# Splice-Aware Multiple Sequence Alignment of Protein Isoforms

Alex Nord*
University of Montana
Missoula, Montana
alexander.nord@umontana.edu

Peter Hornbeck
Cell Signaling Technology, Inc.
Danvers, Massachusetts
phornbeck@cellsignal.com

Kaitlin Carey
University of Montana
Missoula, Montana
kaitlin1.carey@umconnect.umt.edu

Travis Wheeler
University of Montana
Missoula, Montana
travis.wheeler@umontana.edu

## ABSTRACT

Multiple sequence alignment (MSA) is a classic problem in computational genomics. In typical use, MSA software is expected to align a collection of homologous genes, such as orthologs from multiple species or duplication-induced paralogs within a species. Recent focus on the importance of alternatively-spliced isoforms in disease and cell biology has highlighted the need to create MSAs that more effectively accommodate isoforms. MSAs are traditionally constructed using scoring criteria that prefer alignments with occasional mismatches over alignments with long gaps. Alternatively spliced protein isoforms effectively contain exon-length insertions or deletions (indels) relative to each other, and demand an alternative approach. Some improvements can be achieved by making indel penalties much smaller, but this is merely a patchwork solution. In this work we present *Mirage*, a novel MSA software package for the alignment of alternatively spliced protein isoforms. *Mirage* aligns isoforms to each other by first mapping each protein sequence to its encoding genomic sequence, and then aligning isoforms to one another based on the relative genomic coordinates of their constitutive codons. *Mirage* is highly effective at mapping proteins back to their encoding exons, and these protein-genome mappings lead to extremely accurate intra-species alignments; splice site information in these alignments is used to improve the accuracy of inter-species alignments of isoforms. *Mirage* alignments have also revealed the ubiquity of dual-coding exons, in which an exon conditionally encodes multiple open reading frames as overlapping spliced segments of frame-shifted genomic sequence.

## CCS CONCEPTS

• **Applied computing → Molecular sequence analysis**; *Recognition of genes and regulatory elements*;

## KEYWORDS

Multiple sequence alignment, protein isoforms, alternative splicing, dual-coding exons

**ACM Reference Format:**
Alex Nord, Peter Hornbeck, Kaitlin Carey, and Travis Wheeler. 2018. Splice-Aware Multiple Sequence Alignment of Protein Isoforms. In *ACM-BCB'18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, August 29-September 1, 2018, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3233547.3233592

## 1 INTRODUCTION

The task of multiple sequence alignment (MSA [5]) is to organize a set of related biological sequences into a matrix wherein each row represents a single input sequence and each column represents a set of related residues, with the gap character ('-') used to signify insertions or deletions (collectively "indels") that are required for placement of residues into columns.

Multiple sequence alignments (MSAs) are traditionally constructed using software tools that proceed through three characteristic phases [3, 12, 16, 38]. First, the set of input sequences is rapidly clustered to produce a bifurcating "guide" tree, wherein each sequence is represented by a leaf node and the structure of internal nodes approximates the evolutionary relationships between the sequences. An iterative method of profile-to-profile alignment then traces the guide tree from its leaves to its root, such that each internal node stores an MSA representing every sequence among the leaves in that node's subtree, eventually placing an MSA that represents the full set of input sequences at the root of the tree. Finally, a series of refinements attempts to improve the root MSA, which is ultimately returned as the output of the software.

The score of an alignment depends on a scoring scheme that rewards alignments of identical or similar residues and penalizes alignments of dissimilar residues. Importantly, these heuristics also penalize gap characters such that the cumulative penalty for a run of consecutive gap characters is a linear function of the length of that gap. For this reason, MSA tools have traditionally exhibited a strong preference for alignments with occasional mismatches over alignments with long gaps [5].

This approach has proven effective at generating accurate and informative alignments in most contexts, but certain biological phenomena can contradict the logic that guides traditional MSA software. One such challenging phenomenon for protein MSAs is alternative splicing, wherein identical genomic sequence will

*Author to whom correspondence should be addressed


*ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA*

ACM ISBN 978-1-4503-5794-4/18/08.
https://doi.org/10.1145/3233547.3233592

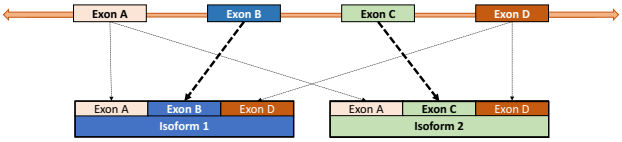


**Figure 1: Diagram of Alternative Splicing with Mutually Exclusive Exons**

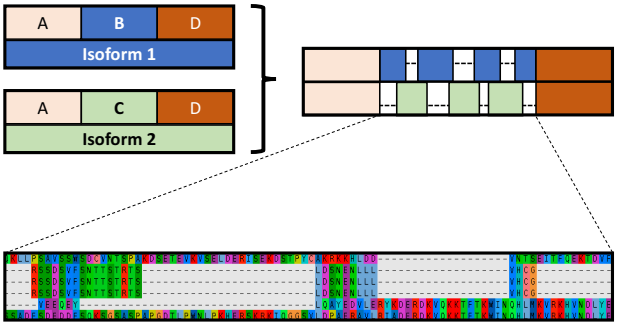


**Figure 2: Diagram Illustrating How Alternative Exon Utilization Confounds Traditional MSA Methods, With Example Alignment of BPAG1 Isoforms by *MUSCLE***

conditionally select protein-coding exons from the available exon pool [34]. The capacity for alternative splicing is biologically advantageous, as it allows cells to dynamically adjust the functional dispositions of their genes in response to environmental changes. An estimated 95% of human genes undergo some form of alternative splicing [28]. Alternatively spliced products of the same gene are commonly referred to as "isoforms," and can rewire protein-protein interactions [37] and signaling pathways [1] in oncogenesis [8] and differentiation [26, 35].

Understanding the relationships between isoforms is useful in a variety of contexts, such as analyzing mass spectrometry results, post-translational modification (PTM) annotation[13] and refinement of genome annotations[22]. Isoform MSAs play an integral role in establishing this understanding. Enzymatically-regulated PTMs play key roles in directing the networks that control biological processes including cancer [2] and other diseases [24]. Isoform MSAs indicate novel target sites for PTMs by transitively annotating residues placed in columns that contain residues with known PTMs (*i.e.*, if a PTM is identified at an amino acid in one isoform, the MSA allows us to predict the possibility of PTM at the same amino acid in another isoform, or in the homologous amino acid in another organism). Genome annotations are largely produced using automated tools that, in part, predict the locations of protein-coding sequences. Manual assessment of predicted coding regions is impractical due to the scale of genomic data, but accuracy can be automatically assessed by constructing MSAs for predicted isoforms and evaluating the occurrences of long indels and the conservation of splice boundaries [7]. The availability of accurate isoform MSAs is indispensable for these and other research applications.

In many proteins, there are exons that are not observed to co-occur in any isoform (so-called Mutually Exclusive Exons [10]). When neighboring exons are mutually exclusive (Figure 1), traditional MSA tools will correctly align the sequences corresponding to shared flanking exons, but struggle to reconcile the non-homologous subsequences corresponding to the exclusive exons. This is because correctly offsetting these exons results in two consecutive exon-length runs of gap characters. Traditional scoring schemes achieve their highest score (lowest penalty) by instead producing alignments with short gaps interleaved with mismatched residues (Figure 2). The resulting alignment of neighboring mutually exclusive exons thus communicates a series of false homologies that can interfere with successive alignment stages during MSA generation and propagate misinformation during downstream analysis.

In this work we present *Mirage*, a novel MSA software package for the alignment of alternatively spliced protein isoforms. The foundational principle of *Mirage* is that isoforms can be aligned to each other by first mapping each protein isoform to its encoding genomic sequence, and then aligning isoforms to one another based on the relative genomic coordinates of their constitutive codons (Figure 3). We refer to this process as "transitive alignment," since residues are aligned to each other through an intermediate. When aligning isoforms from the same species, transitive alignment is expected to produce exact MSAs with columns containing only identical residues, since the isoforms are being mapped back to their actual coding DNA on their shared genome. In the case of inter-species MSAs involving isoforms, evolutionary divergence will mean columns will not be identical, but improved intra-species alignments are expected to improve interspecies accuracy. Further, *Mirage*'s method of intra-species alignment imbues its MSAs with explicit splice site markers that contribute to the scoring scheme used during inter-species alignment, such that *Mirage* alignment acts with an awareness of exonic structure that is not available to standard MSA methods. *Mirage* represents a next-generation approach to bioinformatic algorithm design, insofar as it solves the classical problem of multiple sequence alignment through the use of high-quality reference genomes and curated gene annotation databases that have only recently become widely available.

In addition to the immediate virtues of *Mirage*'s novel approach to multiple isoform alignment, *Mirage* alignments have highlighted the prevalence of dual-coding exons, in which an exon may encode multiple open reading frames as overlapping spliced segments of frameshifted genomic sequence. We describe preliminary analyses of *Mirage*'s dual-coding exon annotations, suggesting that they are much more common than previously recognized and significantly conserved within mammals.

## 2 RESULTS

*Mirage* produces MSAs of all isoforms within one species (an intra-species MSA) by mapping protein sequences to their species' genomes and using the mapped coordinates as an intermediate for aligning sequences that belong to the same gene family. *Mirage* then merges intra-species MSAs to produce inter-species MSAs through the

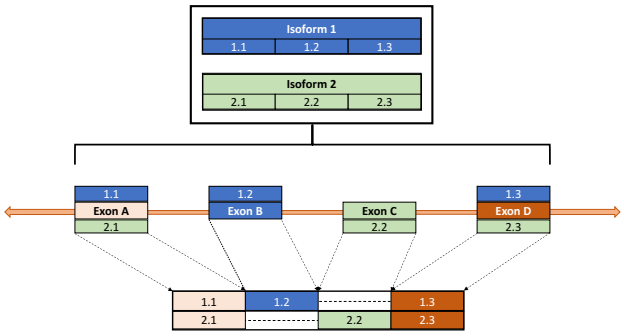


**Figure 3: *Mirage* Aligns Isoforms By Mapping Them to Their Encoding Genome. Each protein sequence is mapped to its encoding genome, then amino acids from these mapped proteins are placed in the same MSA column if they are encoded by the same codon.**

use of a profile-to-profile alignment tool that incorporates splicing information obtained from the transitive intra-species MSAs.

We demonstrate that *Mirage* is highly effective at mapping proteins to their encoding exons, and that these protein-genome mappings result in extremely accurate intra-species alignments. Further, we demonstrate the gain in inter-species MSA quality resulting from *Mirage*'s use of splice site locations in alignment scoring. We show that the runtime of *Mirage* is competitive with the most popular MSA tools, and that the memory overhead of *Mirage* is low enough to run on a standard desktop computer, while its underlying algorithms enable *Mirage* to take advantage of powerful computing resources.

## 2.1 Benchmarking Resources

We tested *Mirage*'s performance along with those of three leading MSA tools: *Clustal-Omega*[32], *MAFFT*[16], and *MUSCLE*[3]. These are generic MSA tools, not designed for the specific task of isoform alignment. To our knowledge, the only existing software designed specifically to align isoforms is from [7], which requires custom data preparation/organization and runs roughly 1000x slower than the tools tested here – due to these constraints, we have not included it in this analysis.

We built alignments of 21,980 gene families made up of 80,779 protein sequences from UniProtKB[31]. We downloaded genomes for the 10 most common species in our protein database from the UCSC Genome Browser[20] (versions in Table 1). GTF annotation files, used to accelerate protein mapping, were acquired from Ensembl[39] (public release 87) for the human, mouse, and rat genomes.

## 2.2 *Mirage*'s Protein-to-Genome Mapping is Effective and Efficient

The *Mirage* pipeline is designed to convert a large database of protein sequences into a set of MSAs, where each MSA is specific to a gene family represented in the input database. Because the proteins in the database may belong to a variety of species, *Mirage* splits the database into a collection of species-specific databases. *Mirage* iterates through these, mapping each protein to that species' genome. *Mirage* is able to map 98% of human and mouse sequences to their genomes, as well as most of the sequences belonging to the other species for which we downloaded genomes (Table 2). On close inspection of several unmatched human isoforms, we found that most contain short internal peptides that have no apparent encoding genomic sequence – it is unclear if this is due to error in isoform or genomic sequence. Because the work invested in curating the human and mouse genomes is unmatched in other species, we suspect that minor errors in other species reduce *Mirage*'s ability to map proteins back to those genomes, explaining the <95% mapping success for other organisms.

| Species | Genome Version |
|---|---|
| Human | GRCh38/hg38 |
| Mouse | GRCm38/mm10 |
| Rat | RGSC 6.0/rn6 |
| Chicken | Gallus_gallus-5.0/galGal5 |
| Cow | UMD_3.1.1/bosTau8 |
| Dog | Broad CanFam3.1/canFam3 |
| Horse | Broad/equCab2 |
| Pig | SCSC Sscrofa11.1/susScr11 |
| Rabbit | Broad/oryCun2 |
| Sheep | ISGC Oar_v3.1/oviAri3 |

**Table 1: Genome Versions Used for Testing. GTF annotation files (from Ensembl) were also downloaded for the three species above the horizontal line.**

| Species | Number of Sequences | Sequences Mapped (%) | Time to Map (hours) |
|---|---|---|---|
| Human | 42,459 | 98.0 | 0.68 |
| Mouse | 27,368 | 98.0 | 1.15 |
| Rat | 10,260 | 94.5 | 0.57 |
| Chicken | 93 | 90.3 | 0.03 |
| Cow | 274 | 77.7 | 0.08 |
| Dog | 43 | 74.4 | 0.06 |
| Horse | 4 | 75.0 | 0.04 |
| Pig | 107 | 66.4 | 0.07 |
| Rabbit | 85 | 60.0 | 0.13 |
| Sheep | 15 | 53.3 | 0.04 |

**Table 2: *Mirage* Protein-to-Genome Mapping Success Rates. All were matched to the corresponding genome from Table 1; mappings for the first three species were aided by use of GTF annotation files.**

## 2.3 *Mirage* Improves Alignment Accuracy

In this section, we discuss the accuracy of alignments produced by *Mirage* and other tools. Numerous benchmarks for alignment quality exist (though see [4, 14] for concerns), but none are designed to

| Species | MSA Column Identity for Mapped Sequences (%) | MSA Column Identity for All Sequences (%) |
|---|---|---|
| Human | 99.97 | 99.91 |
| Mouse | 99.98 | 99.95 |
| Rat | 99.99 | 99.98 |
| Chicken | 99.99 | 99.99 |
| Cow | 99.99 | 99.99 |
| Dog | 100 | 100 |
| Horse | 100 | 100 |
| Pig | 100 | 99.97 |
| Rabbit | 100 | 100 |
| Sheep | 100 | 100 |

**Table 3: Percents Column Identity of *Mirage* Intra-Species MSAs. Barring post-transcriptional modification of the mRNA, we expect 100% of columns to be identical.**

assess efficacy of aligning alternative splicing products. We have opted not to develop such a benchmark out of concerns that development of such a benchmark would create circular dependencies. Instead, we use percent identity as a proxy for accuracy, since all amino acids produced by a codon are expected to be aligned (and identical) in the same column. Additional measures of common error symptoms are provided.

*2.3.1 Column Identity.* After intra-species MSAs have been constructed for the mapped proteins, unmapped proteins are iteratively added (aligned) to the intra-species MSAs. The resulting intra-species MSAs exhibit nearly perfect column identity (Table 3), where column identity is calculated as the fraction of MSA columns in which all non-gap characters are identical. Column identity should be 100% in isoform MSAs because any aligned characters are presumably encoded by exactly the same genomic sequence, although phenomena such as A-to-I editing and sequencing errors can produce nonidentity[25].

Following intra-species MSA construction, *Mirage* iteratively merges the intra-species MSAs to produce the output set of inter-species MSAs. Intra-species MSAs are merged using a simple profile-to-profile alignment [17] algorithm modified to include a penalty for mis-aligned splice sites, thus preserving the exonic structures identified for the intra-species alignments. As shown in Table 4 (and exemplified in Figure 4), *Mirage* MSAs consistently display improved column identity when compared to MSAs produced by the most popular MSA tools. Importantly, the loss in percent identity is greater in other tools than in *Mirage*, presumably because errors within one intra-species alignment are compounded by errors in another intra-species alignment.

*2.3.2 Exon Span.* Ideally, alignments of protein isoforms within one species should display long runs of ungapped amino acids or long runs of gap characters where an exon has been excluded. Assuming that the protein-to-genome mappings produced by *Mirage* are correct, then *Mirage*'s intra-species MSAs can be used to identify the first and last amino acids of each exon in a protein sequence. We refer to the distance between an exon's first column and last column in an MSA as that exon's "span." Ideally, every exon's span

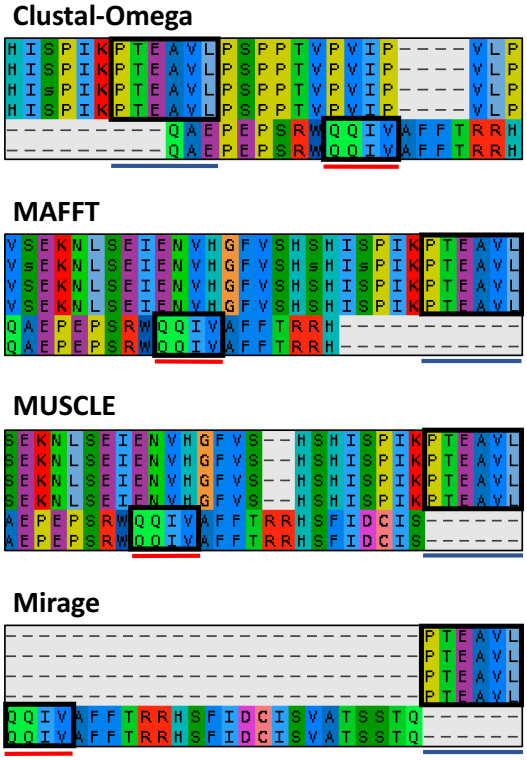


**Figure 4: Comparison of Human DLG1 Alignments Produced by Leading MSA Tools and *Mirage*, with Anchor Motifs Highlighted. Red-underlined motif (QQIV) corresponds to amino acids 54-57 of exon 5, and blue-underlined motif (PTEAVL) corresponds to amino acids 1-6 of exon 6. In an isoform MSA, each instance of each motif should be aligned only to other instances of the same motif. *Mirage* produces this desired result, while other tools force mismatches to be aligned.**

| Alignment Method | Human (%id) | Mouse (%id) | Rat (%id) | Interspecies (%id) | Runtime (hours) |
|---|---|---|---|---|---|
| *Mirage* | **99.91** | **99.95** | **99.98** | **87.12** | 3.00 |
| *Clustal-O* | 96.99 | 97.70 | 98.28 | 82.94 | 9.38 |
| *MAFFT* | 97.66 | 98.28 | 96.68 | 83.82 | 3.26 |
| *MUSCLE* | 97.13 | 97.81 | 94.93 | 83.32 | **2.32** |

**Table 4: Performance of *Mirage* and Popular MSA Tools. Each tool was run with default parameters. In the first three data columns, "%id" indicates the percentage of all columns that have no letter variability. Runtime is the wallclock time when alignments are performed using 32 threads.**

should be equal to the number of amino acids in that exon, but traditional methods increase the spans of exons by inserting gaps in cases of adjacent mutually exclusive exons (MXEs, as illustrated in Figure 1). We examined the extent to which traditional MSA

| Alignment Method | | Human (%) | Mouse (%) | Rat (%) |
|---|---|---|---|---|
| *Clustal-O* | All Exons | 3.1 | 2.9 | 3.9 |
| | Only MXEs | 53.4 | 57.0 | 64.9 |
| *MAFFT* | All Exons | 24.8 | 26.0 | 13.0 |
| | Only MXEs | 198.0 | 227.1 | 108.4 |
| *MUSCLE* | All Exons | 15.4 | 16.6 | 8.3 |
| | Only MXEs | 190.9 | 215.2 | 106.6 |

**Table 5: Increases in Exon Span for Other Tools. Using *Mirage*-learned intron boundaries as a basis for estimating exon length, we show the increase for other alignment tools in the number of columns spanned by an exon. Results are shown for all exons (most of which are not impacted by alignment errors), and also for the focused set of exons most likely to experience alignment error: adjacent mutually exclusive exons (MXEs).**

tools extended the spans of exons by collecting the average percent span increase for all exons and for exons identified as mutually exclusive (Table 5). Acknowledging that these values presuppose that *Mirage*'s protein-to-genome mappings are correct, they demonstrate that traditional MSA methods unwittingly break apart exons. Focusing on mutually exclusive exons, we see that these exons are routinely forced to align to other sequence such that their span in MSA columns is more than 50% greater than the exon sequence length.

*2.3.3 Splice Site Errors.* Information about intron locations from *Mirage* intra-species alignments also allows us to examine how frequently other tools bleed exons into one another. We do this by tracking what we call "splicing pinch points," where splice sites are flanked on both sides by non-gap characters in the same sequence (Figure 5A). Pinch points are informative because they represent the first pieces of coding sequence on either side of an intron, such that errors in the alignments of pinch point amino acids can be interpreted as intron misplacements. If a gap is placed between amino acids that constitute a pinch point, the alignment tool has effectively aligned part of another sequence to the intron implied by the pinch point (depicted in Figure 5B), which we refer to as a "split pinch point" error. Another pinch point error involves unaligning a column of pinch points, in which case the alignment tool is duplicating the intron implied by the pinch point (depicted in Figure 5C), which we refer to as an "unaligned pinch point" error. Table 6 records the frequency with which MSAs show pinch point errors. Similarly to our evaluation of exon span increases, this presupposes that *Mirage* has correctly identified true protein-to-genome mappings.

## 2.4 Runtime

*Mirage*'s runtime is competitive with traditional methods, as shown in Table 4 by the total wallclock runtimes required for each of the alignment programs to produce MSAs for all 21,980 UniProtKB families, consisting of 80,799 sequences. Results are reported using

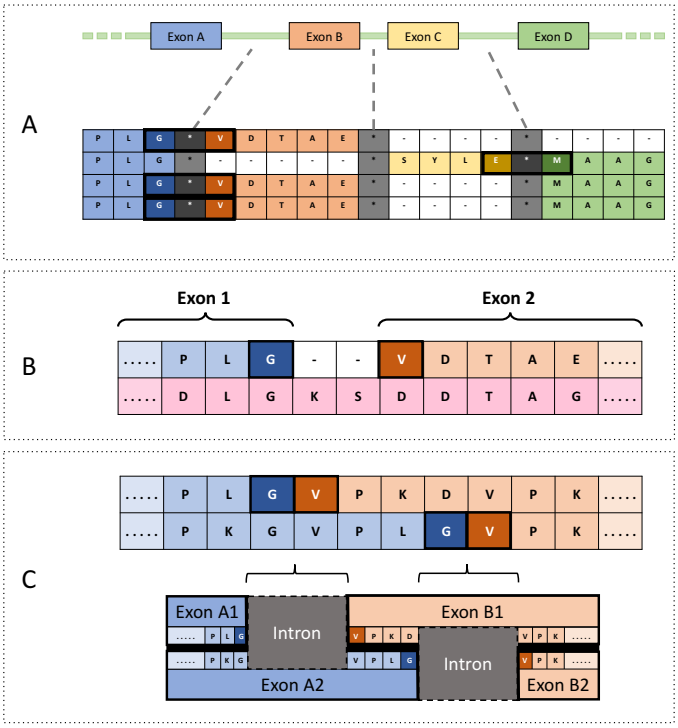


**Figure 5: Pinch Points. (A) A pinch point in sequences 1, 3, and 4 uses the *G* in the 3rd column and the *V* in the 5th column, separated by the intron marker in the 4th column. Breaking apart pinch points like this suggests an isoform alignment error. (B) The top sequence contains a pinch point made up of *G* and *V*, surrounding an intron. The bottom sequence has aligned in a way that essentially aligns two letters (*K* and *S*) to the intron between *G* and *V* – we say this "splits" the pinch point. (C) The placement of conceptual introns implied by the pinch points illustrates how unaligned pinch points confuse the exonic structures of the aligned sequences. All sequences in this figure were developed for purposed of demonstration.**

32 compute cores, though similar runtime ratios are seen for smaller numbers of cores.

Table 7 provides an insight into how time is spent during *Mirage*'s protein-to-genome mapping phase, the most computationally expensive piece of its pipeline. Specifically, Table 7 itemizes the expenses attached to each of the protein-to-genome mapping methods that *Mirage* employs (described in Sections 3.3). While *FastMap* and *SPALN*[15] have similar average execution times per program call, the average associated time per *SPALN* program call is consistently at least one order of magnitude greater than the associated time per *FastMap* program call. This reflects the substantial difference in parsing and post-processing work that *Mirage* requires to extract protein-to-genome mappings from each of the programs. *SPALN* also exhibits a variety of software bugs (detailed in Section 3.6) that *Mirage* is able to detect and repair at a cost to runtime.

| Alignment Method | Species | MSAs with Pinch Point Errors (%) | MSAs with Split Pinch Points (%) | MSAs with Unaligned Pinch Points (%) |
|---|---|---|---|---|
| *Clustal-O* | Human | 5.0 | 4.2 | 1.0 |
| | Mouse | 5.3 | 4.6 | 0.7 |
| | Rat | 11.7 | 9.6 | 3.4 |
| *MAFFT* | Human | 10.0 | 9.1 | 1.1 |
| | Mouse | 9.7 | 9.0 | 0.7 |
| | Rat | 9.7 | 7.6 | 2.7 |
| *MUSCLE* | Human | 10.4 | 9.0 | 2.0 |
| | Mouse | 10.8 | 9.5 | 1.9 |
| | Rat | 11.8 | 10.4 | 4.1 |

**Table 6: MSAs with Exon Bleeding Detected by "Pinch Point" Errors. This shows the percent of all human, mouse, and rat intra-species alignments that contain at least one pinch point error. The first column is a union of the latter two columns.**

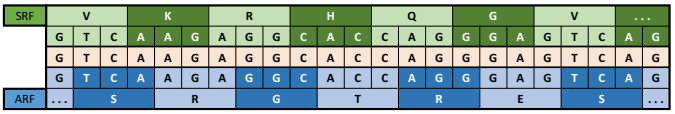


**Figure 6: Toy example Illustrating a Dual-Coding Exon Found in Human ADCY4 Isoforms. The "standard" reading frame peptide is labeled SRF, whereas the "alternative" reading frame peptide is labeled ARF.**

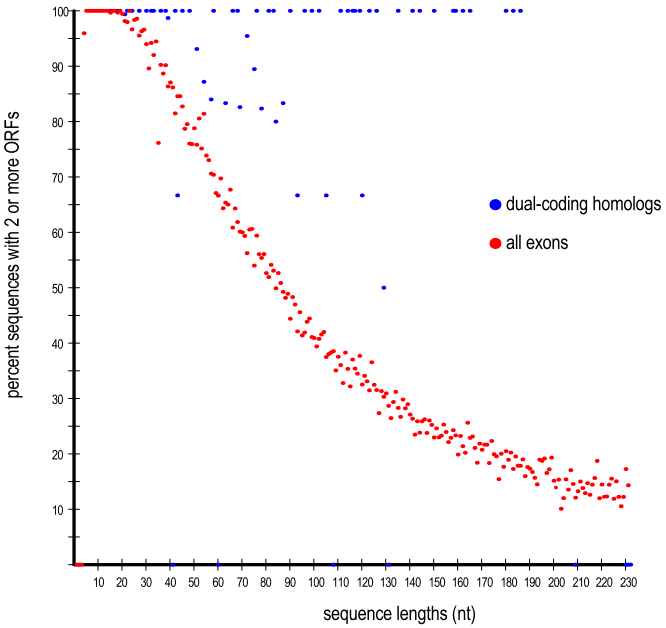


**Figure 7: Conservation of Multiple Open Reading Frames. In general, exons in mouse (red dots) show a decreasing probability of containing multiple open reading frames as length increases. However, nearly all mouse exons that are orthologous to human dual-coding exons have at least two open reading frames (regardless of exon length).**

## 2.5 Dual Coding Exons

During development of *Mirage*, we noticed many instances in which two completely different peptides were mapping to the same region of the genome. Inspection showed that these were instances of a phenomenon know as "dual-coding exons"[9, 36], in which one stretch of DNA can encode more than one protein sequence by using codons that are offset in frame, as illustrated in Figure 6. Dual-coding exons are somewhat common in viruses, but thought to be rare in eukaryotes, with estimates on the order of 100 dual-coding exons in human[21, 23].

*Mirage* identifies dual-coding exons by detecting cases in which two same-species peptides map to an overlapping-but-frameshifted region of the genome. We have identified apparent dual-coding exons in 2,799 (13%) of the 21,980 gene families represented in the protein database we used for testing. The majority of putative dual-coding exons use two overlapping reading frames, but *Mirage* identified 68 that use all three reading frames in various isoforms.

To confirm that dual-coding exons are not simply the result of noisy splicing, we used UCSC whole genome alignments to inspect mouse homologs to human dual-coding exons. We found that >97% of all mouse exon homologs of human dual-coding exons encode at least two open reading frames (Figure 7, blue dots). This stands in stark contrast to the low frequency with which typical mouse exons encoded multiple open reading frames (Figure 7, red dots).

Most dual-coding regions are short (only 18% are longer than 20 amino acids) and consist of a single exon (82% are 1-exon long, 14% involve two consecutive exons). Even taking this length distribution into account, the evolutionary conservation of genomic sequence encoding multiple overlapping open reading frames over approximately 90 million years[6] is striking, and strongly suggests that many of the dual-coding exons identified here may play an important role in biological processes.

# 3 METHODS

## 3.1 Algorithm Overview

*Mirage* constructs an MSA for a family of sequences in four phases. In the first phase, protein sequences are mapped to their respective genome. In the second phase, these mappings are used to produce intra-species alignments, aligning protein residues that map to the same codon on the same organismal genome; unmapped sequences are added to their respective intra-species alignments using sequence-to-profile global alignment. In the third phase, intra-species alignments are merged into a final multi-species MSA using a profile-to-profile alignment algorithm that penalizes misaligned splice sites. In the final phase, unmapped singleton sequences are merged into the MSA.

| Mapping Method | Associated Time (seconds) | Associated Time per Call (seconds) | Execution Time (seconds) | Execution Time per Call (seconds) | Number of Program Calls | Percent of Mapping Successes |
|---|---|---|---|---|---|---|
| *FastMap* | 2,801.58 | 0.036 | 501.21 | 0.0065 | 76,881 | 42.8% |
| *SPALN* | 35,921.85 | 0.810 | 296.27 | 0.0067 | 44,341 | 51.5% |
| *BLAT+SPALN* | 6,961.06 | 24.425 | 1,493.41 | 5.2400 | 285 | 5.7% |

**Table 7: Runtime Comparison of *Mirage*'s Three Mapping Methods, Summed over Human, Mouse, and Rat. The unix command 'time' was used to assess the runtime consumed by each of the three phases of the sequence mapping stage of *Mirage*. Both the "associated" time (a measure that includes the time required for input preparation, program execution, and output parsing) and the "execution" time (the wallclock time elapsed during program execution) are displayed in total for a standard run of *Mirage*, and with their per-program-call averages.**

## 3.2 Input and Preprocessing

*Mirage* was designed to process large-scale sequence files – for example, producing MSAs for all sequence families extracted from UniProtKB[29]. It accepts as input (*i*) a FASTA-formatted protein sequence database and (*ii*) a reference file. The sequence database may contain sequences from many species and many sequence families; the name field indicates the sequence's species and gene family in a prescribed format. The reference file consists of tuples associating species with the locations of their FASTA-formatted genome files and corresponding GTF-formatted gene index files. Genome files are optional, but required for the mapping phase; gene index files are optional, but aid in mapping speed. Upon initialization, the protein database is immediately divided into a collection of species-specific sequence files – one per species listed in the reference file, plus an extra file for sequences belonging to unlisted species. Within each file, sequences are clustered by gene family. This preprocessing results in a set of species, each represented by a unique sequence file, genome, and gene index file.

## 3.3 Mapping Protein Sequences to their Encoding Genome

Following the preprocessing phase, *Mirage* iterates through the list of species, generating spliced genomic mappings for each of the protein sequences. If a species has an associated gene index file, *Mirage* attempts to map the protein to exon regions listed in the index file in two steps: (*i*) mapping peptides to exons using our new tool *FastMap* (Section 3.3.1), then (*ii*) stitching these mapped peptides together to form full-protein mappings using the method described in Section 3.3.2. If *FastMap*+stitching does not identify a full-protein mapping, we depend on the spliced alignment tool *SPALN*[15]. Mapping a protein sequence to a full genome with *SPALN* is prohibitively slow, so *Mirage* uses two strategies to limit the range of genomic sequence searched by *SPALN*. If a gene index file is provided, *SPALN* searches for a spliced mapping of the protein to the region of the genome suggested by that index file (Section 3.3.3). If this fails, or if the gene index file is not available, *Mirage* uses *BLAT*'s translated search option [18] to locate coding regions; the range of these regions is used to constrain the range of *SPALN* alignment (Section 3.3.4).

*3.3.1* FastMap *is A Rapid Method for Identifying Candidate Exons. FastMap* performs highly pruned ungapped alignment of the protein sequence to each exon defined by the gene index file. It seeks a mapping in which a subsequence of the protein aligns with no more than one mismatch to the full length of an exon. This global/local alignment is used because a protein may be encoded by multiple exons.

The input to *FastMap* is a single length-$n$ protein sequence $P$ from species $Q$ and a set $\mathbf{T}$ of $x$ DNA sequences $T_1, T_2, ..., T_x$ that are the complete set of exons listed in $Q$'s gene index file for protein $P$. For notation, the $j^{\text{th}}$ residue of a sequence $T_i$ is represented as $T_i[j]$. *FastMap* iterates through the DNA sequences in $\mathbf{T}$, translating them into the three forward reading frames. For DNA sequence $T_i$, the three forward reading frames are protein sequences $T_i^1$, $T_i^2$, and $T_i^3$. For $1 \leq j \leq 3$ we search for a valid mapping of $T_i^j$ to $P$ as follows:

Let $m$ be the length of $T_i^j$. Initialize an empty queue $A$ of tuples. Each tuple represents a candidate ungapped alignment, and consists of three integers: $D_{start}$, $D_{length}$, and $D_{mismatches}$. For $0 \leq k \leq n-2$, a tuple is placed into the queue with values $D_{start} = k$, $D_{length} = 1$, and $D_{mismatches} = 0$ if $P[k] = T_i^j[0]$ and 1 otherwise, such that each tuple in $A$ represents the beginning of an ungapped alignment of $T_i^j$ to a portion of $P$.

While $A$ is nonempty, let $D = A.dequeue()$. Increment $D_{mismatches}$ if $P[D_{start} + D_{length}] \neq T_i^j[D_{length}]$, then increment $D_{length}$. $D$ is analogous to a cell in an ungapped alignment dynamic programming matrix, insofar as it represents an ungapped alignment of the sequences $P[D_{start}..D_{start}+D_{length}-1]$ and $T_i^j[0..D_{length}-1]$. If $D_{mismatches} < 2$ and $T_i^j[D_{length}-1]$ is not a stop codon, enqueue $D$ into $A$, unless $D_{length} = m$ or $D_{start} + D_{length} = n$.

If $D_{mismatches} < 2$ and either $D_{length} = m$ or $D_{start}+D_{length} = n$, then *FastMap* identified an alignment that is (*i*) global with respect to the exon and local with respect to the protein or (*ii*) local with respect to both but reaches the end of the protein sequence (i.e. may end at the 3' UTR), and has at most one amino acid mismatch, triggering a "hit" to be reported. The final output of a complete *FastMap* run is a list of all ungapped alignments between segments of the input protein sequence and reading frames extracted from the exonic DNA listed in the gene index file.

While the theoretical worst-case runtime of *FastMap* is equal to computing a full dynamic programming matrix to search each reading frame against the full protein sequence, requiring $O(nmx)$ time (assuming $x$ exons each of length $m$), *FastMap*'s expected (and observed) runtime is $\theta(nx)$. The number of tuples added to the initialized queue is $nx$, while the number of computed tuples in a

successful ungapped peptide-exon mapping is $\theta(mx)$ (note: $m < n$), and the number of tuples mapping a peptide to a uniformly random sequence of length $r$ with at most 1 mismatch is $20^{1-r}$, so that the number of tuples along non-matching diagonals is typically dominated by $n$. While the distribution of amino acids in actual protein sequences is not perfectly random, this simplified model roughly characterizes *FastMap*'s observed dropout rate. By maintaining a list of only viable candidate mappings, *FastMap* space usage is typically limited to $O(n)$.

*3.3.2 Candidate Exon Stitching.* If gene index files contained a perfect annotation of constitutive exons for each protein family, mapping would be straightforward: each isoform would map to a subset of a sequential series of exons in substring increments, and the precise alternative exon use would naturally fall out of this mapping. The gene index files available from Ensembl, however, display a dizzying array of annotated putative exons. It is common, for example, to find several exons in the index file that map to essentially the same genomic region, but with very slight offsets from each other (*e.g.*, one starting a nucleotide or two before another, and a third sharing the same start position, but slightly offset end positions). One extreme example is the human nebulin gene (NEB), which is described in the UCSC genome browser database[20] as having 183 known exons, but has over 1,300 exon annotations in the human GTF index. Because of these cases where amino acids can be mapped to several partially-overlapping exons, *Mirage* is required to select a sequence of mappings from the full suite of matched peptides that can account for the full protein sequence. This selected sequence of peptide matches is a splice-aware mapping of the protein to its genome.

Once *FastMap* has identified a set of exons that each partially map to a protein sequence, those hits are sorted according to their genomic starting positions and stored in a list $L$. A graph is created such that each *FastMap* hit is a vertex, and an edge is created between vertices $(u,v)$ if the associated hits are consistent (*i.e.*, the first mapped residue of $v$ begins one amino acid after the last mapped residue of $u$, and $v$ maps to the same genomic strand as $u$, starting in a genomic position downstream of $v$). A depth-first traversal over this graph identifies paths that correspond to concatenations of peptides that cover the entire protein. If multiple such concatenations are found, one with the fewest mismatches is selected (ties are broken arbitrarily). In this way, *Mirage* identifies protein-to-genome mappings using the exon mappings produced by *FastMap*.

*3.3.3 Spliced Mapping with* SPALN. If *FastMap*+stitching fails to identify a set of exon hits that covers the full length of a protein sequence, *Mirage* employs the spliced alignment tool *SPALN*[15] as a fallback mapping method. *SPALN* identifies spliced alignments of protein sequences to DNA. In our tests, it was faster and more accurate than a competitor tool Exonerate[33], with fewer output errors to confound *Mirage*'s parsing efforts.

Though *SPALN* is relatively fast, it is prohibitively slow when used to search a protein sequence against an entire mammalian genome. To reduce runtime, *Mirage* captures the gene location range suggested in the index file and extracts a window of genomic sequence around that range, extending 100Kb in each direction to allow for inclusion of near-boundary exons. The genomic and protein sequences are provided to *SPALN*, which searches for a spliced alignment that *Mirage* can parse into an amino acid-resolution protein-to-genome mapping (for discussion of challenges related to *SPALN* parsing, see Section 3.6). If *SPALN* cannot locate an alignment with ≥95% character identity, a second run is attempted using a larger window of genomic sequence for the DNA input (1 Mb around the indicated coding region), in case a large intron is isolating one or more outlier exons. If *SPALN* fails to identify a 95%-identity alignment after its second run, the protein sequence is designated "unmapped" and is temporarily set aside.

*3.3.4 Spliced Mapping with* BLAT+SPALN. Protein sequences that could not be mapped to the genome in earlier stages are written to a separate sequence file, which is searched against the genome using *BLAT*[18]. Within *Mirage*, *BLAT*'s role is to quickly identify a mapping seed – a portion of the query protein sequence that aligns well to a genomic region. Once this seed is found, *Mirage* employs a series of *SPALN* searches targeting increasingly large windows around the putative coding region. The combination of *BLAT+SPALN* is an effective way to map sequences that either failed to map to the regions of the genome suggested by the gene index file or that belong to a species that was not provided with a gene index file.

Any protein sequences that still fail to map to the genome after the *BLAT+SPALN* phase are designated as "misses," and a file listing all missed sequences is produced. These sequences will eventually be integrated into their gene families' MSAs as described in Section 3.5.

## 3.4 Intra-Species Alignments

Following the protein-to-genome mapping phase, *Mirage* constructs transitive intra-species MSAs for sequences that successfully mapped to the genome. Mirage places amino acids in the same MSA column if they map to the same codon in genomic sequence. For a single gene family, shared codon mapping is tracked by storing mapping information for all intra-species sequences in a hash table. Each amino acid is stored in the table, using the amino acid's mapped location (specifically, the codon's middle nucleotide) as key, and capturing sequence identifier and amino acid index as the value. Once all sequences have been captured in the hash table, production of the alignment is guided by sorting the hash keys (genomic positions) and storing the resulting sequences to a 2-dimensional alignment matrix (Figure 8). In this way, *Mirage* is able to quickly convert its protein-to-genome mappings into splice-aware MSAs.

Following construction of intra-species MSAs from mapped sequences, *Mirage* adds unmapped sequences to their gene family MSAs through a global sequence-to-profile alignment stage. "Splice site" columns such as column 5 in Figure 8 are retained after this stage for use in the next stage.

## 3.5 Inter-Species MSAs Through Splice-Aware Profile Alignment

Once the full set of intra-species MSAs has been constructed, *Mirage* merges them into a complete inter-species MSA. This merging phase follows the order in which species appear in the reference file (essentially following a profile-based chained guide tree[32]). Alignments are progressively merged using a splice-site-aware global profile-to-profile alignment method, implemented in our tool

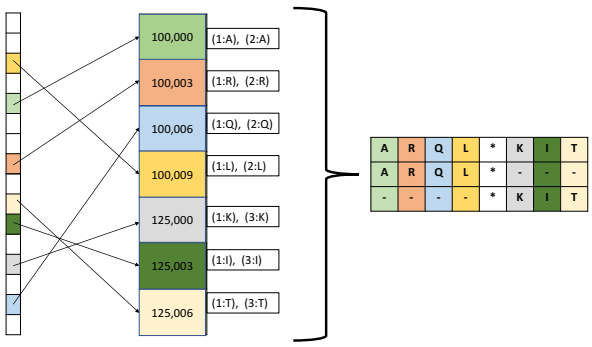


**Figure 8: Illustration of Intra-Species MSA Construction by Hashing. Each entry in the hash table converts naturally into an MSA column, as the identifier-amino acid pairs communicate which characters are placed in which columns and gap characters are placed in every row that is not represented. Additionally, whenever the difference between adjacent keys is larger than 3, a splice junction is inferred and a column consisting of special splice site markers (asterisks) is appended to the MSA (representing where introns are located).**

called *MultiSeqNW* (MSNW). The inputs to MSNW are two MSAs (treating a single sequence as an "alignment" of one sequence), and a variant of the Needleman-Wunsch[27] recurrence is employed to produce a global alignment of the input MSAs using the following scoring scheme:

(1) Match scores between residues are computed using half-bit BLOSUM62 [11] scores. Characters in heterogeneous columns contribute to the match scores of their columns proportionally to their representation in the column.
(2) Affine gap penalties are used with a gap open cost of $-11$ and a gap extension cost of $-1$.
(3) Aligning a splice site to a column containing an amino acid is prohibited.
(4) A gap at a splice site incurs a special gap-start penalty that scales with the distance $d$ to the closest splice site in the other MSA under the function $MAX(-5 \cdot 2^d, -200)$ (this rule is ignored if either input consists solely of unmapped sequences).

The first and second aspects of the scoring scheme reflect familiar features of profile-to-profile alignment implementations, but the third and fourth aspects are unique to *Mirage*'s access to MSAs with splice site markers. Fundamentally, these rules encourage alignments to agree with the observed exonic structures in the existing alignments and, as much as possible, to align sequences in exon-scale blocks (*i.e.*, to deter aligments such as those shown in Figure 2). The parameters were developed based on a desire to allow small shifts in splice site location across species, but strongly discourage major shifts. They have not been optimized.

Once all species MSAs have been merged, splice site markers are removed and minor post-processing is used to fix errors that routinely occur around splice boundaries, after which a directory containing the full output set of splice-aware MSAs (inter- and intra-species) is returned to the user.

## 3.6 *SPALN* Frequently Exhibits Recoverable Errors

During its protein-to-genome mapping phase, *Mirage* relies on the spliced alignment tool *SPALN*[15] to supplement the mappings produced by *FastMap* specifically because *SPALN* is a fast and accurate tool. There is, however, a handful of common errors associated with *SPALN*'s output that *Mirage* repairs. Simple-to-address errors include inconsistent nucleotide indexing, miscalculation of percent identity, and mislabeling of serines. A more pernicious error is the frequent annotation of "micro-exons," putative exons that encode 1 to 3 amino acids. While micro-exons are a known phenomenon, *SPALN* frequently outputs clusters of multiple consecutive micro-exons; we find that these are nearly always properly replaced by extending one or both of the longer exons on either side of a micro-exon cluster. The bookkeeping and alignment work that *Mirage* needs to detect and fix micro-exon errors is one of the most significant contributions to the disparity between program time and associated time reported for *SPALN* in Table 7.

## 3.7 Detection and Verification of Dual-coding Exons

*3.7.1 Intra-Species Transitive Alignment Detects Dual-coding Exons Expressed in Multiple Alternative Reading Frames.* Dual-coding exons are detected during the ordered traversal of codon coordinates in intra-species MSA construction. During traversal, *Mirage* tracks the distances between each pair of adjacent codon coordinate. In the same way that distances larger than 3 indicate splice sites, distances shorter than 3 indicate frameshifts, and extended runs of frameshifted entries in the list of codon coordinates represent dual-coding exons. Whenever a dual-coding exon is detected, the less common of the two reading frames is designated the alternative reading frame (ARF), and an annotation is added to the name fields of each sequence containing the alternative frame. This annotation records the range of amino acids in the ungapped sequence that correspond to the putative ARF.

*3.7.2 Conservation in Mouse.* Lift-over files[19] (files that compactly represent whole-genome alignments) were downloaded from the UCSC genome browser[20] so that positions on the human reference genome could be mapped to their homologs on the mouse genome. Because *Mirage* preserves the protein-to-genome mappings that it uses for transitive intra-species alignment as part of its output, it is straightforward to extract the specific genomic coordinates of putatively dual-coding exons. These genomic coordinates were lifted-over to the mouse genome, and the homologous genomic sequence was examined for multiple open reading frames (ORFs). To provide a sense of deviation from typical rate of multiple ORFs in exons of various lengths, the complete collection of mouse exons was inspected for multiple ORFs. The results of these paired examinations provided the data for Figure 7.

### 3.8 Assessment

Testing was performed on an Ubuntu server housed at the University of Montana with 32 Intel® Xenon® cores (2.40 GHz) and 64 GB of available RAM. *Mirage* was run with the option `-n 32` to request 32 threads and the flag `--time` to report timing data. *MAFFT* and *MUSCLE* do not support multi-threading, so our testing script spawned 32 processes that were assigned roughly equal subsets of the collection of family-specific databases for testing those methods (equality of databases being defined in terms of the total number of amino acid residues in the assigned databases). We ran *Clustal-Omega* using the `--threads=32` option. Aside from setting the number of threads used by *Clustal-Omega*, all tools were run using default parameters. Results were stored using a directory structure that mirrors *Mirage*'s output, so any analysis scripts designed to evaluate *Mirage* output could also be applied to evaluate the MSAs produced by the other tools.

All methods for evaluating the quality of the isoform MSAs were also implemented as Perl scripts.

## 4 DISCUSSION

### 4.1 Integration into the PhosphoSitePlus Web Service

PhosphoSitePlus®[13] (PSP) is an expertly curated proteomic resource of experimentally observed post-translational modifications (PTMs). Tens of thousands of new sites are added per year, requiring precise protein alignments across all members of a protein group, including all isoforms of multiple species of a protein. The error inherent in traditional protein alignment algorithms has required significant manual intervention to identify and fix broken alignments. At the time of submission, the PSP backend is being modified to use *Mirage* to compute its MSAs. We expect this to lead to accompanying improvements in the speed of curation of proteomic data, more accurate site assignments, and ultimately a more accurate understanding of the biology of PTMs.

### 4.2 Dual-coding Regions and Alternative Reading Frames

Dual-coding genomic regions are known to occur in eukaryotes [21, 30], but are thought to be relatively rare. Our identification of over 2,000 genes with putative ARFs, supported by strong evolutionary conservation, suggests that they play an underappreciated role in protein function or regulation. Structural studies of ARFs indicate that they are commonly composed of intrinsically disordered sequence[21], suggesting that ARFs may perhaps be preferred to simple exon removal in alternative splicing due to a role as a spacer – the original function of the exon(s) may be lost, but the remaining disordered sequence maintains separation of flanking domains. This simple notion is complicated by observation that at least 68 putative ARFs utilize all 3 open reading frames, and that a large fraction of ARF exons are N- or C-terminal to the protein. Clearly, further exploration is called for. We are currently investigating sequence and structural characteristics of dual-coding regions, and the distribution of ARFs in RNA-Seq and mass spectrometry datasets.

### 4.3 Availability

*Mirage* is released under the BSD-3-Clause open source license.
Software and documentation are available at
https://github.com/TravisWheelerLab/Mirage.
All benchmarks are available at
https://wheelerlab.org/publications/Nord18/Nord18.suplement.tar.gz.

## ACKNOWLEDGMENTS

We gratefully acknowledge Jon Kornhauser for his analysis of multiple sequence alignments during the development of *Mirage*, and Bin Zhang for her work configuring and running the *Mirage* package on CST AWS service. Funding by NIH P20GM103546, NIH 1R15GM123487, BD2K LINCS DCIC : U54-HL127624-02. Conflicts of interest: none declared.

## REFERENCES

[1] Helena Block, Anika Stadtmann, Daniel Riad, Jan Rossaint, Charlotte Sohlbach, Giulia Germena, Dianqing Wu, Scott I Simon, Klaus Ley, and Alexander Zarbock. 2016. Gnb isoforms control a signaling pathway comprising Rac1, Plc$\beta$2, and Plc$\beta$3 leading to LFA-1 activation and neutrophil arrest in vivo. *Blood* 127, 3 (2016), 314–324.
[2] John Brognard and Tony Hunter. 2011. Protein kinase signaling networks in cancer. *Current Opinion in Genetics & Development* 21, 1 (2011), 4–11.
[3] Robert C Edgar. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32, 5 (2004), 1792–1797.
[4] Robert C Edgar. 2010. Quality measures for protein alignment benchmarks. *Nucleic acids research* 38, 7 (2010), 2145–2153.
[5] Robert C Edgar and Serafim Batzoglou. 2006. Multiple sequence alignment. *Current Opinion in Structural Biology* 16, 3 (2006), 368–373.
[6] Walid H Gharib and Marc Robinson-Rechavi. 2011. When orthologs diverge between human and mouse. *Briefings in bioinformatics* 12, 5 (2011), 436–441.
[7] Osamu Gotoh, Mariko Morita, and David R Nelson. 2014. Assessment and refinement of eukaryotic gene structure prediction with gene-structure-aware multiple protein sequence alignment. *BMC Bioinformatics* 15, 1 (2014), 189.
[8] Emanuela Grassilli, Fabio Pisano, Annamaria Cialdella, Sara Bonomo, Carola Missaglia, Maria Grazia Cerrito, Laura Masiero, Leonarda Ianzano, Federica Giordano, Vittoria Cicirelli, et al. 2016. A novel oncogenic BTK isoform is overexpressed in colon cancers and required for RAS-mediated transformation. *Oncogene* 35, 33 (2016), 4368.
[9] M Kamrul Hasan, Tomoko Yaguchi, Yasumasu Minoda, Takashi Hirano, Kazunari Taira, Renu Wadhwa, et al. 2004. Alternative reading frame protein (ARF)-independent function of CARF (collaborator of ARF) involves its interactions with p53: evidence for a novel p53-activation pathway and its negative feedback control. *Biochemical Journal* 380, 3 (2004), 605–610.
[10] Klas Hatje, Raza-Ur Rahman, Ramon O Vidal, Dominic Simm, Björn Hammesfahr, Vikas Bansal, Ashish Rajput, Michel Edwar Mickael, Ting Sun, Stefan Bonn, et al. 2017. The landscape of human mutually exclusive splicing. *Molecular Systems Biology* 13, 12 (2017), 959.
[11] Steven Henikoff and Jorja G Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* 89, 22 (1992), 10915–10919.
[12] Desmond G Higgins and Paul M Sharp. 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* 73, 1 (1988), 237–244.
[13] Peter V Hornbeck, Bin Zhang, Beth Murray, Jon M Kornhauser, Vaughan Latham, and Elzbieta Skrzypek. 2014. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Research* 43, D1 (2014), D512–D520.
[14] Stefano Iantorno, Kevin Gori, Nick Goldman, Manuel Gil, and Christophe Dessimoz. 2014. Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. In *Multiple Sequence Alignment Methods*. Springer, 59–73.
[15] Hiroaki Iwata and Osamu Gotoh. 2012. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Research* 40, 20 (2012), e161–e161.
[16] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* 30, 14 (2002), 3059–3066.
[17] John D Kececioglu and Weiqing Zhang. 1998. Aligning alignments. In *Annual Symposium on Combinatorial Pattern Matching*. Springer, 189–208.
[18] W James Kent. 2002. BLAT–the BLAST-like alignment tool. *Genome Research* 12, 4 (2002), 656–664.
[19] W James Kent, Robert Baertsch, Angie Hinrichs, Webb Miller, and David Haussler. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse

and human genomes. *Proceedings of the National Academy of Sciences* 100, 20 (2003), 11484–11489.

[20] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. 2002. The human genome browser at UCSC. *Genome Research* 12, 6 (2002), 996–1006.

[21] Erika Kovacs, Peter Tompa, Karoly Liliom, and Lajos Kalmar. 2010. Dual coding in alternative reading frames correlates with intrinsic protein disorder. *Proceedings of the National Academy of Sciences* 107, 12 (2010), 5429–5434.

[22] Hong Li, Xiaobin Xing, Guohui Ding, Qingrun Li, Chuan Wang, Lu Xie, Rong Zeng, and Yixue Li. 2009. SysPTM: a systematic resource for proteomic research on post-translational modifications. *Molecular & Cellular Proteomics* 8, 8 (2009), 1839–1849.

[23] Han Liang and Laura F Landweber. 2006. A genome-wide study of dual coding regions in human alternatively spliced genes. *Genome research* 16, 2 (2006), 190–196.

[24] Kiersten A Liddy, Melanie Y White, and Stuart J Cordwell. 2013. Functional decorations: post-translational modifications and heart disease delineated by targeted proteomics. *Genome medicine* 5, 2 (2013), 20.

[25] Stefan Maas, Alexander Rich, and Kazuko Nishikura. 2003. A-to-I RNA editing: recent news and residual mysteries. *Journal of Biological Chemistry* 278, 3 (2003), 1391–1394.

[26] Jason Merkin, Caitlin Russell, Ping Chen, and Christopher B Burge. 2012. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* 338, 6114 (2012), 1593–1599.

[27] Saul B Needleman and Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48, 3 (1970), 443–453.

[28] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* 40, 12 (2008), 1413–1415.

[29] Sangya Pundir, Maria J Martin, and Claire O'Donovan. 2017. UniProt protein knowledgebase. *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics* (2017), 41–55.

[30] Corinne Rancurel, Mahvash Khosravi, A Keith Dunker, Pedro R Romero, and David Karlin. 2009. Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. *Journal of virology* 83, 20 (2009), 10719–10736.

[31] Chris Sander and Reinhard Schneider. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Bioinformatics* 9, 1 (1991), 56–68.

[32] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7, 1 (2011), 539.

[33] Guy St C Slater and Ewan Birney. 2005. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics* 6, 1 (2005), 31.

[34] Dorothee Staiger and John WS Brown. 2013. Alternative splicing at the intersection of biological timing, development, and stress responses. *The Plant Cell* 25, 10 (2013), 3640–3656.

[35] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 7221 (2008), 470.

[36] Youchun Wang, Huayuan Zhang, Roger Ling, Hemin Li, and Tim J Harrison. 2000. The complete sequence of hepatitis E virus genotype 4 reveals an alternative strategy for translation of open reading frames 2 and 3. *Journal of General Virology* 81, 7 (2000), 1675–1686.

[37] Robert J Weatheritt, Norman E Davey, and Toby J Gibson. 2012. Linear motifs confer functional diversity onto splice variants. *Nucleic acids research* 40, 15 (2012), 7123–7131.

[38] Travis J Wheeler and John D Kececioglu. 2007. Multiple alignment by aligning alignments. *Bioinformatics* 23, 13 (2007), i559–i568.

[39] Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, et al. 2017. Ensembl 2018. *Nucleic Acids Research* 46, D1 (2017), D754–D761.