues. The longest pairs aligned 3750 amino acid residues. This is as expected, given the well-known distribution of polypeptide lengths in folded domains, but also reflects the fact that pairwise matches with fewer than 50 conserved two-fold redundant sites were not considered.

**Table 1.** One gap pairs of homologous sequences at various $f_2$ and PAM ranges

| Bins | | PAM and $f_2$ ranges | | No. of pairs | No. of aligned positions | No. of deleted AA | Average gap length | Categories[a] |
|---|---|---|---|---|---|---|---|---|
| 1 | A | $f_2 > 0.95$ | PAM < 10 | 1006 | 304,213 | 6069 | 6.033 | typical |
| | B | $f_2 > 0.95$ | 10 < = PAM = < 100 | 247 | 57,626 | 958 | 3.879 | adaptive |
| | C | $f_2 > 0.95$ | PAM > 100 | 0 | 0 | 0 | 0 | very adaptive |
| 2 | A | $0.80 <= f_2 =< 0.95$ | PAM < 10 | 1079 | 409,403 | 5752 | 5.331 | conserved |
| | B | $0.80 <= f_2 =< 0.95$ | 10 < = PAM = < 100 | 906 | 280,627 | 3491 | 3.853 | typical |
| | C | $0.80 <= f_2 =< 0.95$ | PAM > 100 | 0 | 0 | 0 | 0 | adaptive |
| 3 | A | $f_2 < 0.80$ | PAM < 10 | 278 | 122,887 | 1605 | 5.773 | very conserved |
| | B | $f_2 < 0.80$ | 10 < = PAM = < 100 | 1427 | 481,701 | 4526 | 3.172 | conserved |
| | C | $f_2 < 0.80$ | PAM > 100 | 9 | 3482 | 26 | 2.889 | typical |
| | | Total | | 4952 | 1,659,939 | 22,427 | 4.529 | |

[a] Categories of bins are based on comparisons within this database. "Adaptive" proteins are observed to be evolving at a rate greater than what is observed as "Typical" for this dataset. Similarly, proteins are deemed "Conserved" based on their measured evolutionary metrics to be evolving at a slower rate as compared to typical.

## Gap length analysis

To understand how gaps arise in pairwise alignments, we first asked whether the distribution of gaps having different lengths was better described by an exponential function or a Zipfian function. With an exponential distribution, the number of gaps ($n$) of length ($L$) decreases according to the expression $n = c_1 \exp(-c_2 L)$, where $c_1$ and $c_2$ are parameters empirically selected to best fit the data. With a Zipfian distribution, the number of gaps ($n$) of length ($L$) decreases according to the expression $n = c_1(L^{-c_2})$.

Table 2 depicts the gap length ($L$) and the number of pairs of proteins ($n$) separated by bins reflecting their evolutionary distance within the OGD. The gap length distribution of the entire database fits remarkably well with a Zipfian expression over the entire gap length distribution (Figure 1A, sum of squares = 12,630). Comparatively, the best exponential fit is poorer (Figure 1B, sum of squares = 120,500). The sum of squares measurement provides an indication of how far the data depart from the fitted curve. It is clearly evident from this value that the exponential function fits the data much more loosely. The best exponent for the Zipfian fit is 1.821. This is consistent with the value obtained in preliminary work on a smaller data set, where the exponent in the Zipfian distribution is 1.7.[12]

This means that in real proteins, the number of very long gaps can be underestimated by an exponential distribution. As the widely used tools for scoring gaps assume an exponential distribution (it is easily incorporated into a Smith–Waterman algorithm), current gap-scoring heuristics penalizes long gaps more than is appropriate based on empirical evidence.

Significantly, the nature of the gap length distribution and the parameters of the best Zipfian equation are not substantially different when differing bins representing different protein pairs suffering different levels and types of selective pressure are examined separately. The Zipfian exponent ranged from 1.65 to 1.91 in the different bins, with no obvious trend (data not shown). From these results, we concluded that the gap length distribution was not different in proteins suffering an episode of adaptive evolution compared with those that are diverging under strong functional constraints.

## Amino acid analysis in and around the gap

We then asked whether different amino acids were found preferentially in regions within and around a gap. Here, the fractional representation of each of the 20 standard amino acids at positions 1, 2, 3, 4, 5, 6, and Z relative to the gap (defined in Table 3) were recorded. The observed values are shown in Table 3. To obtain propensities, the fractional representation of each of the 20 amino acids was also determined for the OGD as a whole.

The $\chi^2$ analyses were performed to evaluate the significance of the differences in amino acid usage at positions around the gap as compared to the amino acid usage of the complete OGD. Strikingly, a large number of the 20 amino acids exhibit significantly different propensities at positions surrounding the gap. The amino acids in Table 3 are organized by their physical properties ranging from hydrophobic to charged and hydrophilic residues. From this organization, it is immediately clear that the proportions of hydrophilic and hydrophobic residues are the most different as compared to that observed in the entire database.

The patterns were readily rationalized based on general concepts that relate protein sequence to protein fold, and the assumption that any indel event will need to survive natural selection to appear in our database. The amino acids having the highest propensity to appear within and around a gap are Ala, Asp, Gln, Glu, Gly, Pro, and Ser (ADQEGPS). The amino acids having the lowest propensity to appear within and around a gap are Phe, Met, Ile, Leu, Tyr, Val, Trp, and Cys

**Table 2.** Gap length of gene pairs at various evolutionary distances

| Gap length $L$ | $f_2 > 95$ PAM $< 10$ | $f_2 > 95$ $10 <= \text{PAM} =< 100$ | $f_2 > 95$ PAM $> 100$ | $80 <= f_2 =< 95$ PAM $< 10$ | $80 <= f_2 =< 95$ $10 <= \text{PAM} =< 100$ | $80 <= f_2 =< 95$ PAM $> 100$ | $f_2 < 80$ PAM $< 10$ | $f_2 < 80$ $10 <= \text{PAM} =< 100$ | $f_2 < 80$ PAM $> 100$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 461 | 131 | 0 | 576 | 475 | 0 | 157 | 833 | 5 |
| 2 | 125 | 39 | 0 | 134 | 143 | 0 | 33 | 193 | 2 |
| 3 | 73 | 22 | 0 | 90 | 72 | 0 | 22 | 141 | 1 |
| 4 | 59 | 13 | 0 | 45 | 43 | 0 | 7 | 65 | 0 |
| 5 | 30 | 7 | 0 | 23 | 26 | 0 | 8 | 31 | 0 |
| 6 | 41 | 3 | 0 | 23 | 24 | 0 | 2 | 31 | 0 |
| 7 | 24 | 1 | 0 | 20 | 12 | 0 | 2 | 21 | 0 |
| 8 | 22 | 3 | 0 | 7 | 22 | 0 | 3 | 16 | 0 |
| 9 | 11 | 1 | 0 | 11 | 8 | 0 | 3 | 6 | 0 |
| 10 | 8 | 3 | 0 | 16 | 7 | 0 | 4 | 9 | 0 |
| 11 | 11 | 5 | 0 | 3 | 9 | 0 | 1 | 2 | 0 |
| 12 | 8 | 1 | 0 | 14 | 10 | 0 | 3 | 8 | 0 |
| 13 | 6 | 1 | 0 | 9 | 5 | 0 | 3 | 1 | 0 |
| 14 | 9 | 0 | 0 | 6 | 3 | 0 | 1 | 9 | 1 |
| 15 | 11 | 3 | 0 | 7 | 2 | 0 | 3 | 4 | 0 |
| 16 | 9 | 1 | 0 | 4 | 6 | 0 | 1 | 4 | 0 |
| 17 | 6 | 0 | 0 | 2 | 2 | 0 | 1 | 7 | 0 |
| 18 | 7 | 1 | 0 | 5 | 2 | 0 | 1 | 1 | 0 |
| 19 | 1 | 2 | 0 | 2 | 3 | 0 | 0 | 3 | 0 |
| 20 | 7 | 2 | 0 | 3 | 0 | 0 | 2 | 4 | 0 |
| 21 | 6 | 1 | 0 | 7 | 3 | 0 | 1 | 4 | 0 |
| 22 | 3 | 0 | 0 | 6 | 1 | 0 | 0 | 2 | 0 |
| 23 | 3 | 1 | 0 | 0 | 4 | 0 | 0 | 1 | 0 |
| 24 | 5 | 0 | 0 | 2 | 1 | 0 | 1 | 3 | 0 |
| 25 | 1 | 1 | 0 | 4 | 1 | 0 | 1 | 4 | 0 |
| 26 | 3 | 0 | 0 | 4 | 3 | 0 | 0 | 3 | 0 |
| 27 | 2 | 0 | 0 | 6 | 1 | 0 | 2 | 2 | 0 |
| 28 | 2 | 0 | 0 | 7 | 1 | 0 | 1 | 0 | 0 |
| 29 | 3 | 0 | 0 | 2 | 2 | 0 | 2 | 1 | 0 |
| 30 | 1 | 0 | 0 | 4 | 0 | 0 | 1 | 2 | 0 |
| 31 | 3 | 0 | 0 | 1 | 1 | 0 | 2 | 1 | 0 |
| 32 | 6 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 0 |
| 33 | 3 | 1 | 0 | 2 | 1 | 0 | 0 | 0 | 0 |
| 34 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 35 | 3 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 |
| 36 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 1 | 0 |
| 37 | 5 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| 38 | 3 | 0 | 0 | 1 | 0 | 0 | 0 | 3 | 0 |
| 39 | 2 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 40 | 4 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 41 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 |
| 42 | 0 | 2 | 0 | 2 | 0 | 0 | 1 | 1 | 0 |
| 43 | 2 | 0 | 0 | 3 | 1 | 0 | 0 | 1 | 0 |
| 44 | 1 | 0 | 0 | 1 | 4 | 0 | 0 | 1 | 0 |
| 45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 46 | 2 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| 47 | 1 | 0 | 0 | 1 | 1 | 0 | 2 | 0 | 0 |
| 48 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 49 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 50 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 51–60 | 5 | 0 | 0 | 7 | 1 | 0 | 3 | 0 | 0 |
| 61–70 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 71–80 | 3 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| >81 | 2 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 0 |

The first column indicates gaps of length $L$. Pairs of proteins in the one gap database (OGD) were binned into categories based on their respective evolutionary metric (PAM and $f_2$). The numbers in each column indicate the number of pairs in the OGD that display the indicated behavior.

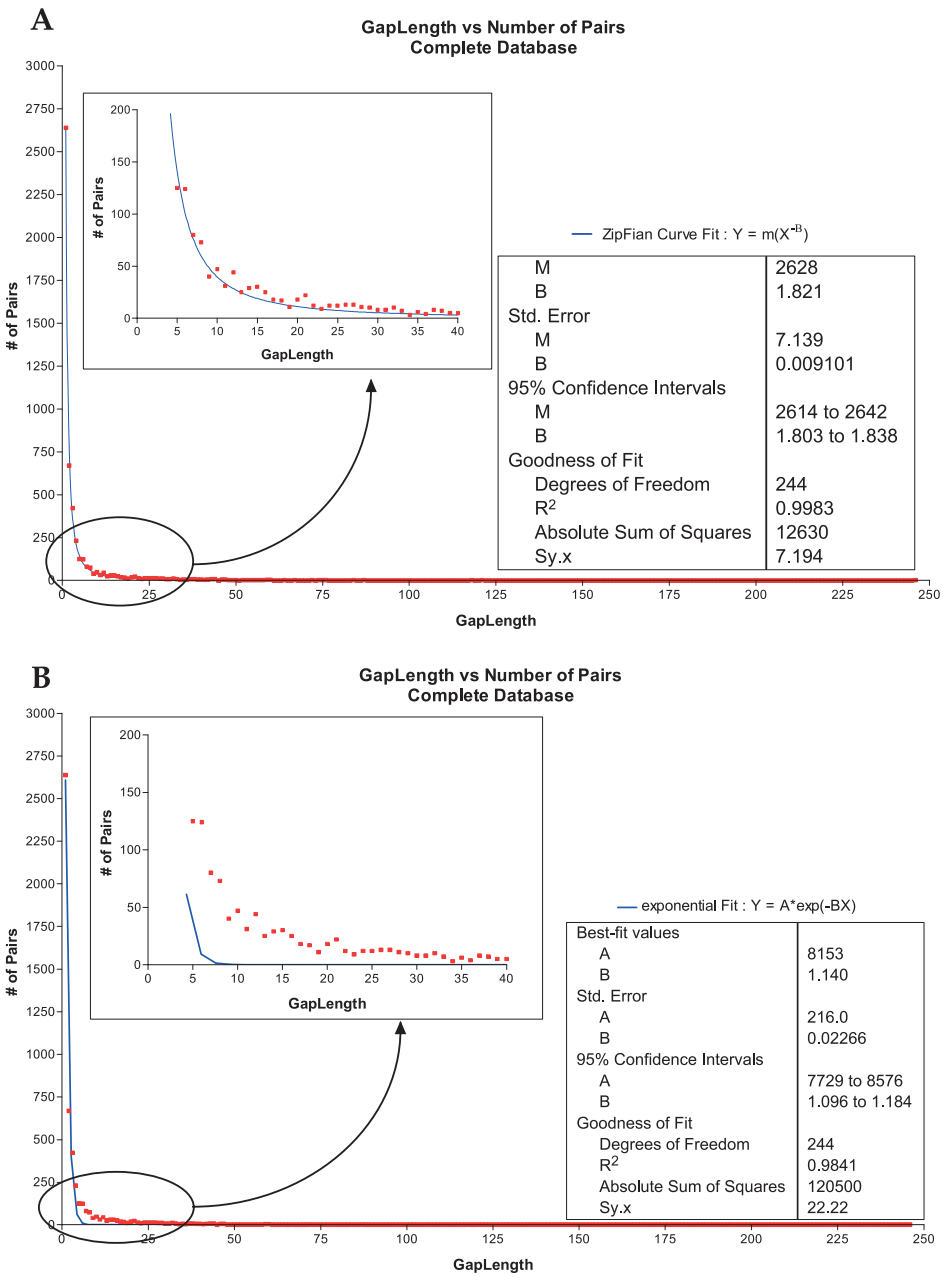


**Figure 1**. Fitting the gap length distribution. A, Gap length distribution of one gap database (OGD) fitted using the Zipfian expression $n = c_1(L^{-c_2})$. B, The same distribution fitted to the exponential decay expression $n = c_1 \exp(-c_2 L)$.

(FMILYVWC). The first includes amino acids that are well known to disrupt secondary structure (DNGPS) as well as two having hydrophilic side-chains that frequently appear at the surface of a protein fold (QE). The second includes amino acids that carry hydrophobic side-chains that frequently lie with within the folded protein structure (FAMILYVW).

These observations are consistent with the notion that indels are accepted by natural selection primarily if they occur in loops and turns, which themselves frequently occur on protein surfaces.[13] Even the "exceptions" are consistent with this notion. Thus, Ala (A) is the least hydrophobic of the amino acid side-chains that are traditionally classed as such,[17] and appears around gaps with higher propensity. Likewise, Cys (C) is not classically grouped with side-chains that form the interior core of protein folds. Of the amino acids having intermediate polarities (CHQST), however, it is among the most hydrophobic. Last, Asn (N) is frequently a structure-disruptor. Although within this database, Asn does not appear in the list of amino acids most likely to be found in or flanking gaps, it is clear that the propensity for Asn is higher in all positions with the exception of position 6. Therefore, Asn conforms with our expectation as a residue likely to be associated with indels.

We next determined whether the propensities are different in proteins suffering different levels of selection pressure (Table 4). Protein pairs

**Table 3.** Distribution of amino acids in and around the gap of the one gap database (OGD)

| | 1 | 2 | 3 | 4 | 5 | 6 | *Z* | Usage (*f*) |
|---|---|---|---|---|---|---|---|---|
| F | ***2.282*** | ***2.585*** | ***2.161*** | ***2.686*** | ***2.979*** | ***2.553*** | ***3.423*** | 4.183 |
| I | ***3.130*** | ***3.716*** | ***3.009*** | ***3.796*** | ***4.012*** | ***3.161*** | ***4.040*** | 5.406 |
| L | ***6.341*** | ***6.422*** | ***6.058*** | ***6.159*** | ***6.687*** | ***6.079*** | ***7.704*** | 9.420 |
| Y | ***2.242*** | ***2.080*** | ***2.181*** | ***2.080*** | ***2.432*** | ***1.945*** | ***2.812*** | 3.303 |
| V | ***4.584*** | 5.755 | ***4.443*** | ***5.372*** | ***7.112*** | ***5.532*** | ***5.253*** | 6.523 |
| W | ***0.666*** | 0.929 | ***0.687*** | ***0.808*** | ***0.973*** | ***0.729*** | ***0.915*** | 1.305 |
| M | 1.393 | ***1.252*** | ***1.555*** | ***1.313*** | ***1.520*** | ***1.641*** | ***1.584*** | 2.211 |
| A | 8.098 | ***9.006*** | ***8.461*** | ***9.855*** | ***8.328*** | ***11.246*** | 7.406 | 7.265 |
| C | ***0.989*** | 1.434 | ***0.989*** | ***1.191*** | ***2.067*** | ***0.851*** | 1.662 | 1.772 |
| R | 4.968 | 4.927 | 4.443 | 4.503 | ***3.647*** | ***4.559*** | 4.741 | 4.988 |
| N | 5.331 | 5.210 | ***5.574*** | ***5.473*** | ***4.559*** | ***3.891*** | 4.751 | 4.570 |
| H | 2.201 | 2.282 | 1.939 | 2.221 | ***2.067*** | ***1.763*** | 2.383 | 2.321 |
| K | 6.422 | 5.957 | ***6.926*** | 5.998 | ***6.018*** | ***5.775*** | ***5.509*** | 5.928 |
| T | 6.078 | 5.897 | 5.876 | 5.755 | ***5.228*** | ***5.714*** | 5.822 | 5.587 |
| D | 6.159 | ***6.987*** | ***6.361*** | ***7.027*** | ***6.809*** | ***6.687*** | 5.185 | 5.391 |
| P | ***6.664*** | 5.594 | ***6.321*** | 5.674 | ***6.018*** | ***5.653*** | ***7.500*** | 5.090 |
| G | ***8.582*** | ***8.057*** | ***8.279*** | 7.573 | ***8.511*** | ***7.416*** | ***7.887*** | 6.715 |
| S | ***9.471*** | ***8.744*** | ***9.249*** | ***9.107*** | ***8.936*** | ***9.119*** | ***9.424*** | 7.287 |
| E | ***8.542*** | ***8.623*** | ***9.269*** | ***8.704*** | ***7.842*** | ***10.517*** | ***6.862*** | 6.527 |
| Q | ***5.775*** | 4.503 | ***6.179*** | 4.665 | ***4.195*** | ***5.046*** | ***5.138*** | 4.207 |

Amino acid occurrence at specific positions in and around the gap were analysed. Positions are defined as:

...XXX**1**_____..._____**2**XXX...

...XXX**35**ZZZ...ZZZ**64**XXX...

Columns 1–6 and *Z* indicate the frequency of occurrence (in %) of specific amino acids at the respective position within the complete one gap database (OGD). The values were calculated from the complete OGD consisting of 4952 sequence pairs. The $\chi^2$ test was performed to determine if the usage of amino acids from each of the positions was a significant departure from the amino acid usage in the complete OGD. Italicized and bold text denotes significance, $P < 0.05$.

separated by PAM and $f_2$ combinations were binned separately using the criteria described in Table 1. Pairs in bin 1A, experiencing "typical" selection pressure, were compared with pairs in bins 2A and 3B, which diverged under strong selection constraints. Pairs in bin 1A were then compared with pairs in bin 2B, which are also diverging under "typical" selection pressure. Pairs in bin 1B, separated by a putative episode of adaptive evolution, were compared with those in bin 3A, which diverged under strong selection constraints. In all, four comparisons were made (1A–2A, 1A–2B, 1A–3B and 1B–3A) by subtracting the propensity of amino acids to appear in each position relative to the gap in the first bin from the corresponding propensity in the second. No large differences were observed (data not shown).

## Seeking compensatory adaptive changes in the flanking segments

Another feature expected for episodes of evolution that include an indel event is a corresponding change in amino acid sequence. Perhaps naively, we might expect that a dramatic change in protein structure (the addition or removal of a segment of polypeptide sequence) would need to be compensated for and refined by smaller changes in the protein sequence (single amino acid replacements) near the indel. We therefore expected the sequence of a pair of proteins to be less conserved near the gapped segment than in the pair overall.

To test this hypothesis, we examined polypeptide segments five positions in length flanking the gap. We first asked whether the percentage identity in the regions flanking the gap was larger or smaller than the percent identity in the pair overall. This was illustrated by plotting the percentage identity in the five amino acids in the left (front) flanking and right (rear) flanking regions, *versus* global percentage identity. Consequently, if differences in the flanking region exist, we will attempt to ascertain whether the difference depended on the global percentage identity of the proteins.

As shown in Figure 2, the polypeptide segment five sites preceding and following the gapped region are not markedly less conserved than the sequence pair overall. This was true also when examining conservation at ten sites preceding and following the gap (data not shown).

We then compared the regions flanking longer *versus* shorter gaps, exploring the hypothesis that the longer the indel, the more likely the protein must compensate for the dramatic change caused by the indel event through point mutation. Interestingly, conservation of amino acids in regions flanking indels covering one, two, three, four, and more than four sites was very similar (data not shown).

In addition, compensation in proteins with lower $f_2$ values was sought, to test the hypothesis that protein pairs with high $f_2$ may not have had time to compensate for the indel event. Again, no major difference in conservation was observed (data not shown).

### Analysis of repeating elements

An unexpected discovery was made during these studies. A large number of indels were associated with amino acid repeats in the non-gapped member of the pair, more than expected by random chance. The presence of repeating amino acids does not appear to depend on gap length or evolutionary distance. For gaps covering a single site, 37.8% are associated with repeats. Similarly, gaps of length two, three and four amino acids are associated with repeats ca 32% of the time in each case. Gaps with length greater than four amino acids are associated with repeats 25.6% of the time. Repeats associated with gaps in the OGD therefore appear five to seven times more often than expected by chance.

In all, approximately 33.8% of the gaps in the OGD are associated with repeating amino acids at their flanking ends (Table 5). If all 20 amino acids were equally abundant in the gapped regions, we might expect a 5% probability of finding any of the 20 amino acids as a dipeptide repeat.

However, the frequency of occurrence of amino acids in the database is not equal (Table 3). This means that the probability for encountering a repeat is higher than would be the case if all amino acids were equally represented in the database. Based on the amino acid usage of the OGD (Table 3), we calculated the probability of encountering repeating amino acids by chance to be 5.75%. This suggests that unequal amino acid usage cannot account for the number of amino acid repeats that are associated with gaps.

Further, the length of the gap will determine the likelihood of finding a repeat within it. The average length of the indel region (flanking residues plus the gap itself) is 6.53. For these and longer indels, it is possible that the level of observed repeats arises from random chance, even though this is not possible for the short indels (for example, those where the indel region covers only three sites, the gapped site and the two flanking sites), where the 37% hit rate cannot be due to chance.

If repeats are a signature of the process underlying indel generation, repeats should also be evident at the DNA level. To test this hypothesis, DNA sequences encoding amino acid repeats that are associated with gaps were extracted and examined to determine if codon usage created a repeat at the DNA level as well. For repeats of length greater than two, a "repeat" was scored if two or more codons within the repeating segment are the same. The results (Table 6) showed that the DNA sequence was also repetitive.

To discern the significance of this bias for repeating codons at the DNA level, we further analysed dipeptide repeats for amino acids encoded by twofold redundant coding systems (FYCNHKDEQ). Here, codon bias is typically small, meaning that if the repeats arise from random chance, repeating codons should be found in 50% of the cases. For four amino acids (FYCH), the number of repeats was too small to be statistically significant. For the remainder, repeating codons were found in 70–86% of the dipeptide repeats encoded by twofold redundant codon systems. This suggests that repeating DNA codons in amino acid repeat segments in and around gaps are significantly favored over non-repeating DNA codons (Table 7).

## Discussion

The principal problem encountered in the attempt to align protein sequences as part of an evolutionary analysis involves the placement of gaps. Without gaps, the only issue faced by an alignment process is the decision where to begin and end the alignment. With gaps, however, alternative alignments that position gaps differently between anchors in a sequence must be considered. The number of alternative gap placements can be large, especially when the protein sequences have diverged to an extent sufficient to remove most of the "anchors" in the alignment.

Gapping in a pairwise alignment is frequently transferred to gapping in multiple sequence alignments.[11] The difficulties encountered are well known. Frequently, users will manually adjust placement of gaps in a multiple sequence alignment, a process that lacks rigor, both mathematical and empirical. This drives the need to develop empirical tools for the proper placement of gaps in pairwise alignments, as a first step towards achieving reliable gap placement in multiple alignments.

Several previous studies of gaps attempted to provide this empirical understanding, and to provide a better consensus for developing tools for placing gaps in alignments.[12,18] These studies illuminated the complexities of gaps, in part because they presented conflicting findings. Most notable was a disagreement on the relationship between the typical length of the gap and the extent of sequence divergence. The disagreements can easily have arisen, however, from the different databases used, their relatively small sizes, and different ways of defining indels.

To build an empirical model for placing gaps requires the user to solve a type of "chicken-or-egg" of circularity. A database containing alignments with indels is a prerequisite for an empirical study to learn the rules for placing gaps. However, rules to place gaps are required to place gaps when creating the database. To manage this problem, we deliberately exploited the large post-genomic database to select for pairs that had suffered (in all likelihood) just one indel. We hoped to diminish the number of gaps that arose from multiple indel events in the same region of the sequence, making standard tools more likely to place gaps correctly in the database.

We also took steps to reduce redundancies by imposing specific cutoffs defined by evolutionary

**Table 4.** Distribution of amino acid in and around the gap separated by evolutionary distance

| $f_2 > 95$ PAM $< 10$ | | | | | Number of pairs in bin: 1006 | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | *Z* | Usage (*f*) |
| F | ***2.087*** | ***2.187*** | 2.684 | 3.082 | ***2.857*** | ***2.381*** | 3.176 | 3.912 |
| I | ***2.883*** | ***3.181*** | ***3.082*** | 3.877 | ***5.238*** | ***2.619*** | ***4.286*** | 5.261 |
| L | ***6.064*** | 6.461 | ***4.771*** | ***5.865*** | ***5.476*** | ***4.048*** | ***7.023*** | 8.820 |
| Y | 3.181 | 1.988 | 2.584 | ***1.690*** | ***3.095*** | ***1.429*** | 2.928 | 3.352 |
| V | 5.169 | 6.362 | 5.567 | 5.567 | ***7.619*** | ***4.762*** | 5.473 | 6.280 |
| W | 0.696 | 0.994 | 0.795 | 0.795 | ***1.667*** | ***1.190*** | 1.263 | 1.137 |
| M | 1.491 | 0.994 | 1.690 | 1.292 | ***0.714*** | ***1.667*** | 1.569 | 2.022 |
| A | 7.455 | ***10.835*** | 7.356 | ***10.636*** | ***8.095*** | ***13.095*** | 6.946 | 7.655 |
| C | 1.491 | 1.590 | 1.889 | 0.895 | ***2.857*** | ***1.190*** | 1.875 | 1.680 |
| R | 5.666 | 4.672 | 4.970 | 4.175 | ***4.048*** | ***4.048*** | 4.650 | 4.757 |
| N | 6.958 | 6.262 | 6.859 | 6.461 | ***3.810*** | ***4.524*** | 4.956 | 5.426 |
| H | 2.187 | 2.684 | 1.988 | 3.082 | ***1.667*** | ***3.095*** | 2.296 | 2.249 |
| K | 4.871 | 4.573 | 4.970 | 4.573 | ***5.238*** | ***5.000*** | 5.434 | 5.886 |
| T | 6.064 | 5.964 | 5.666 | 5.865 | ***5.238*** | ***6.429*** | 6.334 | 5.942 |
| D | 4.871 | 7.058 | 5.368 | 6.262 | ***5.000*** | ***6.667*** | 4.899 | 5.439 |
| P | 4.970 | 6.561 | 4.771 | 6.759 | ***7.381*** | ***6.905*** | ***7.252*** | 5.123 |
| G | 7.952 | 7.256 | 8.052 | 6.660 | ***8.571*** | ***6.190*** | ***8.075*** | 6.908 |
| S | 8.748 | 8.250 | 8.748 | 9.642 | ***8.095*** | ***8.333*** | ***8.458*** | 7.559 |
| E | 8.052 | ***8.449*** | ***8.449*** | ***9.046*** | ***8.810*** | ***10.714*** | ***6.869*** | 6.118 |
| Q | ***8.847*** | 3.579 | ***9.543*** | 3.678 | ***4.286*** | ***5.476*** | ***6.238*** | 4.475 |

| $f_2 > 95$ $10 <= \text{PAM} =< 100$ | | | | | Number of pairs in bin: 247 | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | *Z* | Usage (*f*) |
| F | 2.429 | 2.429 | 1.619 | 1.619 | ***3.896*** | ***2.597*** | 3.731 | 4.296 |
| I | 5.668 | 6.478 | 4.453 | 5.668 | ***6.494*** | ***5.195*** | 3.856 | 5.452 |
| L | 8.502 | 6.883 | 8.907 | 6.883 | ***9.091*** | ***9.091*** | 8.209 | 9.037 |
| Y | 1.215 | 1.619 | 2.429 | 1.619 | ***1.299*** | ***1.299*** | 2.612 | 3.392 |
| V | 6.073 | 3.644 | 5.668 | 4.049 | ***1.299*** | ***3.896*** | 4.602 | 6.717 |
| W | 0.810 | 0.810 | 0.810 | 0.405 | ***0.000*** | ***0.000*** | 0.871 | 1.431 |
| M | 0.405 | 3.644 | 1.619 | 1.619 | ***1.299*** | ***2.597*** | 2.488 | 1.940 |
| A | 9.717 | 10.931 | 8.502 | ***17.004*** | ***10.390*** | ***14.286*** | 6.716 | 7.341 |
| C | 0.810 | 1.215 | 1.215 | 1.215 | ***1.299*** | ***1.299*** | 0.995 | 2.033 |
| R | 6.478 | ***10.121*** | 4.049 | 4.049 | ***1.299*** | ***2.597*** | 6.343 | 4.626 |
| N | 6.883 | 4.049 | 7.287 | 5.668 | ***1.299*** | ***2.597*** | 4.229 | 4.883 |
| H | 0.405 | 2.024 | 1.215 | 0.405 | ***2.597*** | ***0.000*** | 1.990 | 2.181 |
| K | 6.073 | 3.644 | 4.049 | 5.668 | ***3.896*** | ***6.494*** | 6.219 | 6.090 |
| T | 5.668 | 3.644 | 4.858 | 6.478 | ***10.390*** | ***3.896*** | 5.473 | 6.151 |
| D | 5.668 | 8.502 | 6.478 | 4.858 | ***11.688*** | ***7.792*** | 5.846 | 5.119 |
| P | 4.858 | 4.453 | 4.453 | 3.239 | ***6.494*** | ***2.597*** | ***7.214*** | 4.982 |
| G | 3.239 | 8.907 | 4.049 | 6.883 | ***6.494*** | ***5.195*** | 8.831 | 6.813 |
| S | 9.312 | 6.478 | 8.097 | 8.502 | ***5.195*** | ***6.494*** | 7.214 | 7.313 |
| E | ***12.551*** | 7.692 | ***14.980*** | ***10.931*** | ***9.091*** | ***15.584*** | 7.214 | 6.152 |
| Q | 3.239 | 2.834 | 5.263 | 3.239 | ***6.494*** | ***6.494*** | 5.348 | 4.051 |

| $80 <= f_2 =< 95$ PAM $< 10$ | | | | | Number of pairs in bin: 1079 | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | *Z* | Usage (*f*) |
| F | ***1.112*** | 2.502 | ***1.297*** | ***2.132*** | ***2.168*** | ***2.168*** | 3.311 | 3.983 |
| I | ***2.966*** | 3.429 | ***2.966*** | ***2.780*** | ***3.252*** | ***2.168*** | ***3.750*** | 5.045 |
| L | ***6.487*** | 7.322 | ***5.653*** | ***6.209*** | ***4.878*** | ***6.504*** | 8.378 | 9.307 |
| Y | 1.761 | 2.317 | ***1.483*** | 2.317 | ***1.897*** | ***2.168*** | ***2.234*** | 3.144 |
| V | ***3.985*** | 4.727 | ***3.336*** | 4.541 | ***6.775*** | ***5.691*** | ***4.927*** | 6.395 |
| W | 0.649 | 0.927 | 0.556 | 0.649 | ***0.813*** | ***0.542*** | ***0.738*** | 1.222 |
| M | 1.483 | 1.019 | 1.019 | 1.668 | ***2.168*** | ***1.897*** | ***1.456*** | 2.213 |
| A | 7.878 | ***9.639*** | ***9.731*** | ***11.214*** | ***7.859*** | ***11.111*** | ***8.458*** | 7.320 |
| C | 1.112 | 1.483 | 1.019 | 1.390 | ***2.168*** | ***0.813*** | 1.775 | 1.814 |
| R | 5.283 | 5.375 | 5.283 | 5.468 | ***3.523*** | ***6.504*** | 5.286 | 5.004 |
| N | 4.078 | 5.097 | 4.078 | 4.171 | ***5.691*** | ***3.523*** | 4.229 | 4.430 |
| H | 2.132 | 2.132 | 2.595 | 2.317 | ***0.813*** | ***1.355*** | 2.613 | 2.346 |
| K | 6.673 | 5.931 | 7.507 | 5.746 | ***5.962*** | ***5.962*** | 5.087 | 5.977 |
| T | 6.951 | 5.746 | 5.561 | 5.931 | ***5.149*** | ***5.962*** | 5.825 | 5.493 |
| D | 4.634 | 6.766 | 5.468 | ***7.414*** | ***6.775*** | ***7.588*** | 4.768 | 5.321 |
| P | ***7.600*** | 5.468 | 7.414 | 5.746 | ***7.859*** | ***4.607*** | ***7.840*** | 5.536 |
| G | ***11.677*** | 7.322 | ***10.843*** | 8.063 | ***9.485*** | ***9.214*** | ***7.680*** | 6.738 |
| S | 7.414 | 9.453 | 8.526 | 8.804 | ***10.027*** | ***7.046*** | ***9.515*** | 7.566 |
| E | ***9.639*** | 8.897 | ***9.824*** | 8.712 | ***9.214*** | ***11.111*** | 7.102 | 6.798 |
| Q | ***6.395*** | 4.449 | 5.839 | 4.634 | ***3.523*** | ***3.794*** | 5.027 | 4.346 |

(*continued*)

**Table 4** continued

$80 <= f_2 = < 95\ 10 <= \text{PAM} = < 100$ Number of pairs in bin: 906

| | 1 | 2 | 3 | 4 | 5 | 6 | Z | Usage (*f*) |
|---|---|---|---|---|---|---|---|---|
| F | 3.863 | 3.422 | 2.759 | 3.863 | ***3.472*** | ***3.125*** | ***3.190*** | 4.336 |
| I | 3.642 | ***3.201*** | ***3.201*** | 3.753 | ***3.472*** | ***3.125*** | ***3.636*** | 5.303 |
| L | ***5.850*** | ***6.512*** | ***6.291*** | 6.954 | ***8.681*** | ***7.639*** | 8.473 | 9.561 |
| Y | 1.876 | 2.318 | 2.097 | 2.649 | ***3.125*** | ***2.431*** | 3.431 | 3.316 |
| V | 4.636 | 4.525 | 4.746 | 4.967 | ***7.986*** | ***4.167*** | ***5.146*** | 6.621 |
| W | 0.662 | 1.104 | 0.993 | 0.773 | ***0.347*** | ***0.694*** | ***0.480*** | 1.394 |
| M | 1.325 | 1.435 | 1.876 | 1.214 | ***0.000*** | ***1.736*** | 1.509 | 2.119 |
| A | ***9.492*** | 7.285 | 9.051 | 8.168 | ***9.375*** | ***9.028*** | 6.792 | 6.960 |
| C | 1.214 | 1.766 | 0.993 | 1.545 | ***1.736*** | ***1.389*** | 1.475 | 1.959 |
| R | 4.415 | 4.084 | 3.422 | 3.091 | ***4.167*** | ***3.472*** | 4.494 | 4.923 |
| N | 5.629 | 5.519 | 5.740 | 5.298 | ***4.861*** | ***3.472*** | ***6.003*** | 4.486 |
| H | 1.766 | 2.208 | 1.876 | 1.656 | ***3.472*** | ***1.389*** | 2.161 | 2.292 |
| K | 7.395 | 6.291 | 7.726 | 5.960 | ***6.944*** | ***5.208*** | 5.626 | 5.897 |
| T | 5.077 | 6.291 | 6.402 | 6.181 | ***4.167*** | ***5.903*** | 5.935 | 5.537 |
| D | 6.843 | 7.285 | 7.174 | ***7.616*** | ***7.986*** | ***7.986*** | 4.803 | 5.401 |
| P | ***7.506*** | 6.623 | 6.623 | 6.071 | ***4.861*** | ***6.597*** | ***8.370*** | 5.186 |
| G | 7.506 | 7.506 | 7.506 | 7.506 | ***7.986*** | ***7.292*** | 7.581 | 6.782 |
| S | ***10.044*** | 9.713 | 9.161 | 9.603 | ***7.292*** | ***12.153*** | ***10.223*** | 7.331 |
| E | 7.064 | 7.616 | 7.285 | 8.278 | ***5.208*** | ***7.986*** | 6.244 | 6.422 |
| Q | 4.194 | 5.188 | 5.077 | 4.857 | ***4.861*** | ***5.208*** | 4.425 | 4.174 |

$f_2 < 80\ \text{PAM} < 10$ Number of pairs in bin: 278

| | 1 | 2 | 3 | 4 | 5 | 6 | Z | Usage (*f*) |
|---|---|---|---|---|---|---|---|---|
| F | 3.597 | 3.597 | 2.518 | 3.597 | ***4.545*** | ***5.682*** | 3.499 | 4.070 |
| I | 2.518 | 6.115 | 1.799 | 2.878 | ***2.273*** | ***4.545*** | ***3.289*** | 5.109 |
| L | 6.475 | 6.835 | 5.755 | 5.396 | ***9.091*** | ***5.682*** | 9.517 | 9.652 |
| Y | 2.878 | 0.719 | 3.237 | 1.439 | ***5.682*** | ***1.136*** | 2.379 | 3.050 |
| V | 3.597 | 6.115 | 3.597 | 8.273 | ***5.682*** | ***7.955*** | 5.528 | 6.428 |
| W | 0.360 | 0.719 | 0.000 | 0.360 | ***2.273*** | ***0.000*** | 1.190 | 1.269 |
| M | 1.439 | 1.079 | 1.439 | 2.158 | ***2.273*** | ***2.273*** | 1.749 | 2.439 |
| A | 10.072 | 10.432 | ***11.151*** | 10.791 | ***7.955*** | ***10.227*** | 7.838 | 6.768 |
| C | 1.079 | 0.360 | 1.079 | 0.000 | ***1.136*** | ***0.000*** | 1.749 | 1.807 |
| R | 3.237 | 4.676 | 2.878 | 5.755 | ***4.545*** | ***6.818*** | 5.738 | 5.252 |
| N | 3.597 | 4.317 | 3.957 | 3.237 | ***0.000*** | ***2.273*** | 3.709 | 4.190 |
| H | 1.439 | 3.237 | 1.079 | 3.597 | ***2.273*** | ***3.409*** | 2.729 | 2.356 |
| K | 3.957 | 3.957 | 5.036 | 4.317 | ***9.091*** | ***4.545*** | 4.689 | 6.062 |
| T | 5.396 | 5.396 | 5.036 | 6.475 | ***5.682*** | ***4.545*** | 6.088 | 5.431 |
| D | 6.115 | 6.475 | 2.158 | ***9.353*** | ***3.409*** | ***4.545*** | 5.948 | 5.427 |
| P | ***10.791*** | 4.317 | ***10.791*** | 5.036 | ***4.545*** | ***1.136*** | 6.438 | 5.368 |
| G | 7.554 | 9.353 | 7.914 | 6.475 | ***9.091*** | ***6.818*** | 6.788 | 6.388 |
| S | 10.072 | 8.993 | ***12.590*** | 8.273 | ***10.227*** | ***11.364*** | 9.237 | 7.560 |
| E | 8.273 | 7.554 | 10.432 | 7.194 | ***6.818*** | ***11.364*** | 6.998 | 6.998 |
| Q | 7.554 | 5.755 | 7.554 | 5.396 | ***3.409*** | ***5.682*** | 4.899 | 4.374 |

$f_2 < 80\ 10 <= \text{PAM} = < 100$ Number of pairs in bin: 1427

| | 1 | 2 | 3 | 4 | 5 | 6 | Z | Usage (*f*) |
|---|---|---|---|---|---|---|---|---|
| F | ***2.032*** | ***2.242*** | ***2.102*** | ***2.102*** | ***2.993*** | ***1.995*** | 3.974 | 4.456 |
| I | ***2.803*** | ***3.714*** | ***2.873*** | 4.415 | ***3.741*** | ***3.990*** | ***4.753*** | 5.932 |
| L | ***6.307*** | ***5.536*** | ***6.657*** | ***5.886*** | ***7.232*** | ***6.234*** | ***6.337*** | 9.791 |
| Y | 2.242 | 2.172 | 2.242 | ***2.032*** | ***1.247*** | ***2.244*** | 3.142 | 3.459 |
| V | ***4.485*** | 7.008 | ***4.275*** | 5.746 | ***7.731*** | ***6.983*** | ***5.505*** | 6.729 |
| W | 0.701 | 0.841 | 0.561 | 1.121 | ***0.748*** | ***0.748*** | 0.913 | 1.427 |
| M | 1.472 | ***1.121*** | 1.682 | ***0.841*** | ***2.743*** | ***0.748*** | ***1.531*** | 2.354 |
| A | 7.148 | 7.779 | 7.358 | 7.989 | ***7.980*** | ***10.723*** | 7.116 | 7.257 |
| C | ***0.420*** | 1.331 | ***0.280*** | 1.261 | ***1.746*** | ***0.249*** | 1.477 | 1.651 |
| R | 4.625 | 4.415 | 4.485 | 4.765 | ***3.242*** | ***3.990*** | ***3.571*** | 5.121 |
| N | 5.046 | 4.695 | 5.676 | ***6.237*** | ***5.736*** | ***4.239*** | 4.699 | 4.254 |
| H | 3.013 | 2.032 | 1.752 | 1.892 | ***2.494*** | ***0.998*** | 2.309 | 2.372 |
| K | 7.218 | 7.498 | ***8.129*** | 7.568 | ***5.985*** | ***6.983*** | 6.230 | 5.872 |
| T | 6.307 | 6.237 | 6.307 | 5.046 | ***4.489*** | ***5.237*** | 4.995 | 5.452 |
| D | ***7.849*** | 6.797 | ***8.059*** | 6.797 | ***7.731*** | ***5.237*** | 6.015 | 5.439 |
| P | ***6.167*** | 4.835 | 5.886 | 5.186 | ***3.990*** | ***6.234*** | ***7.143*** | 4.596 |
| G | ***8.479*** | ***9.180*** | 7.779 | 8.199 | ***8.229*** | ***7.731*** | ***8.405*** | 6.626 |
| S | ***11.002*** | 8.199 | ***9.741*** | ***8.970*** | ***10.474*** | ***9.726*** | ***10.607*** | 6.794 |
| E | ***8.409*** | ***9.530*** | ***9.460*** | ***8.619*** | ***7.481*** | ***10.474*** | 6.928 | 6.518 |
| Q | 4.275 | 4.835 | 4.695 | 5.326 | ***3.990*** | ***5.237*** | 4.350 | 3.898 |

Data collected from the one gap database (OGD) separated into specific bins. Amino acid occurrence at specific positions in and around the gap was analysed. The positions are defined as:

...XXX**1**_____..._____**2**XXX...

...XXX**35**ZZZ...ZZZ**64**XXX...

Columns 1–6 and *Z* indicate the frequency of occurrence (in %) of specific amino acids within the database represented by the Table. The values were calculated from the complete OGD consisting of 4952 sequence pairs. The $\chi^2$ test was implemented to determine if the usage of amino acids from each of the positions were a significant departure from the amino acid usage in the complete OGD. Italicized and bold text denotes significance of $P < 0.05$.

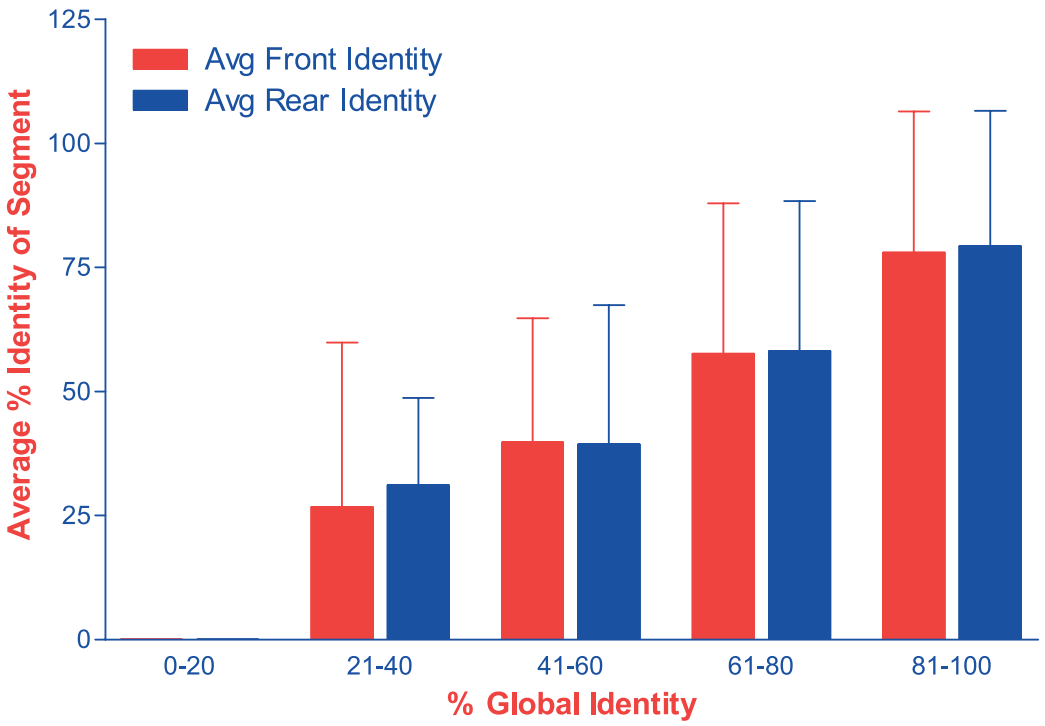


**Figure 2**. Gap flanking region (five amino acids) analysis. A histogram comparing the identity of amino acids flanking the gapped region *versus* the complete sequence (global) identity of each sequence pair. Pairs in the OGD were sorted into bins based on their global identity. Subsequently, the percentage identity of five amino acids before the gap (front/towards the $NH_2$ end) is averaged for the respective bins. Similar calculations were performed for the five amino acids immediately following the gap (rear/towards the COOH end). The error bars reflect the standard deviation of segment identity within that bin.

distance. The cutoffs imposed reduced the number of gaps introduced by entry errors, in sequence fragments, and other features of the database that may create false indels. In each case, we excluded many sequence pairs from the analysis; therefore, a sufficiently large database was required to enable these procedures.

The length of the gaps in the OGD is independent of the evolutionary distance separating the two sequences, as measured using either the $f_2$ value (a good measure of time) as a metric for distance or the PAM distance metric. This is consistent with the view that gaps in the OGD arise from single indel events.

Given a reliable set of pairs separated by a single indel event, various features of the indel event could be inferred. These could be conveniently divided into two classes, those that reflected selection pressure acting at the level of the protein, and those that reflected the mechanism occurring at the DNA level by which the indel event arises.

The first concerns the types of indels that can be suffered by a protein without its functional behavior being destroyed, which would typically cause it to be lost in subsequence history through purifying selection. This, in turn, would prevent it from appearing in any modern database (except, perhaps, as a pseudogene). This discussion focuses on proteins, their folding and their reactivity.

Signals of purifying selection were easy to find. First, amino acids found in and around gaps were quite different from those found in the database as a whole. This suggests that natural selection strongly influences the acceptance of gaps in a genome. Further, the predominant amino acids are secondary structure breakers, consistent with the notion that an indel survives selective pressure most frequently if it involves a segment that is neither a standard alpha helix nor a standard beta strand.

Further, the physical chemical behavior of the side-chains was easily correlated with their propensities to appear in or near a gap. Gapped regions prefer hydrophilic residues and classic structure-breaking amino acids such as Pro, Gly, Asp and Ser. This is consistent with the notion that segments on the surface of proteins, which tend to be hydrophilic, as well as breaks in secondary structure, are more tolerant of indel events. Consistent with this notion are decreased propensities for hydrophobic residues such as Phe, Met, Ile, Leu, Tyr, Val, Trp and Cys in and around gaps.

Likewise, the Zipfian distribution in the gap length distribution is consistent with the notion that indels occur in coils and loops.[12] This explanation is based on three assumptions: (a) that the ends of indeled regions must lie near in space; (b) the sequence between the ends adopts a random coil structure; and (c) that the Flory rules of statistical mechanics applied to polymers govern the conformation of those coils. If true, the volume

**Table 5.** Determining proportion of repeating elements in the one gap database

| Gap length $L$ | Total number of pairs | | Gaps containing repeats | | Gaps lacking repeats | |
|---|---|---|---|---|---|---|
| **1** | 2638 | **Avg PAM** 18<br>**Avg $f_2$** 83 | 998 | **Avg PAM** 13<br>**Avg $f_2$** 85 | 1640 | **Avg PAM** 20<br>**Avg $f_2$** 81 |
| **2** | 669 | **Avg PAM** 17<br>**Avg $f_2$** 83 | 212 | **Avg PAM** 15<br>**Avg $f_2$** 83 | 457 | **Avg PAM** 18<br>**Avg $f_2$** 83 |
| **3** | 421 | **Avg PAM** 19<br>**Avg $f_2$** 83 | 136 | **Avg PAM** 18<br>**Avg $f_2$** 84 | 285 | **Avg PAM** 19<br>**Avg $f_2$** 83 |
| **4** | 232 | **Avg PAM** 17<br>**Avg $f_2$** 85 | 76 | **Avg PAM** 16<br>**Avg $f_2$** 87 | 156 | **Avg PAM** 17<br>**Avg $f_2$** 84 |
| **>4** | 992 | **Avg PAM** 14<br>**Avg $f_2$** 86 | 254 | **Avg PAM** 15<br>**Avg $f_2$** 86 | 738 | **Avg PAM** 13<br>**Avg $f_2$** 86 |
| **Total** | 4952 | | 1676 | | 3276 | |

The ODG is binned based on gap length. Indels with repeats are defined by the following five criteria: **1** = **2**, **3** = **5**, **6** = **4**, **5** = **7** and **6** = **8**. For gaps of length one, only compare three criteria: **1** = **2**, **3** = **Z** and **6** = **Z**. Gaps of length two compares four criteria: **1** = **2**, **3** = **5**, **6** = **4** and **5** = **6**.

...XXX**1**_____..._____**2**XXX...
...XXX**35**ZZZ...ZZZ**64**XXX...

If any of the five criteria were satisfied, then it is considered to have repeats in the indel pair. In addition, the average distance separating pairs of protein sequences in each bin has been calculated based on both PAM and $f_2$.

occupied by a typical peptide scales with the length of the peptide raised to the 1.8 power,[19,20] and the probability that their ends will be together scales inversely with the 1.8 power of the length of the indeled region. This is remarkably close to what is observed empirically.[12]

However, some of the expected signatures of purifying selection were not observed in these data. For example, indel events are not typically accompanied by episodes of rapid amino acid replacement. We expected to see this, assuming that the (presumed) dramatic gain or loss of a segment of polypeptide chain would cause the selection of compensatory amino acid replacements.

In fact, we found no strong evidence to support compensatory amino acid sequence change in regions flanking indels as a response to the gain/loss of peptide sequence. Compensatory amino acid replacement might in fact occur, of course, in segments distant in the linear sequence but close to the indel in the three-dimensional fold. This hypothesis is much more difficult to test, as it requires structural information currently not available to many of the protein sequence pairs within the OGD.

Likewise, only modest signals were seen when we compared protein sequence pairs that have diverged under strong purifying selection, or under adaptive changes. Amino acid propensities in and around gaps differ, but not strongly, depending on whether the gap is being introduced during an episode of adaptive evolution or under strong functional constraints. This suggests that indels are not particularly important elements of adaptive evolution. This implies that gap placement tools need not intensely consider the impact of adaptive or constrained evolution on gaps. It may, however, suggest that subtle differences may indicate levels of adaptation during protein sequence divergence, especially when combined with other metrics.[21]

Based on earlier studies, we expected that the signatures of natural selection operating at the level of the protein would overwhelm any signature that might arise from the mechanism by which indels occur at the DNA level. Well known, for example, is the appearance of indels in noncoding, repetitive sequences.[22,23] Approximately 2% of the human genome is estimated to be in the form of tandem repeats.[24] The fundamental role of tandem repeats currently remains a mystery. It has been observed, however, that within repeat regions slippage occurs more frequently than point mutations.[25] This suggests that slipstrand mispairing, and indelling alike, may play an important role in DNA sequence evolution. Repeats have been occasionally reported within coding regions in studies targeted at specific genomes.[25,26]

This work establishes that the association of gaps in protein alignments with repeats in the DNA sequence is widespread. We believe that this association is best explained as a signature of the mechanism by which indels arise. When examining initial indels, this signature is not yet lost by purifying selection operating at the level of the protein.

The molecular mechanism for the introduction of indels into DNA is not yet fully understood, although several theories have been postulated. For small indels (involving $<10$ nucleotides), slipstrand mispairing is frequently hypothesized.[27–29]

Other mechanisms have been postulated for longer indels. The presence of long stem–loop structures in introns has been observed to promote indels.[30,31] Recombination events, such

**Table 6.** Gap associated repeating amino acids

| Length of Repeat | F | | I | | L | | Y | | V | | W | | M | | A | | C | | R | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R | N | R | N | R | N | R | N | R | N | R | N | R | N | R | N | R | N | R | N |
| 2 | 14 | 7 | 18 | 14 | 34 | 52 | 9 | 7 | 32 | 29 | 2 | | 10 | | 69 | 77 | 1 | 3 | 13 | 29 |
| 3 | 3 | | 4 | | 8 | 8 | 2 | | 3 | 1 | | | 1 | | 33 | 7 | 1 | | 5 | 3 |
| 4 | 1 | | | | 12 | | 1 | | 3 | | | | | | 25 | | | | 4 | 1 |
| 5 | | | 1 | | 3 | | | | 1 | | | | | | 12 | | | | 1 | |
| 6 | | | | | 5 | | | | | | | | | | 1 | | | | | |
| 7 | | | | | 3 | | | | | | | | | | 1 | | | | | |
| 8 | | | | | 1 | | | | | | | | | | 1 | | | | | |
| 9 | | | | | 1 | | | | | | | | | | 3 | | | | | |
| 10 | | | | | | | | | | | | | | | | | | | | |
| 11 | | | | | | | | | | | | | | | | | | | | |
| 12 | | | | | | | | | | | | | | | | | | | | |
| 13 | | | | | | | | | | | | | | | | | | | | |
| 14 | | | | | | | | | | | | | | | | | | | | |
| 15 | | | | | | | | | | | | | | | | | | | | |
| 16 | | | | | | | | | | | | | | | | | | | | |
| 17 | | | | | | | | | | | | | | | | | | | | |
| 18 | | | | | | | | | | | | | | | | | | | | |
| 24 | | | | | | | | | | | | | | | | | | | | |
| 28 | | | | | | | | | | | | | | | | | | | | |
| Total | 18 | 7 | 23 | 14 | 67 | 60 | 12 | 7 | 39 | 30 | 2 | 0 | 11 | 0 | 145 | 84 | 2 | 3 | 23 | 33 |
| AA Total | 25 | | 37 | | 127 | | 19 | | 69 | | 2 | | 11 | | 229 | | 5 | | 56 | |
| %AATotal | 72 | 28 | 62 | 38 | 53 | 47 | 63 | 37 | 57 | 43 | 100 | 0 | 100 | 0 | 63 | 37 | 40 | 60 | 41 | 59 |

| Length of Repeat | N | | H | | K | | T | | D | | P | | G | | S | | E | | Q | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R | N | R | N | R | N | R | N | R | N | R | N | R | N | R | N | R | N | R | N | Total |
| 2 | 45 | 20 | 4 | 7 | 53 | 32 | 35 | 32 | 79 | 31 | 38 | 35 | 73 | 60 | 70 | 122 | 82 | 32 | 61 | 22 | 1353 |
| 3 | 13 | | 4 | | 17 | | 16 | 3 | 20 | | 18 | 3 | 42 | 7 | 28 | 14 | 48 | | 20 | | 332 |
| 4 | 4 | | 1 | | 5 | | 7 | | 5 | | 8 | | 14 | | 15 | | 23 | | 7 | | 136 |
| 5 | 4 | | | | | | 1 | | 3 | | 1 | | 7 | | 7 | | 15 | | 6 | | 62 |
| 6 | 1 | | 2 | | | | | | 2 | | 5 | | 6 | | 3 | | 6 | | 7 | | 38 |
| 7 | 1 | | | | 1 | | | | 2 | | 1 | | 4 | | 1 | | 4 | | 2 | | 20 |
| 8 | 2 | | 2 | | | | 1 | | 1 | | 1 | | 2 | | | | 2 | | 1 | | 14 |
| 9 | 1 | | | | | | 1 | | 2 | | 1 | | | | 3 | | 2 | | 1 | | 15 |
| 10 | 1 | | | | | | | | | | | | | | 2 | | 2 | | 3 | | 8 |
| 11 | | | 2 | | | | | | | | | | | | | | | | 9 | | 11 |
| 12 | | | | | | | | | | | 2 | | 1 | | 2 | | | | 1 | | 6 |
| 13 | | | | | | | | | | | 1 | | 1 | | | | 9 | | 3 | | 14 |
| 14 | | | | | | | | | | | 1 | | 1 | | | | 1 | | 2 | | 5 |
| 15 | | | | | | | | | | | | | | | | | | | 3 | | 3 |
| 16 | | | | | | | | | | | | | 1 | | | | | | | | 1 |
| 17 | | | | | | | | | | | 1 | | 1 | | | | | | | | 2 |
| 18 | | | | | | | | | | | | | | | 1 | | | | | | 1 |
| 24 | | | | | | | | | | | | | 1 | | | | | | | | 1 |
| 28 | | | | | | | | | 1 | | | | | | | | | | | | 1 |
| Total | 72 | 20 | 15 | 7 | 76 | 32 | 61 | 35 | 115 | 31 | 78 | 38 | 154 | 67 | 132 | 136 | 194 | 32 | 126 | 22 | **2023** |
| AA Total | 92 | | 22 | | 108 | | 96 | | 146 | | 116 | | 221 | | 268 | | 226 | | 148 | | |
| %AATotal | 78 | 22 | 68 | 32 | 70 | 30 | 64 | 36 | 79 | 21 | 67 | 33 | 70 | 30 | 49 | 51 | 86 | 14 | 85 | 15 | |

The amino acid identity of each repeating element associated with gaps in the ODG is scored. Each repeating element is also checked for repeating codons at the DNA level: R, denoting DNA codon repeating or N, denoting DNA codon not repeating. For repeats of length greater than two, the element is designated as R if any codon within the element is repeated.

as unequal crossing-over, have also been proposed as sources for longer indels.[32–34] The mechanism for unequal recombination currently proposed is also dependent on the presence of tandem repeats.[29]

Slip-strand mispairing and unequal crossing-over differ, in that the latter is an interhelical event involving DNA from two different chromosomes. Indels arising from recombination can potentially give rise to gaps of any length without bias towards short or long.

Molecular mechanisms for gapping at the DNA level are likely to be more easily detected in non-coding DNA, since it need not suffer through evolutionary adaptation or constraints. Consistent with this idea, tandem repeats have been observed to be more numerous in non-coding regions as compared to coding regions.[35–37] To date, relatively little work has examined indels in non-coding DNA. Such studies are a challenge in any case, as regions of non-coding DNA are not well cataloged. Further, repeats in these regions frequently make confident alignment difficult. As a consequence, the size of the database analysed is typically small. As more genomic sequences from closely related organisms (such as chimp and human)

**Table 7.** Determine the proportion of DNA repeats for dipeptides composed of twofold redundant amino acids

| | Repeating DNA | Non-repeating DNA |
|---|---|---|
| F | 14 | 7 |
| Y | 9 | 7 |
| C | 1 | 3 |
| N | 45 | 20* |
| H | 4 | 7 |
| K | 53 | 32* |
| D | 79 | 31* |
| E | 82 | 32* |
| Q | 61 | 22* |
| Total | 348 | 161* |

Dipeptide repeats in and around gap regions were analysed. Dipeptides from amino acids encoded by two DNA codons (twofold redundant amino acids) were scored to ascertain if the DNA codon is repeating within this element. *$P > 0.005$.

become available, it should be possible to explore this issue in greater depth and to learn more about the molecular mechanisms by which DNA gains and loses segments.

In some cases, a simple experiment can be done with the OGD to address the different hypotheses. For example, repeats in the protein sequence may reflect a process occurring at the DNA level, as opposed to a process that involves selection. If the repeats were a DNA-base phenomenon, one should see the repeats extended to the DNA level. This is in fact observed.

Mechanisms for creating indels need not be the same in different organisms, of course. This would be especially so if the intrinsic structure of DNA, which is common to all forms of life, is not the dominant feature in creating indels. In this work, we aggregate proteins from all forms of life on Earth, where the only bias is that created in the primary database itself. Future work, again enabled by genomic sequencing more densely across the tree of life, will help identify these differences.

Why are these results important from a technological perspective? Much of the structure of various genomes is inferred directly from DNA sequence without confirmation by experimental work. This includes identifying open reading frames, finding start and stop signals, placing introns within coding regions, and using evolutionary homology to assign coding regions in new genomes based on the assignment of coding regions in known genomes.

Each of these processes is well known to contain errors. Manual curation of databases is possible, of course, with organisms that have been explored in great detail by experimentalists. In the yeast genome, for example, where manual curation and experimental work have both been intensive, a large number of the automatically found ORFs can be labeled "dubious". In less well-studied genomes, however, this is not possible.

Misplaced introns, introns/exon boundaries, and start and stop signals create gaps in multiple sequence alignments. Given a strong empirical understanding about where gaps should be placed in real proteins diverging under functional constraints, it should be possible to develop automated gene analysis packages that distinguish real gaps from false gaps. This has been done anecdotally as well, most notably in the identification of an intron's misassignment in calcineurin.[38] This approach can be extended to analyse families that are entirely false. These arise, for example, through the translation of the incorrect strand. Such errors propagate through the database, creating "faux families", entire collections of putative protein sequences assigned to new genomes from an incorrect assignment to an original genome. Here again, the patterns of evolution, including gap placement, in faux families are not as expected for a real family. Thus, this effort holds the possibility of improving automated family assignments.

## Methods

We began with a collection of families of independently evolving modules[39] contained within the MasterCatalog (EraGen Biosciences, Madison WI).[40] These were built from the Genbank (version 1.20), Ensemble4, PIR4, and PRF4 databases. The MasterCatalog delivered a set of pre-computed families, trees and multiple sequence alignments for every nuclear family in the known global proteome. Individual gap placements from the MasterCatalog were recomputed for specific analyses as described below.

All pairs of proteins within each family were then characterized by two distance metrics, the classical point accepted mutation (PAM) distance (reflecting distance in protein sequence space)[3] and an $f_2$ value,[14] which represents the fraction of conserved nucleotides at the third position of twofold redundant codon systems where the encoded amino acid itself is conserved.[14] This has been shown to be an effective tool for ordering dates of events in a molecular system.[14]

Pairs of homologous proteins were then sought whose intervening evolutionary history most likely included exactly one indel event. The strategy to do so was designed to avoid common database deficiencies, including incorrect gene finding, misplaced start/stop points, fragmentary entries, and overlooked/over-predicted introns. Pairs of proteins whose amino acid sequences were separated by less than four PAM units were discarded, as these might include duplicate entries of the same protein, where gapping appears to be dominated by entry errors. A similar filter removed all gene pairs having an $f_2$ value greater than 0.98. Pairs that had fewer than 50 characters available for the $f_2$ analysis were also discarded, to ensure that the $f_2$ values calculated had an acceptably small standard error.

From the remaining pairs, a OGD was created as follows. For each family, a multiple sequence alignment was constructed. In the order of decreasing $f_2$ distances, the alignment for each pair that met the conditions listed above was extracted from the MSA. The pairwise alignments were then inspected for gaps. If the pairwise alignment contained exactly one gap and the gap was not within six amino acid residues of the beginning or end of the pairwise alignment, then it was retained for the next step. This filter eliminated many gaps that arise

from sequence fragments or modularization artifacts. Protein sequence pairs were retained if and only if neither of the two sequences was represented in a pair previously retained. This process was continued until all pairs within the family were examined.

The resulting database held non-redundant pairs of proteins that contained a single gap in their pairwise alignments. Choosing the pair with the highest $f_2$ when more than one pair within the same gene family was available ensured the selection of the pair within the family that diverged most recently, assuming that silent transitions provide an accurate clock.[14]

The selected sequences were then gathered into a MySQL database resource. This resource, available electronically†, is a set of pairwise alignments for protein pairs that are separated by their first indel event recorded in GenBank 1.20.

Since it is not possible to examine all the pairs in the OGD to validate the gaps (due to lack of genomic information in some species as well as time restrictions), we have sampled 15 pairs of protein from species with existing genomic database (primarily man, mouse and rat). Our analyses of long gaps revealed that several are from splice variants, while gaps of length ranging from one to four residues are primarily indels. Our comparisons were chosen from interspecies protein pairs. In all cases examined, short indels were found to be embedded within corresponding exon regions. This suggests that the gap reflect true indels. As with gaps giving rise to splice variants, at the protein level, these are also true indels and not gap artifacts introduced by alignment algorithms.

Patterns of insertion and deletion might conceivably be different in proteins that are diverging slowly due to strong functional constraints, *versus* proteins that are rapidly evolving under positive selection pressure. We therefore binned protein sequence pairs in the OGD into different categories based on the extent of protein sequence change, as measured by PAM distances,[3] and time since the point of divergence, as measured by the fraction of conserved twofold redundant silent sites in the DNA sequence ($f_2$). The boundaries between the bins were chosen to allow each bin to contain a useful number of pairs. Resulting binned pairs are illustrated in Table 1.

The PAM metric is a distance between two protein sequences. As a distance, it obeys additive rules and the triangle inequality. It is an excellent metric for judging how similar two protein sequences are. However, protein sequences are generally under strong selective pressure. As a consequence, they evolve rapidly or slowly in response to selective pressure and adaptive change. This episodic (start–stop) sequence evolution means that the PAM metric is an imperfect measure of time.[41]

For this reason, metrics based on silent nucleotide substitution in the gene are generally used to estimate the time of divergence. The $f_2$ metric has recently been shown to be useful in both vertebrates and fungi for this purpose.[14] This is presumably because the rate of accumulation of silent substitutions is largely free of selective pressure, unless codon bias is strong.[16] The $f_2$ metric can be converted to a distance if codon bias is known. For the purpose of this work, the $f_2$ metric was used directly.

The two metrics, PAM and $f_2$, are roughly inversely correlated. Protein pairs that are separated by large evolutionary distance as indicated by high PAM distances (indicating many amino acid replacements in the history separating them) are also frequently found to have low $f_2$ values (indicating many silent nucleotide substitutions). These were designated to be "typical". Also "typical" are proteins that have only recently diverged, as indicated by a high $f_2$ value, which are associated with short PAM distances.

This correlation between PAM and $f_2$ is expected to be violated in proteins that are subjected to exceptionally high levels of purifying selection, which cause their protein sequences to tolerate very few amino acid replacements over long periods of time. These "constrained" proteins are separated by low $f_2$ values, indicating more ancient divergence, and shorter PAM distances as compared to typical protein pairs.

The PAM and $f_2$ correlation is also not fit by proteins that are evolving under positive selection pressure. Here, non-synonymous substitutions in the gene accumulate with greater chance than synonymous substitutions, creating pairs of proteins with unusually large PAM distances and high $f_2$ values. These pairs are termed "adaptive".

† http://www.scinq.org

## Acknowledgements

We are indebted to the Department of Defense (DAAD13-02-C-0080), to the Agouron Foundation for providing a postdoctoral fellowship to M.C., and to EraGen Biosciences for providing the MasterCatalog. We thank Raphael LaFrance for his assistance in programming issues arising from this research. We thank Dr Michael Bradley for his insights in discussion of results relevant to this paper, and Dr Gina M Cannarozzi for her assistance in statistical analyses.

## References

1. Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.
2. Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
3. Dayhoff, M. O. & National Biomedical Research Foundation (1978). A model for evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, p. 345. National Biomedical Research Foundation, Silver Spring, MD.
4. Gonnet, G. H., Cohen, M. A. & Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science*, **256**, 1443–1445.
5. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
6. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. & Thompson, J. D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucl. Acids Res.* **31**, 3497–3500.

7. Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucl. Acids Res.* **16**, 10881–10890.
8. Notredame, C. & Higgins, D. G. (1996). SAGA: sequence alignment by genetic algorithm. *Nucl. Acids Res.* **24**, 1515–1524.
9. Notredame, C., Holm, L. & Higgins, D. G. (1998). COFFEE: an objective function for multiple sequence alignments. *Bioinformatics*, **14**, 407–422.
10. Notredame, C., Higgins, D. G. & Heringa, J. (2000). T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217.
11. Notredame, C. (2002). Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics*, **3**, 131–144.
12. Benner, S. A., Cohen, M. A. & Gonnet, G. H. (1993). Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol.* **229**, 1065–1082.
13. Benner, S. A. & Gerloff, D. (1991). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. *Advan. Enzyme Regul.* **31**, 121–181.
14. Benner, S. A. (2003). Interpretive proteomics—finding biological meaning in genome and proteome databases. *Advan. Enzyme Regul.* **43**, 271–359.
15. Fukami-Kobayashi, K., Schreiber, D. R. & Benner, S. A. (2002). Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. *J. Mol. Biol.* **319**, 729–743.
16. Benner, S. A., Caraco, M. D., Thomson, J. M. & Gaucher, E. A. (2002). Planetary biology—paleontological, geological, and molecular histories of life. *Science*, **296**, 864–868.
17. Black, S. D. & Mould, D. R. (1991). Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. *Anal. Biochem.* **193**, 77–82.
18. Pascarella, S. & Argos, P. (1992). Analysis of insertions/deletions in protein structures. *J. Mol. Biol.* **224**, 461–471.
19. Brant, D. A. & Flory, P. J. (1965). The configuration of random polypeptide chains. *J. Am. Chem. Soc.* **87**, 2788.
20. Flory, P. J. (1953). *Principles of Polymer Chemistry*, Cornell University Press, Ithaca.
21. Gaucher, E. A., Gu, X., Miyamoto, M. M. & Benner, S. A. (2002). Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem. Sci.* **27**, 315–321.
22. Blaisdell, B. E. (1983). A prevalent persistent global nonrandomness that distinguishes coding and noncoding eucaryotic nuclear DNA sequences. *J. Mol. Evol.* **19**, 122–133.
23. Tautz, D. & Renz, M. (1984). Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucl. Acids Res.* **12**, 4127–4138.
24. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
25. Borstnik, B. & Pumpernik, D. (2002). Tandem repeats in protein coding regions of primate genes. *Genome Res.* **12**, 909–915.
26. Tompa, P. (2003). Intrinsically unstructured proteins evolve by repeat expansion. *BioEssays*, **25**, 847–855.
27. Takaiwa, F. & Sugiura, M. (1982). Nucleotide sequence of the 16S–23S spacer region in an rRNA gene cluster from tobacco chloroplast DNA. *Nucl. Acids Res.* **10**, 2665–2676.
28. Levinson, G., Marsh, J. L., Epplen, J. T. & Gutman, G. A. (1985). Cross-hybridizing snake satellite, *Drosophila*, and mouse DNA sequences may have arisen independently. *Mol. Biol. Evol.* **2**, 494–504.
29. Levinson, G. & Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4**, 203–221.
30. Buroker, N. E., Brown, J. R., Gilbert, T. A., O'Hara, P. J., Beckenbach, A. T., Thomas, W. K. & Smith, M. J. (1990). Length heteroplasmy of sturgeon mitochondrial DNA: an illegitimate elongation model. *Genetics*, **124**, 157–163.
31. Learn, G. H., Jr, Shore, J. S., Furnier, G. R., Zurawski, G. & Clegg, M. T. (1992). Constraints on the evolution of plastid introns: the group II intron in the gene encoding tRNA-Val(UAC). *Mol. Biol. Evol.* **9**, 856–871.
32. Ohno, S. (1970). *Evolution by Gene Duplication*, Springer, Berlin.
33. Smith, G. P. (1976). Evolution of repeated DNA sequences by unequal crossover. *Science*, **191**, 528–535.
34. Anderson, P. & Roth, J. (1981). Spontaneous tandem genetic duplications in *Salmonella typhimurium* arise by unequal recombination between rRNA (rrn) cistrons. *Proc. Natl Acad. Sci. USA*, **78**, 3113–3117.
35. Toth, G., Gaspari, Z. & Jurka, J. (2000). Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* **10**, 967–981.
36. Metzgar, D., Bytof, J. & Wills, C. (2000). Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res.* **10**, 72–80.
37. Field, D. & Wills, C. (1998). Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. *Proc. Natl Acad. Sci. USA*, **95**, 1647–1652.
38. Jenny, T. F., Gerloff, D. L., Cohen, M. A. & Benner, S. A. (1995). Predicted secondary and supersecondary structure for the serine-threonine-specific protein phosphatase family. *Proteins: Struct. Funct. Genet.* **21**, 1–10.
39. Riley, M. & Labedan, B. (1997). Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module. *J. Mol. Biol.* **268**, 857–868.
40. Benner, S. A., Chamberlin, S. G., Liberles, D. A., Govindarajan, S. & Knecht, L. (2000). Functional inferences from reconstructed evolutionary biology involving rectified databases–an evolutionarily grounded approach to functional genomics. *Res. Microbiol.* **151**, 97–106.
41. Ayala, F. J. (1999). Molecular clock mirages. *BioEssays*, **21**, 71–75.


***Edited by F. E. Cohen***

*(Received 10 November 2003; received in revised form 17 May 2004; accepted 24 May 2004)*