Alignment of Phylogenetically Unambiguous Indels in *Shewanella*

JOHN P. McCROW

ABSTRACT

High levels of alignment errors associated with gaps have generally meant their exclusion from phylogenetic analysis. Conserved inserts and deletions (indels) may in some cases be less subject to errors than amino acid substitutions for inferring the history of genomes and identifying recently laterally transferred genes, but alignment error near gaps must be evaluated prior to using indels as phylogenetic characters. A method is presented for evaluating the phylogenetic unambiguity of gaps in multiple sequence alignments by allowing a defined amount of pairwise alignment ambiguity. This work considers the bacterial genus *Shewanella*, which is of particular interest for applications of bioremediation and environmental engineering. Understanding the genetic history of these species is vital for these applications. A set of pairwise dynamic programming alignments is constructed to test positions in multiple alignments for phylogenetic unambiguity, and a whole genome scan is done on protein sequences from 11 sequenced species of the bacterial genus *Shewanella*. The splits defined by phylogenetically unambiguous indels are then used as characters for phylogenetic analysis, and results are compared to whole genome Maximum Likelihood phylogeny. A comparable description of the history of the species is found, as well as a set of lateral gene transfer candidates undetectable by traditional analysis of amino acid substitutions. This analysis is applicable to other taxonomic units at all levels and has the potential to allow cataloging of clear genome-wide phylogenetic markers for taxonomic profiling down to the species level.

Key words: algorithms, computational molecular biology.

1. INTRODUCTION

GENES HAVING UNDERGONE RECENT LATERAL TRANSFER may be of particular importance to the evolution of microbial organisms (Syvanen, 1985; Doolittle, 1999; Jain et al., 2003), but even with complete genomic sequences it is difficult to accurately infer the history of species or of individual genes. Processes of lateral genetic transfer are thought to be much more prevalent among prokaryotes, and they challenge our ability to understand and classify them accurately (Bapteste et al., 2004; Beiko et al., 2005). Levels of genetic exchange may be especially high among shewanellae, as there is evidence of widespread mobile genetic elements and sequences specific to the uptake of foreign DNA found in *Shewanella onidensis*

MR-1 (Romine et al., 2008). Our ability to detect lateral gene transfer (LGT) is dependent on our ability to infer the history of the genome as a whole. Traditional methods of bacterial taxonomy involving primary sequence phylogeny give a reasonable overview of the evolutionary relationships among *Shewanella* (Venkateswaran et al., 1999; Hau and Gralnick, 2007). However, incongruence between phylogenetic trees derived from different genes, as well as varying levels of inconsistency among the sequence characters within each gene, obfuscates the true history and relationships of these species. The use of rare genomic changes (RGCs) as phylogenetic characters has been proposed as potentially yielding higher quality information about the evolutionary history of genomes than collections of nucleotide or amino acid substitutions (Rokas and Holland, 2000; Gribaldo and Philippe, 2002). High-level genomic information—including gene content, gene order, conserved relative inserts or deletions (indels), and nucleotide composition such as GC content—has been used to track evolutionary relationships between organisms within common core gene sets. However, the use of RGCs is limited in mainstream phylogenetic analysis because, while statistical models exist for sequences, no such models are known that describe these high-level genomic events (Rokas and Holland, 2000).

This article focuses on 11 completely sequenced species of *Shewanella*, and a strain of *Vibrio cholerae* is used as an outgroup. Using publicly available protein sequences, a phylogeny is constructed of sequenced *Shewanella* based on conserved indels selected automatically from single family orthologous proteins. An indel, as evidence of one or more historic insertion or deletion event, refers to a single sequence locus with more than one relative indel state. An indel state refers to a gap of a particular size, including no gap, and an indel may in general have two or more states. It is assumed that the rate of independent reversion within conserved indels is intrinsically very low, and certainly lower than single amino acid substitutions. Although indels appear in lower quantity than individual amino acids, they may be much higher quality phylogenetic characters and exhibit less homoplasy—due either to independent convergent events (Meyer et al., 1986; Lloyd and Calder, 1991) or to amelioration (Lawrence and Ochman, 1997). Efforts to explain the historical relationships among the main divisions of bacteria (Gupta, 2001; Gupta and Griffiths, 2002; Murphy et al., 2007) have relied on few, and in some cases single, indels as evidence of particular branches in the phylogeny. Use of these indels requires that independent convergence of indel states be extremely rare. Other studies (Baptiste and Philippe, 2002) suggest that a great deal of homoplasy may exist in indels due to independent reversion, and that more than a few indels are needed to accurately describe any phylogeny. Remarkably, these studies rely on the single multiple alignment of sequences to utilize indels as phylogenetic characters, and none explores the extent to which homoplasy might be due to alignment error and the resulting ambiguity of indel states.

Alignment ambiguity may be negligible in typical phylogenetic analysis based on highly conserved regions of sequence away from indels, but is of utmost importance when aligning indels themselves. Gaps in multiple alignments are rarely unambiguous in their exact location across different sets of alignment parameters, and efforts to strictly avoid alignment ambiguity leave few characters to use for further analysis (Gatesy et al., 1993). Studies that focus on reducing phylogenetic ambiguity rely on single multiple alignments, ignoring potential hidden alignment ambiguity (Gupta, 2001; Gupta and Griffiths, 2002; Murphy et al., 2007). Improvements in sequence alignment to more accurately model gaps, such as those proposed by Simmons and Ochoterena (2000), are crucial but are still subject to alignment ambiguity. Before direct inferences are made from apparent shared gaps in a multiple alignment, the stability of such gaps should be in question with regard to small variations in the alignment model. Rather than suggest another sequence alignment model, it is the purpose of the methods presented here to be somewhat robust to variations in alignment models and parameters, and find indels that are most suitable for phylogenetic inference because they exhibit a consistent phylogenetic signal within a certain neighborhood of alignment variation. To the author's knowledge, there are no previous studies that have developed a set of criteria for balancing alignment and phylogenetic ambiguity that can be used to automatically scan entire bacterial genomes for unambiguous indels.

The present article establishes a practical definition for phylogenetically unambiguous indels and automatically searches over the entire core genome for well-defined synapomorphic characters (Hennig, 1966) from which to construct the likely branching order of sequenced *Shewanella*. By searching a restricted space of suboptimal alignments for phylogenetic ambiguity, it is possible to filter out those indels that are phylogenetically ambiguous without exhaustively analyzing all suboptimal alignments. Phylogenetically unambiguous indels may also suggest LGT among genes where there is not enough phylogenetic signal to support a LGT hypothesis using maximum likelihood methods on primary sequence alone.

2. EVALUATION OF PHYLOGENETIC UNAMBIGUITY

For the evaluation of phylogenetic unambiguity, it is important to find as many indels as possible while avoiding those that are not clear in their implied grouping of species. To require no alignment ambiguity would be perhaps too restrictive (Gatesy et al., 1993) and would be unnecessary in cases where some acceptably low amount of alignment ambiguity exists, but where there is no phylogenetic ambiguity (Fig. 1 left). Conversely, interpreting all gaps in a multiple alignment as being equally clear phylogenetic characters is too permissive, and does not address either alignment or phylogenetic ambiguity. Although phylogenetic inference may in some cases be robust enough to handle errors associated with using all of these sets of characters together, it is necessary to address the quality of individual characters and their implications regarding LGT of individual genes. A working definition of phylogenetic unambiguity is proposed, and a set of criteria is established by which to evaluate indels prior to their use as phylogenetic characters.

2.1. Unambiguity criteria

One way of defining phylogenetic unambiguity is to require a single shared gap in each optimal multiple alignment under a defined range of alignment parameters that implies the same partition of sequences. Problems with this approach are that (1) the boundaries of what should be considered reliable alignment parameters are unclear, (2) the challenging parametric alignment problem this implies, and (3) only one multiple alignment under each parameter choice is considered. Different groupings of species inferred from other optimal, or even slightly suboptimal, multiple alignments are not discovered. A more convenient approach that addresses these three problems is to define phylogenetic unambiguity as the existence of a single partition of sequences over all suboptimal alignments within a neighborhood defined by a single

FIG. 1. Hypothetical examples of multiple alignments both displaying alignment ambiguity, with phylogenetic unambiguity (left) and phylogenetic ambiguity (right). In both cases, alternative alignments are shown vertically, and corresponding trees given with minimal changes marked by alignment positions on branches for taxa a–e. (Adapted from Gatesy et al., 1993, Fig. 1.)

making the description of each indel locus robust to alignment ambiguity. Multiple closely located loci may be tested, but only one resulting split is used, as they are highly dependent and likely describe the same indel and should thus result in only one phylogenetic character.

A two-step process is employed to evaluate an indel locus for phylogenetic unambiguity. The first step attempts to partition the sequences into parts such that all pairs of sequences within a part align well without a relative gap in the neighborhood of the indel locus, and no pairs between parts align well without a relative gap (Fig. 2-II). If a valid partition exists, then the second step ensures that a single gap size and location exists between parts by requiring all pairs of sequences between parts align well with a single forced shared gap (Fig. 2-III). Parameter $\delta = 10$ is used in both steps as the distance on either side of $\vec{t}$ in the first step, and the common gap in the second step, where matches are required and no gaps allowed when comparing pairwise sequences to ensure the indel is well separated from any additional nearby indels. Partitions, denoted P , divide homologous sequences from different taxa into disjoint parts, denoted P_x , where x is one of the multiple parts of the partition. Indices x, y will refer to any parts of partitions, while i, j, k will refer to any homologous sequences.

Sequences are partitioned into parts such that sequences have the same indel state within each part and different indel states between parts. Sequences with the same indel state when aligned should not have any gaps in the region near the indel locus, but sequences with different indel states should always have a gap in this region. The optimal pairwise alignment score of sequences i and j is denoted $S_{i,j}^*$. Score M_{i,j,t_j}^* is the optimal pairwise alignment score of sequences i and j , given that no gaps are allowed in the region near the indel defined by $t_j \pm \delta$. Because M^* is a special case of S^* , and both are similarity scores, then $S^* \geq M^*$ and thus $S^* - M^* \geq 0$ is an appropriate measure of suboptimal distance for the pair of sequences to have the same indel state. All pairwise alignment scores are calculated and a partition P is attempted, where parts P_x for all x must contain sequences with the same indel state according to the criterion 1:

Criterion 1

$$\exists P : \quad \forall x \left\{ \begin{array}{l} S_{i,j}^* - M_{i,j,t_j}^* \leq \alpha, \quad \forall i, j \in P_x \\ S_{i,k}^* - M_{i,k,t_k}^* > \alpha, \quad \forall i \in P_x, \forall k \notin P_x \end{array} \right\} \quad (1)$$

The existence of a valid partition P of all sequences implies unambiguity, and if no such partition is possible, then there exists some amount of ambiguity with regard to indel state among the sequences. In addition to a valid partition of sequences based on pairwise comparisons, a single common gap must exist between parts when all sequences are considered together as in a multiple alignment. Another optimal pairwise alignment score, denoted $G_{i,j,l',r'}^*$, is defined as the optimal pairwise alignment score between sequences i and j , such that there must be a gap in sequence i aligned to positions $[l', r']$ on j , and there must be no gap in the regions $[l' - \delta, l' - 1]$ and $[r' + 1, r' + \delta]$, on j . This ensures the gap $[l', r']$ is uniquely defined, and there is clear separation between any additional nearby indels. Formal definitions of the recursions used to efficiently calculate (S^*, M^*, G^*) are given in equations (6, 7, 8), respectively, in Section 2.2.

The values l and r represent the leftmost and rightmost position of a common gap between two parts of a partition. Without loss of generality, part P_x is defined as the part with the gap relative to P_y , then l and r are defined on a reference sequence j_0 in P_y . Reference sequence j_0 is selected to be the sequence with most similarity to one in P_x such that:

$$(i_0, j_0) = \arg \max_{i,j} S_{i,j}^* : i \in P_x, j \in P_y \quad (2)$$

The gap length d is extracted from the pairwise alignment between (i_0, j_0) and $d = r - l + 1$ is maintained for all choices of (l, r) . Then γ different choices of (l, r) in the neighborhood of the indel around $\vec{t}$ are tried. The index h is used to denote each choice of (l, r) in the range $1 \leq h \leq \gamma$, so that:

$$l_h = \lfloor t_{j_0} - d/2 - \gamma/2 + h \rfloor \quad (3)$$

$$r_h = l_h + d - 1 \quad (4)$$

At each choice (l_h, r_h) , all pairs of sequences between $i \in P_x$ and $j \in P_y$ are aligned such that the same common gap is forced on i relative to j . Because (l_h, r_h) are defined on the reference sequence j_0 , (l_h', r_h') is defined as the positions in j comparable to (l_h, r_h) in j_0 . A traceback in the matrix used to compute $M_{j,j_0,t_{j_0}}^*$

gives an alignment representing such a mapping between sequences j_0 and j in P_y . If all pairs of sequences between P_x and P_y satisfy $S_{i,j}^* - G_{i,j,l_h',r_h'}^* \leq \beta$ for any choice of h , then this indel satisfies criterion 2:

Criterion 2

$$\exists h : 1 \leq h \leq \gamma, \quad S_{i,j}^* - G_{i,j,l_h',r_h'}^* \leq \beta, \quad \forall i,j : i \in P_x, j \in P_y, x \neq y \quad (5)$$

It is the purpose of this method to validate indel loci in multiple alignments for their phylogenetic stability prior to use as phylogenetic characters. If even one single orthologous sequence aligns within α of the best pairwise alignment to both a sequence in one part of a partition and to a sequence in another part, then these parts are not distinct enough to reliably infer an indel event between them, and the indel locus in question should be regarded as unreliable. Because a perfect partition of sequences is required by criterion 1, phylogenetic ambiguity may be detected even in cases where a single common gap is shown in the original multiple alignment and such ambiguity is not obvious. In contrast, this procedure is also capable of finding a phylogenetically unambiguous partition of sequences in cases where overlapping gaps of different sizes and/or locations are visible in the original multiple alignment, suggesting a more parsimonious explanation of a single indel.

If parameters α and β are relatively small, such as on the order of single amino acid matches, then indels that fail the first criterion must result in at least two different groupings of sequences with the same or similar alignment scores. This is evidence that the given indel is unstable, as there is a relatively small difference in scores supporting more than one conclusion about the grouping of these sequences. It should be noted that a valid partition is not guaranteed to exist, as there may be no choice for α such that two parts of a partition both contain at least two sequences that together all satisfy criterion 1. The second criterion ensures that a single gap between two groups is likely in the context of all the sequences involved, not just each pair of sequences, suggesting a multiple alignment of sequences with regard to the indel. This multiple alignment need not be explicitly constructed as it is the partitioning of sequences that is used for phylogenetic inference and further analysis. Even with approaches to efficiently list suboptimal alignments within a distance of optimal (Waterman, 1983), the size of such a list may be prohibitively large, and in the special case of detecting shared gaps this becomes unnecessary. Comparing restricted pairwise sequence alignments in this way allows efficient detection of ambiguous alignments within α of optimal without attempting to search among all possible suboptimal alignments. Singleton parts of a partition are phylogenetically uninformative, so only those parts containing two or more sequences need be considered for criterion 2.

Traditional multiple alignment attempts to find the most likely arrangement of all sequences together with no upper limit on the reduction of pairwise likelihood allowed in favor of the overall likelihood, effectively making $\beta = \infty$. In contrast, a small β ensures all pairs of sequences between two groups support a single placement of the indel. These criteria can support any similarity based alignment model, so the most appropriate model should be used according to the scale of relatedness of the taxa involved. This study compares *Shewanella* species using a BLOSUM62 scoring matrix with a gap open penalty $\theta = -11$, and extension $\varepsilon = -1$, with $\alpha = 2$, and $\beta = 10$. While larger values of β are less conservative and always allow at least as many indels to pass the criteria, larger values of α may result in fewer indels that pass criterion 1 because increasing α represents a more inclusive cutoff for partitioning. In this case, the maximum number of phylogenetically unambiguous indels results from $\alpha = 2$ and represents the most natural cutoff for partitioning indel states by these criteria.

2.2. Dynamic programming

A dynamic programming approach is taken for computing alignment scores, because it provides for simple and efficient computation of a well-defined alignment model for comparison of pairwise gap placements. The optimal pairwise alignment score is calculated using similarity-based dynamic programming with a log-odds match function and linear gap penalties (Needleman and Wunsch, 1970; Smith and Waterman, 1981; Gotoh, 1982). Altered forms of these recursions are developed here to calculate the optimal alignment score when matches are forced in the region of the indel, or a specific gap is forced. The recursions necessary for full dynamic programming pairwise global alignment of sequences for evaluation by the criteria for phylogenetic unambiguity are given below. Pairwise recursions are used to compare best pairwise optimal alignments when unrestricted (6), restricted to no gaps in a defined region (7), or restricted to a specific gap (8). Alignments use a match scoring function $s(a_i, b_j)$, based on a given substitution model

between amino acid position i in sequence a and position j in b , with linear gap, open θ and extend ε , penalties.

Optimal Alignment

$$\begin{aligned} E_{i,j} &= \max\{S_{i,j-1} + \theta, E_{i,j-1} + \varepsilon\} \\ F_{i,j} &= \max\{S_{i-1,j} + \theta, F_{i-1,j} + \varepsilon\} \\ S_{i,j} &= \max\{S_{i-1,j-1} + s(a_i, b_j), E_{i,j}, F_{i,j}\} \end{aligned} \quad (6)$$

Forced Match: no gaps in $[l, r]$

$$\begin{aligned} E'_{i,j} &= \begin{cases} M_{i,j-1} + \theta, & i = r + 1 \\ \max\{M_{i,j-1} + \theta, E_{i,j-1} + \varepsilon\}, & \text{otherwise} \end{cases} \\ F'_{i,j} &= \max\{M_{i-1,j} + \theta, F_{i-1,j} + \varepsilon\} \\ M_{i,j} &= \begin{cases} M_{i-1,j-1} + s(a_i, b_j), & l \leq i \leq r \\ \max\{M_{i-1,j-1} + s(a_i, b_j), E'_{i,j}, F'_{i,j}\}, & \text{otherwise} \end{cases} \end{aligned} \quad (7)$$

Forced Gap: gap $[l, r]$, no gaps $\pm\delta$

$$\begin{aligned} E''_{i,j} &= \begin{cases} G_{i,j-1} + \theta, & i = l \\ G_{i,j-1} + \theta, & i = r + \delta \\ \max\{G_{i,j-1} + \theta, E_{i,j-1} + \varepsilon\}, & \text{otherwise} \end{cases} \\ F''_{i,j} &= \max\{G_{i-1,j} + \theta, F_{i-1,j} + \varepsilon\} \\ G_{i,j} &= \begin{cases} E''_{i,j}, & l \leq i \leq r \\ G_{i-1,j-1} + s(a_i, b_j), & l - \delta < i < l \\ G_{i-1,j-1} + s(a_i, b_j), & r < i < r + \delta \\ \max\{G_{i-1,j-1} + s(a_i, b_j), E''_{i,j}, F''_{i,j}\}, & \text{otherwise} \end{cases} \end{aligned} \quad (8)$$

Any system of alignment procedures that can satisfy the criteria given may be used here, and in certain cases, there are opportunities for computational short-cuts. For instance, if a traceback on $S_{i,j}$ reveals an optimal alignment with no gap in the region specified, there is no need to calculate $M_{i,j}^*$ explicitly, because it is known in this case that $S_{i,j}^* - M_{i,j}^* = 0$. However, when there is a gap in $t_j \pm \delta$ in the unrestricted pairwise alignment, then $M_{i,j}^*$ must be calculated explicitly. This situation applies for $G_{i,j,l,r}^*$ as well. Calculation of matrices M and G can be done with minimal operations by transforming S into M or G , by updating the matrix only in the regions where they are restricted and continuing to update downstream of these regions similar to declumping local alignments (Waterman and Eggert, 1987). Matrix $G_{i,j,l,r}$ may be transformed into $G_{i,j,l+1,r+1}$ in a similar manner.

A further computational improvement may be achieved by computing pairwise fit alignments rather than global alignments. For analysis of *Shewanella* species, initialization of the dynamic programming matrices was performed as a fit alignment with $S_{x,y}(0,0) = 0$, $S_{x,y}(i,0) = \theta + i\varepsilon$, $S_{x,y}(0,j) = \theta + j\varepsilon$, and optimal alignment score $S_{x,y}^* = \max_{n-2\delta \leq i \leq n} \{S_{x,y}(i,m)\}$, where n and m are the lengths of sequences x and y respectively. The fit alignment radius $r=40$ was used prior to alignment to truncate all sequence pairs (x, y) around their seeded indel positions to $[t_x - r - \delta, t_x + r + \delta]$ for sequence x , and $[t_y - r, t_y + r]$ for sequence y . This avoids alignment errors associated with the ends of sequences, and sets all pairwise alignment matrices to a constant size. All potential indels marked by shared gaps in multiple alignments of orthologous proteins returned by CLUSTALW were tested for phylogenetic unambiguity by the following parameters: $\alpha = 2$, $\beta = 10$, $\delta = 10$, $\gamma = 20$. Those indels stable enough to pass as phylogenetically unambiguous under these parameters were used as characters for phylogenetic analysis of *Shewanella* species and for detection of specific cases of lateral gene transfer.

The method presented here is quadratic in space and time in the number of sequences. Dynamic pairwise alignment takes $O(n^2)$ space and time in the length of the longest sequence n . All pairwise sequence comparisons are done for each indel giving $O(k^2n^2)$ time in the number of sequences k to partition sequences for a single indel. Once partitioned, a gap length equal to that between the best scoring pair of sequences is assumed, and gap positions are tried within a region of constant size γ centered around seeded indel positions by criterion 2. This takes $O(\gamma n^2)$. Transformation of matrix S into M or G saves computation, but the number of updates required for transformation is not bounded and so this does not result in a decrease in time complexity. The realistic length of sequences n can be bounded, however, because indels are dealt with one at a time in individual protein sequences. A further bound on n is introduced by performing fit alignments on a region of sequence centered around each indel instead of global alignments making n a small constant. This gives a total time complexity of $O(k^2)$.

3. RESULTS

The methods described here for evaluating the phylogenetic unambiguity of indels were used to build a set of phylogenetic characters among conserved core genes in 11 species of the bacterial genus *Shewanella*. These species are used as a basis for comparison of multiple methods of phylogenetic inference, and the resulting species trees are discussed in the following section. A set of genes that appear to have undergone a different history than the rest of the core genome are also found, and the ability to identify potential LGT candidate genes by phylogenetically unambiguous indels is explored.

3.1. Phylogenetic analysis

A reference branching order of species is required against which genes are cataloged that have a significantly different history than the species in general due possibly to lateral gene transfer. However, as expected, slightly different species topologies result from analysis based on different sources of genomic information. A minimal characterization of the genetic relationship of species involves sequencing of highly conserved core RNAs, and/or protein sequences, typically 16S rRNA, DNA Gyrase subunit B (GyrB), and DNA/DNA hybridization experiments on core genes (Venkateswaran et al., 1999; Murray et al., 2001; Hau and Gralnick, 2007). Where whole genome sequences are available, we may be able to more accurately infer species history.

A comparison of branching order derived from concatenated core genome protein sequences and those from phylogenetically unambiguous indels is given in Figure 3. It is debatable whether using phylogenetic analysis designed for single genes is valid for concatenated core genome sequences (Kubatko and Degnan,

FIG. 3. Branching order of *Shewanella* species by concatenated core protein sequences (**left**) and phylogenetically unambiguous indels (**right**). Bootstrap values are given below each branch as a percentage of 1000 samples, all branches on concatenated core protein sequences are 100%. The number of indels exclusive and (inclusive) to each branch are given above. Branch points A–D all have a similar number of indels supporting each branching of species *S. loihica* PV-4.

2007), as there is an implicit assumption of a single genomic history, yet the extent to which this assumption is untrue and will affect the results is unclear. All branches show 100% bootstrap support in the concatenated core protein tree, showing the highest confidence in this topology, but perhaps hiding the locations of any instability in phylogenetic signal across the genomes. There are 665 (12%) phylogenetically unambiguous indels of 5657 potential loci found in 465 (26%) orthologous proteins of 1806 total.

Branching order derived from indels is largely the same as that from the concatenated core proteins, with the exception of the placement of *S. loihica* PV-4, which has reasonable support on four different lineages (A–D) relative to the rest of the tree. Values are given above each branch, indicating the number of indels that appear to have had single insertion or deletion events during the time represented along each branch. Counts of indels exclusive to each branch are given, as well as the count, in parenthesis, of indels inclusive on each branch when not enough species are involved to determine a single branch. Branch D is supported by the largest number of exclusive indels, while the branch in the concatenated core protein topology corresponding to branch B is most likely. An alignment of one indel supporting branch D is shown in Figure 4. Branch A, which has support from 10 indels, is also the favorite, given the GyrB sequence, for example. For the purposes of comparing potential LGT candidate genes, the consensus history of the genomes of these species is considered to be any of the four trees derived from Figure 3.

3.2. Laterally transferred genes

By comparing the highest likelihood tree for each multiply aligned orthologous group of proteins to that of the consensus trees, a Shimodaira-Hasegawa (SH) *p*-value (Shimodaira and Hasegawa, 1999) can be calculated as a measure of the significance that each single gene history differs from the consensus history based on amino acid sequence. Controlling false discovery rate with $q^* = 0.05$ (Benjamini and Hochberg, 1995) gives a *p*-value cutoff of 10^{-3} . The same cutoff results by estimating the null *p*-value frequency

Region of leucyl-tRNA synthetase (LeuS) protein [MR-1 gene SO_1174]

Species	Accession	Protein Sequence Alignment	Indel State
<i>S. oneidensis</i> MR-1	NP_716799	DLDISAAYTEKGILFNSG---EFDGLDFDGA FNVIANKL	1
<i>V. cholerae</i>	NP_230603	ELDVSEAAYTEKGVLFASG---EFDGLDFQA AFNAIAAKL	1
<i>S. sp.</i> MR-7	ZP_00856403	ELDISAAYTEKGVLFNSG---EFDGLDFDGA FNIAIDKL	1
<i>S. sp.</i> ANA-3	ZP_00852833	ELDISAAYTEKGVLFNSG---EFDGLDFDGA FNIAIDKL	1
<i>S. sp.</i> MR-4	ZP_00883936	ELDISAAYTEKGVLFNSG---EFDGLDFDGA FNIAIDKL	1
<i>S. putrefaciens</i> CN-32	ZP_00815542	DVDISEAAYTEKGILFNSG---EFDGLDFEAG FNIAIANKL	1
<i>S. baltica</i> OS155	ZP_00583842	DVDISEAAYTEKGVLFNSG---EFDGLDFEAG FNIAIANKL	1
<i>S. amazonensis</i> SB2B	ZP_00587919	EVDISEAAYTEKGVLFNSA---EFDGLDFDGA FNIAIDKL	1
<i>S. sp.</i> W3-18-1	ZP_00906021	EVDISEAAYTEKGILFNSG---EFDGLDFDGA FNIAIANKL	1
<i>S. frigidimarina</i> NCIMB 400	ZP_00638315	EVDVSKEAYTEKGVVFNSGDAFPELDGLDFQA AFEAIDARL	2
<i>S. denitrificans</i> OS217	YP_561810	EVDISKEAYTEKGVLFNSGTAFFPELDGLEFQA AFDIDAIDARL	2
<i>S. loihica</i> PV-4	ZP_00838966	ELDISAAYTEKGILFNSGDSFPELDGLDFQA AFDIDAIDAKL	2

Region of MtrB protein [MR-1 gene SO_1776]

Species	Accession	Protein Sequence Alignment	Indel State
<i>S. oneidensis</i> MR-1	NP_717385	ILTGQMSQDQALVTDNYRYANQLNTDAVDAKV DLLGMNLKV	1
<i>S. putrefaciens</i> CN-32	ZP_00814363	FLAGQMSQDQALVTDNYIYANQLATNAVDAKV DLLGINLKL	1
<i>S. sp.</i> W3-18-1	ZP_00904289	LLAGQMSQDQALVTSYGYY--QIPTEAVDAKV DLLGMNLKL	2
<i>S. loihica</i> PV-4	ZP_00837777	VHYGQMSQDQALVTSYGYY--QTPVDALDAKV DMDKGANLKV	2
<i>S. sp.</i> MR-4	ZP_00883244	ILTGQMSQDQSLVTSYGYY--QLPTDAVDAKV DLLGMNLKV	2
<i>S. amazonensis</i> SB2B	ZP_00586636	LMAGQMSQDEPLVTTGYGY--ALPVEAVDAKV DIIGMTLKA	2
<i>S. sp.</i> MR-7	ZP_00854119	ILTGQMSQDQSLVTSYGYY--QLPTDAVDAKV DLLGMNLKV	2
<i>S. frigidimarina</i> NCIMB 400	ZP_00640776	LLLGQMTQDQQLTSIGYGY--SLPADSIDAKV DITGMTLKA	2
<i>S. sp.</i> ANA-3	ZP_00850302	ILTGQMSQDQSLVTSYGYY--QLPTDAVDAKV DLLGMNLKV	2
<i>S. baltica</i> OS155	ZP_00583714	LLTGQMSQDQALVTSYGYY--QVPTEALDAKV DLIIGNLKV	2

FIG. 4. Multiple alignment of LeuS and MtrB orthologous proteins in the region of phylogenetically unambiguous indels. Both indels are hypothesized to be the result of single insertions 4 a.a. and 2 a.a., respectively, and that the partitions of species suggested by these indels indicate a potential recent lateral gene transfer event involving MtrB, as well as evidence of a monophyletic clade depicted as branch point D in Figure 3 suggested by the indel in LeuS.

assuming uniformly distributed null p -values, and above this cutoff there is no more ability to detect LGT than background noise. There are 64 genes called significant LGT candidates at the $p = 10^{-3}$ level by their amino acid sequence. A more liberal cutoff of $p = 10^{-2}$ would yield 100 LGT candidate genes. A set of 106 genes have phylogenetically unambiguous indels with 2 states that do not belong to a branch from any of the consensus trees. Of these 106 LGT candidates, there are 82 genes that lack the phylogenetic signal to be called significant even at the $p = 10^{-2}$ level based on amino acid sequence alone.

It is difficult to obtain an estimate of false positives among these candidates, as the true positives are unknown, but comparison of indels to adjacent sequence is possible. For each gene containing a phylogenetically unambiguous indel, the number of bootstrap samples is counted, from permutations of amino acid sequences, that result in a tree containing the split defined by the indel. Low bootstrap values are expected to correspond to disagreement between an indel and the adjacent amino acid sequence. Based on 100 bootstrap replicates for each gene, 6–23% of genes containing phylogenetically unambiguous indels have a bootstrap value at or below 0–50, respectively. Using the phylogenetic signal in amino acid sequence as a benchmark for true positives neglects any consideration of errors due to amelioration of sequence and presumes adequate phylogenetic signal exists in each gene. Either source of error would allow more true positives among indels that disagree with their adjacent sequence. An estimate of 23% should only be regarded as a crude upperbound estimate of false positives among LGT candidates found by phylogenetically unambiguous indels.

An alignment of two genes in the region flanking indels is shown as an example in Figure 4. MtrB is particularly interesting, as it is involved in the unique metal reducing capability of many shewanellae (Beliaev and Saffarini, 1998). A possible transmission of this gene between species *S. oneidensis* MR-1 and *S. putrefaciens* CN-32 is identified by the clear allele they share in the form of a two-amino acid insertion. A series of multiple independent insertions of this same sequence into the same position is regarded as highly unlikely, and so it must be inferred that (1) a single insertion event happened, and that these two lineages were at one time a single cell, or (2) this allele was transmitted from one to the other later. Also regarded as highly unlikely is the reversion of this insertion back to the ancestral state, so it must be concluded that our species tree is wrong or there truly was a recent lateral transmission involving this gene. A potential hazard in this logic would involve situations when either multiple independent indel events, or reversion to ancestral indel states, become significantly more likely under certain conditions. While no adequate model exists for how these indels occur, it is hypothesized that phylogenetically unambiguous indels are rare enough in general to support phylogenetic inference by parsimony.

Protein sequences used in all analysis were downloaded April 19, 2006, from NCBI for all available refseq proteins for the following organisms: *Vibrio cholerae* 01 biovar eltor N16961, *Shewanella amazonensis* SB2B, *S. baltica* OS155, *S. denitrificans* OS217, *S. frigidimarina* NCIMB 400, *S. loihica* PV-4, *S. oneidensis* MR-1, *S. putrefaciens* CN-32, *S. sp.* ANA-3, *S. sp.* MR-4, *S. sp.* MR-7, and *S. sp.* W3-18-1. Orthologous groups were created by reciprocal best blast hits for each species among all protein sequences. Fully connected and disjoint groups, 4483 (84%), were used in further analysis representing single family proteins. BLAST 2.2.13 (Altschul et al., 1990) was used with default parameters BLOSUM62, and gap penalties 11, 1. Multiple alignments of all orthologous proteins were done using CLUSTALW 1.83 (Thompson et al., 1994) with default parameters.

Phylogenetic analysis was done using the PHYLIP 3.65 package of tools (Felsenstein, 2005). Phylogenetic inference on amino acid alignments was accomplished both using Maximum Likelihood and Parsimony with PROML and PROTPARS, respectively. Indel states are treated as discrete morphological characters, and analysis was performed using PARS. Perl scripts were created to view phylogenetically unambiguous indel partitions, and to assign them to branches on the consensus tree or to report them as LGT candidates.

4. CONCLUSION

Presented here is a method for validating splits inferred from gaps in multiple alignments prior to using them as phylogenetic characters. While there are significantly fewer informative indels than amino acids available for phylogenetic inference, there are a few important advantages to using indels. Indels are clear markers of single events, not subject to the same levels of reversion or sequence amelioration as amino acid or nucleotide substitutions. An exact dynamic programming method is used to ensure that, within a defined

distance, the space of suboptimal alignments suggests a parsimonious split, and within this same space no ambiguity exists as to what side of the split any taxa are on. Phylogenetically unambiguous indels result in a topology highly consistent with that inferred by maximum likelihood and parsimony methods on concatenated core genome sequences. Clear evidence of instability of particular branches is detected in the form of indels that suggest multiple well supported placements of species *S. loihica* PV-4 relative to other species. This instability is not as easy to detect among concatenated core protein sequences. Indels that are not consistent with consensus topologies, but are phylogenetically unambiguous by the author's definition, are considered probable candidates of recent LGT. These genes may be of particular interest to functional studies, and many would otherwise have been missed by other phylogenetic methods of LGT detection.

Conserved indels may contain valuable information about the history of genes and genomes, yet this information is often discarded, or worse still inferences are made directly from single multiple alignments. As many approaches to alignment of gaps are possible, the detection of phylogenetic ambiguity is a necessary first step before inferring splits from multiple alignments. Conserved indels appear to be clear phylogenetic characters, and as this presents an advantage over primary sequence, it may also lead some inference astray if alignment errors and specifically phylogenetic ambiguity is not taken into account. This analysis may be easily extended to other taxonomic groups, and used for other purposes than inferring species trees or detection of LGT. The potential exists for the use of minimal sets of phylogenetically unambiguous indels as molecular markers for the characterization of unknown or potentially contaminated samples. Indels, as with other rare genomic changes, are best suited in concert with traditional phylogenetic analysis of primary sequence, not purely in contrast to it. This contribution is intended to be a step among future efforts to successfully unify alignment and phylogenetic inference, from multiple sources of data.

ACKNOWLEDGMENTS

I gratefully acknowledge Michael S. Waterman and Kenneth H. Nealson for their helpful discussion, and editing of the manuscript, as well as the following funding support for this research: NIH Center for Excellence in Genomic Science (NIH P50 HG002790-05) and a USC Viterbi Graduate Fellowship award.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

- Altschul, S.F., Gish, W., Miller, W. et al. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Baptiste, E., and Philippe, H. 2002. The potential value of indels as phylogenetic markers: position of trichomonads as a case study. *Mol. Biol. Evol.* 19, 972–977.
- Baptiste, E., Boucher, Y., Leigh, J. et al. 2004. Phylogenetic reconstruction and lateral gene transfer. *Trends Microbiol.* 12, 406–411.
- Beiko, R.G., Harlow, T.J., and Ragan, M.A. 2005. Highways of gene sharing in prokaryotes. *Proc. Natl. Acad. Sci. USA* 102, 14332–14337.
- Beliaev, A.S., and Saffarini, D.A. 1998. *Shewanella putrefaciens* MtrB encodes an outer membrane protein required for Fe(III) and Mn(IV) reduction. *J. Bacteriol.* 180, 6292–6297.
- Benjamini Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* 57, 289–300.
- Doolittle, W.F. 1999. Phylogenetic classification and the universal tree. *Science* 284, 2124–2129.
- Felsenstein, J. 2005. PHYLIP (phylogeny inference package). Department of Genome Sciences, University of Washington, Seattle.
- Gatesy, J., DeSalle, R., and Wheeler, W. 1993. Alignment-ambiguous nucleotide sites and the exclusion of systematic data. *Mol. Phylogenet. Evol.* 2, 152–157.
- Gotoh, O. 1982. An improved algorithm for matching biological sequences. *J. Mol. Biol.* 162, 705–708.
- Gribaldo, S., and Philippe, H. 2002. Ancient phylogenetic relationships. *Theor. Popul. Biol.* 61, 391–408.
- Gupta, R.S. 2011. The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins. *Int. Microbiol.* 4, 187–202.

- Gupta, R.S., and Griffiths, E. 2002. Critical issues in bacterial phylogeny. *Theor. Popul. Biol.* 61, 423–434.
- Hau, H.H., and Gralnick, J.A. 2007. Ecology and biotechnology of the genus *Shewanella*. *Annu. Rev. Microbiol.* 61, 237–258.
- Hennig, W. 1996. *Phylogenetic Systematics*. University of Illinois Press, Chicago.
- Jain, R., Rivera, M.C., Moore, J.E. et al. 2003. Horizontal gene transfer accelerates genome innovation and evolution. *Mol. Biol. Evol.* 20, 1598–1602.
- Kubatko, L.S., and Degnan, J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 56:17–24.
- Lawrence, J.G., and Ochman, H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* 44, 383–397.
- Lloyd, D.G., and Calder, V.L. 1991. Multi-residue gaps, a class of molecular characters with exceptional reliability for phylogenetic analyses. *J. Evol. Biol.* 4, 9–21.
- Meyer, T.E., Cusanovich, M.A., and Kamen, M.D. 1986. Evidence against use of bacterial amino acid sequence data for construction of all-inclusive phylogenetic trees. *Proc. Natl. Acad. Sci. USA* 83, 217–220.
- Murphy, W.J., Pringle, T.H., Crider, T.A. et al. 2007. Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res.* 17, 413–421.
- Murray, A.E., Lies, D., Li, G. et al. 2001. DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. *Proc. Natl. Acad. Sci. USA* 98, 9853–9858.
- Needleman, S.B., and 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.
- Rokas, A., and Holland, P.W. 2000. Rare genomic changes as a tool for phylogenetics. *TREE* 15, 454–459.
- Romine, M.F., Carlson, T.S., Norbeck, A.D. et al. 2008. Identification of mobile elements and pseudogenes in the *Shewanella oneidensis* mr-1 genome. *Appl. Environ. Microbiol.* 74, 3257–3265.
- Shimodaira, H., and Hasegawa, M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16, 1114–1116.
- Simmons, M.P., and Ochoterena, H. 2000. Gaps as characters in sequence-based phylogenetic analyses. *Syst. Biol.* 49, 369–381.
- Smith, T.F., and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197.
- Syvanen, M. 1985. Cross-species gene transfer; implications for a new theory of evolution. *J. Theor. Biol.* 112, 333–343.
- Thompson, J.D., Higgins, D.G. and 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. *Nucleic Acids Res.* 22, 4673–4680.
- Venkateswaran, K., Moser, D.P., Dollhopf, M.E. et al. 1999. Polyphasic taxonomy of the genus *Shewanella* and description of *Shewanella oneidensis* sp. nov. *Int. J. Syst. Bacteriol.* 49, 705–724.
- Waterman, M.S., 1983. Sequence alignments in the neighborhood of the optimum with general application to dynamic programming. *Proc. Natl. Acad. Sci. USA* 80, 3123–3124.
- Waterman, M.S., and Eggert, M. 1987. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J. Mol. Biol.* 197, 723–728.

Address correspondence to:
Dr. John P. McCrow
J. Craig Venter Institute
10355 Science Center Drive
San Diego, CA 92121

E-mail: jmccrow@jcv.org