**ABSTRACT**

**Motivation:** Accurate alignment of large numbers of sequences is demanding and the computational burden is further increased by downstream analyses depending on these alignments. With the abundance of sequence data, an integrative approach of adding new sequences to existing alignments without their full re-computation and maintaining the relative matching of existing sequences is an attractive option. Another current challenge is the extension of reference alignments with fragmented sequences, as those coming from next-generation metagenomics, that contain relatively little information. Widely used methods for alignment extension are based on profile representation of reference sequences. These do not incorporate and use phylogenetic information and are affected by the composition of the reference alignment and the phylogenetic positions of query sequences.

**Results:** We have developed a method for phylogeny-aware alignment of partial-order sequence graphs and apply it here to the extension of alignments with new data. Our new method, called PAGAN, infers ancestral sequences for the reference alignment and adds new sequences in their phylogenetic context, either to predefined positions or by finding the best placement for sequences of unknown origin. Unlike profile-based alternatives, PAGAN considers the phylogenetic relatedness of the sequences and is not affected by inclusion of more diverged sequences in the reference set. Our analyses show that PAGAN outperforms the reference alignments with fragmented sequences. Moreover, PAGAN-generated alignments of noisy next-generation sequencing (NGS) sequences are accurate enough for the use of RNA-seq data in evolutionary analyses.

**Availability:** PAGAN is written in C++, licensed under the GPL and its source code is available at http://code.google.com/p/pagan-msa.

**Supplementary information:** Supplementary data are available at Bioinformatics online.

Received on January 5, 2012; revised on April 13, 2012; accepted on April 16, 2012

1 INTRODUCTION

Sequence alignment has numerous applications but its role is especially central in evolutionary analyses of molecular sequences. These inferences are based on the identities and differences detected between homologous characters and errors in these homology statements, that is errors in the alignment of the sequences, are likely to lead to errors in any downstream analyses. The generation of high-quality alignments can be computationally laborious and the solutions often require manual assessment or editing. When such alignments need to be extended, e.g. after new sequences become available, it may be preferable to keep the relative alignment of existing sequences intact and have the new sequences aligned to this reference alignment. Such addition of sequences should take into account the evolutionary relationships of all the sequences and be performed in the correct phylogenetic context.

Alignment extension has interesting applications in the analyses of next-generation sequencing (NGS) data. Fast profile-based methods have been used for the alignment of metagenomic sequence reads of unknown origin against a set of reference sequences in phylogenetic placement studies (Katoh et al., 2009). These do not use all information available in the data, however, and flatten the reference alignment into a consensus profile that only models conserved regions shared by most sequences. On the other hand, existing read placement methods based on phylogenetic alignment (Berger et al., 2011) handle the query sequences separately and delete sites inferred as insertions, limiting their use to phylogenetic placement only. Accurate alignment of complete NGS reads is of interest e.g. in analyses of RNA-seq data that come nearly exclusively from the gene regions of the genomes. With appropriate handling of short and noisy reads, RNA-seq data allow for inexpensive large-scale comparative studies of protein-coding genes, such as inferences of selection (Yang et al., 2000), and extend the use of NGS methods to evolutionary analyses of non-model organisms (e.g. http://www.onekp.com).

Popular progressive alignment programs (e.g. Katoh et al., 2008; Larkin et al., 2009) indirectly exploit the connection between alignment and phylogeny (Sankoff, 1980) as they divide the computationally intractable multiple alignment problem into many pairwise tasks. Yet, they ignore the phylogeny during the remainder of the alignment process and produce alignments whose gap patterns are not evolutionarily meaningful (Löytynoja and Goldman, 2008). We showed earlier that phylogenetic information can be used to distinguish insertions from deletions and these two very different mutation events can then be treated correctly in the progressive alignment (Löytynoja and Goldman, 2006). The phylogeny-aware algorithm based on these ideas and implemented in the program PRANK performs very well in evolutionary alignment comparisons (Desai and Goldman, 2001, Fletcher and Yang, 2001, Jordan and Goldman, 2007; Markova-Raina and Petrov, 2011).
The idea of using graphs to represent a sequence alignment is attracting interest, as it provides a method to model the evolutionary context of sequences, especially in cases of complex evolutionary relationships. The optimal alignment path for graphs is not always straightforward, as seen in the example of graphs 2 and 4, to indicate that the presence of vertices 3 and 4 is uncertain. Representing both pyrimidines, and having two incoming edges, from vertices 1 and 2, indicates that they might have been added at a later stage in the evolutionary process.

Deletions can be represented as unused (asterisks) in the skip-over fragment. This method allows the use of dynamic programming algorithms to be extended to alignment of graphs, by using a method that can handle uncertainties in the alignment process. For example, when aligning graphs representing the positions of insertions and deletions, we can use a method that allows changes in the alignment parameters to be made independently, to better model the evolutionary process.

Here, we outline a new general-purpose method for phylogeny-aware alignment of sequence graphs and apply it to phylogenetic extension of existing alignments. Our method, called PAGAN, is based on the same principle as PRANK and uses evolutionary information to distinguish insertions from deletions. In contrast to the greedy insertion-calling of the original approach, sequence graphs provide a flexible framework to model the phylogenetic evidence from related sequences and allow building a robust progressive aligner that tolerates errors in the guide phylogeny.

The key advantage of our graph representation of sequences is the ability to describe uncertainty regarding the presence of characters at certain sequence positions. This is beneficial during the progressive alignment of sequences (Fig. 1) but it can also be used to represent uncertainties in unaligned sequences or in the inferred ancestral sequences. As our approach considers all sequences as ancestral sequences, especially the tree alignment method (Hein, 1989) and has similarities to earlier methods for global alignment of multiple sequences, it can be used to align more distantly related sequences in the reference alignment.

We developed PAGAN as a progressive aligner that tolerates errors in the guide phylogeny. It reconstructs sequence graphs to represent the ancestral nodes and aligns the new sequences against the ancestor of their sister clade, indicated by X. Each alignment of a sequence, Xn, creates a new parent node, Xn, against which the next sequence is aligned. After finishing the alignment, the sub-tree with the new sequences is inserted back to the tree structure.

Here, we present a new general-purpose method for phylogeny-aware alignment of sequence graphs and apply it to phylogenetic extension of existing alignments. Our method, called PAGAN, is based on the same principle as PRANK and uses evolutionary information to distinguish insertions from deletions. In contrast to the greedy insertion-calling of the original approach, sequence graphs provide a flexible framework to model the phylogenetic evidence from related sequences and allow building a robust progressive aligner that tolerates errors in the guide phylogeny.

We tested PAGAN in the extension of real reference alignments with new protein and DNA sequences and found that it can successfully handle data of great variety in length and evolutionary divergence as well as different sizes of reference alignments. The accuracy of alignment cannot be tested with real data so we simulated datasets representing gene families of closely related paralogues. We used PAGAN to extend both protein and DNA reference alignments with new data and compared its performance with that of alternative extension approaches. To test larger problems, we used PAGAN and the best alternative method, hmmalign (Eddy, 2011), for a re-analysis of metagenomic data of Mirarab et al. (2012) consisting of reference alignments of 500 sequences and addition of 5000 sequence fragments. Finally, we quantified the effects of different factors on alignment accuracy and compared PAGAN and hmmalign under a simplified setup that allowed changing the different parameters independently. Our results show that PAGAN produces exceptionally accurate alignments and its phylogenetic approach can efficiently use the evolutionary information available while remaining unimpaired by more distantly related sequences in the reference alignment.

2 METHODS

In the following sections we first describe the main concepts of our new method for phylogeny-aware alignment of graphs using partial-order graphs. We then outline how we apply this method, called PAGAN, in the extension of DNA and protein alignments with new sequences and compare its performance with existing methods. Finally, we dissect in more detail the impact of the reference alignment and the query sequence on the performance of the best performing methods.

2.1 Phylogeny-aware graph alignment algorithm

The conversion of a regular sequence to a graph is trivial (Supplementary Fig. S1). Two such graphs could be aligned with a standard dynamic-programming algorithm. Partial-order graphs can represent more than one sequence of characters, however, and allow modelling of e.g. evolutionary units of multiple characters, non-linear dependencies among the characters and uncertainties in the input data (Supplementary Figs S1 and S2).

The representation of sequences with graphs is especially attractive in progressive alignment that attempts to backtrack the tree-like hierarchical structure of relatedness among a set of sequences (novelty and Goldman, 2006). Each alignment clusters two sister nodes, representing either single
sequences or previous alignments, and defines a new node to represent this pairwise solution. The challenge of progressive alignment is that insertions cannot be distinguished from deletions at the time of aligning a pair of sequences but failing to account for their different properties is likely to cause alignment error (Löytynoja and Goldman, 2008).

A graph can describe this uncertainty regarding the type of mutation event giving the character associated to a vertex $i$ of graph $x$, the score for matching characters at vertices $x_i$ and $y_j$ is:

$$\log \left( \frac{P(\text{chr}(x_i), \text{chr}(y_j) | t)}{q_a} \right)$$

where $q_a$ is the frequency of $a$, $q_a = (q(\text{chr}(x_i)) + q(\text{chr}(y_j))) / 2$, and $P(t)$ is the substitution probability between characters $a$ and $b$ given the evolutionary distance $r$ and the substitution model. A notable difference to the standard score is the additional term $q_a$, discussed in more detail in the Supplementary Material. As a further simplification, PAGAN uses weighted parsimony reconstruction of ancestral characters states and, for greater speed, the number of alternative character states for amino acid and codon data is limited to two (see Supplementary Material).

PAGAN can reconstruct ancestral sequences for an existing alignment and then extend that by aligning new sequences against the extant or inferred ancestral sequences. The reconstruction of ancestral sequence graphs for a reference alignment is not different from the de novo alignment except that the alignment solution is read from the input data. The addition of new sequences should be performed in their correct phylogenetic context; while PAGAN allows the user to define or constrain the possible phylogenetic positions for sequences coming from a known origin, it can also search for the optimal placement for unknown data. During the extension, PAGAN takes the target nodes, represented by sequence graphs, out of the tree structure and aligns the sequences assigned to each target using a progressive algorithm (Fig. 2). With data from mixed sources, the sequences for each target species paralogue.

We downloaded the simulated test data of Mirarab et al. (2012) and analyzed the first ten replicates of the three different evolutionary scenarios. We used true simulated reference alignments and reference trees with RAxML-estimated branch lengths. We extended these alignments with the 5000 query sequences using PAGAN’s experimental heuristics to quickly assign the query sequences to targets.

The accuracy of the resulting extended alignment was measured as above except that each QS was compared with the closest reference sequence.

2.3 Extension of large alignments

We downloaded the simulated test data of Mirarab et al. (2012) and analyzed the first ten replicates of the three different evolutionary scenarios. We used true simulated reference alignments and reference trees with RAxML-estimated branch lengths. We extended these alignments with the 5000 query sequences using PAGAN’s experimental heuristics to quickly assign the query sequences to targets. These heuristics perform Eberon local alignments (Stute and Hiraoka, 2009) between the 5000 query sequences and 999 target sequences, the latter including extant sequences from the reference alignment and PAGAN-inferred ancestral sequences. The queries were assigned to their best-scoring target nodes and those not producing significant hits were discarded. We also aligned the same datasets with hmmalign and analyzed the resulting alignments (Löytynoja et al., 2005) using OrthoMCL. The accuracy of the resulting extended alignment was measured as above except that each QS was compared with the closest reference sequence.

2.4 Impact of reference alignment

The impact of the RA composition on the accuracy of alignment extension was tested using the two best-performing methods from the first set of tests (see Section 3). PAGAN and hmmalign. The data were simulated using ultrametric trees that differed in the following parameters: (i) composition of the QS; (ii) size of the ingroup; (iii) evolutionary divergence; and (iv) (for hmmalign) size of the outgroup. The simulation trees consisted of three 32-sequence sub-trees with one additional query sequence placed at different positions within the central sub-group (Fig. 6 and c). We call these basic topologies ‘close’, ‘intermediate’ and ‘distant’ and their ingroup (the central
subgroup most closely related to the query 'large'. To study the effect of reduced phylogenetic information in the RA, further datasets were created with the ingroup cut down to two maximally divergent sequences. These three topologies are subsets of the full topologies but have 'small' ingroups. As the profile hidden Markov models (HMMs) of HMMER are affected by study, we focus on a novel feature of the method and a task that has no future, is meant to replace our earlier method PRANK. In this protein and codon sequences when a guide tree is provided and, in this study, we focus on a novel feature of the method and a task that has no

3 RESULTS

3.1 Phylogeny-aware graph alignment algorithm

PAGAN is capable of inferring de novo multiple alignments of DNA, protein and codon sequences when a guide tree is provided and, in the future, is meant to replace our earlier method PRANK. In this study, we focus on a novel feature of the method and a task that has no satisfactory existing solution, the extension of multiple alignments with new data in a phylogeny-aware manner.

We tested our new method in the extension of EnsemblCompara GeneTrees alignments with new sequences, focusing on plausible use cases such as update of an alignment repository after inclusion of new species [Supplementary Fig. S2] and analysis of RNA-seq data from a non-model organism [Supplementary Figs S3 and S9]. The resulting alignments and the assembled sequence contigs are highly similar to the original ones but, as the original data may also contain errors, the analyses do not allow assessment of the true accuracy of the method. To further understand and illustrate the performance of the method we tested it with simulated data.

3.2 Comparison of methods for alignment extension

We compared methods for alignment extension using simulated data representing a mammalian gene family [Fig 3b]. The codon data, analyzed both as DNA and protein, were simulated under purifying selection [model M0 with $\omega = 0.15$] and many substitutions were synonymous on the protein level: despite the three-times greater number of sites for DNA, the sequence identity of DNA and protein sequences were similar, 84–90% between human and Primate and 70–80% between mouse and Rodent over the three levels of evolutionary divergence. Each reference alignment (RA), consisting of 67 sequences, was extended with 250 fragments of 30, 60 or 120 bases/amino acids (50 fragments per query sequence) or with five full-length sequences, and the accuracy of homology inference was measured.

We tested five alignment methods for both data types and two additional methods that only support DNA or protein data. The methods tested for both were PAGAN/guided, PAGAN/free (v.0.33), hmmalign (from the HMMER package, v.3.0), ClustalW (v.2.1), MAFFT (v.6.860b); for DNA we also used PaPaRa (RAxML v.7.2.6); Berger and
The accuracy of alignment of DNA (top row) and protein (bottom row) QSs against the corresponding reference alignment using different alignment methods. The x-axis indicates the length of the fragments aligned and the sub-panels show two of the five query species analyzed. Columns (a)–(c) correspond to trees with branch lengths multiplied by 1.5, 2.0 and 2.5, respectively. The accuracy is measured as the correctness of the site-wise homology inference with respect to the closest human/mouse reference sequence.

Fig. 4. The performance of other methods shows that classical global alignment methods for the extension of alignments with new DNA and protein sequences (Fig. 3). The two modes of running PAGAN, the ‘guided’ approach with pre-defined locations to place the sequences and the ‘free’ approach using heuristic local alignment to find the best location, had slightly different performance on different datasets. PAGAN/guided did better on the alignment of very short fragments that contain little information to infer their correct placement. Despite the prior information, also the guided approach was affected by multiple target nodes and a proportion of shorter sequences were misplaced (Supplementary Fig. S7). PAGAN/free slightly outperformed the guided approach in the alignment of longer fragments (Supplementary Fig. S9) and it did this despite a high fraction of the new sequences being placed to incorrect nodes (Supplementary Fig. S9). An explanation of this considers the long branches around the query sequence in our simulation tree: it is often advantageous to use outgroup information to resolve the mutation events that have taken place in the descendants of the true target node and place the sequence at a deeper location. On the other hand, the placement algorithm visits the tip nodes first and, if no mutations have occurred in the descendants, the greedy approach places the sequences to nodes visited earlier: up to 7% of short fragments from Primate_1 are misplaced (Supplementary Fig. S7).

PAGAN and hmmalign were consistently the two most accurate methods for the extension of alignments with new DNA and protein sequences (Fig. 3). The two modes of running PAGAN, the ‘guided’ approach with pre-defined locations to place the sequences and the ‘free’ approach using heuristic local alignment to find the best location, had slightly different performance on different datasets. PAGAN/guided did better on the alignment of very short fragments that contain little information to infer their correct placement. Despite the prior information, also the guided approach was affected by multiple target nodes and a proportion of shorter sequences were misplaced (Supplementary Fig. S7).

PAGAN/free slightly outperformed the guided approach in the alignment of longer fragments (Supplementary Fig. S9) and it did this despite a high fraction of the new sequences being placed to incorrect nodes (Supplementary Fig. S9). An explanation of this considers the long branches around the query sequence in our simulation tree: it is often advantageous to use outgroup information to resolve the mutation events that have taken place in the descendants of the true target node and place the sequence at a deeper location. On the other hand, the placement algorithm visits the tip nodes first and, if no mutations have occurred in the descendants, the greedy approach places the sequences to nodes visited earlier: up to 7% of short fragments from Primate_1 are misplaced (Supplementary Fig. S7).

The performance of other methods shows that classical global alignment methods, MAFFT, ClustalW and ClustalO, struggle in the alignment of short sequence fragments (Fig. 3). MAFFT performed relatively well with full length sequences whereas the accuracy of two Clustal variants was unacceptably low. In contrast to these, hmmalign aligned short fragments nearly as well as full length sequences and was consistently one of the best performing methods. Hmmalign’s excellent results should be taken with a grain of salt, however, as our accuracy score is based on sites shared by the query and reference only. By ignoring sites inserted since the species split, this score is overly lenient with profile-based methods that leave many insertions unaligned.

The performance of PaPaRa, based on a variant of phylogeny-aware algorithm, was good in the alignment of closely related DNA sequences but as it does not produce real multiple alignments—it deletes sites in the query sequences that are inferred as insertions—the method is only meaningful for its original task, phylogenetic placement of sequences. On the other hand, results for PaPaRa are based on reconstructed full-length query sequences (see the Supplementary Material for details) and in phylogenetic placement analyses its error may be smaller than reported.

We repeated the PAGAN and hmmalign analyses of DNA fragment data after adding NGS-like errors. The added noise had only a small negative impact on their accuracy (Supplementary Fig. S9). The run times and peak memory usage of different methods were compared on a workstation with 2.40 GHz Intel Xeon CPUs. Two of the methods, MAFFT and hmmalign, are exceptionally fast and perform an alignment in seconds (Table 1). In contrast, PAGAN/free and PaPaRa needed 3.3 and 4.7 min for the most time-consuming alignment. These are the first versions of each software, however: the authors of PaPaRa have reported forthcoming speed-ups for their method and we are also working to accelerate ours. Of the methods tested, ClustalO is the most memory hungry but even that can easily be used on a standard personal computer. The memory usage of PAGAN is dominated by the dynamic programming matrix and, due to a more complex data structure needed for the graph representation, the requirements are, for example, somewhat higher than that of MAFFT with a comparable alignment strategy.

3.3 Extension of large alignments

Our graph alignment algorithm is not yet well-optimized for speed and the first version of PAGAN is mainly targeted at small and
Table 1. Execution time and peak memory usage for the extension of EnsTr2 reference alignment with 250 sequence fragments

| Method        | DNA fragments (s) | Protein fragments (s) |
|---------------|------------------|----------------------|
|               | 60 nt            | 120 nt               | 60 aa                | 120 aa               |
| Mafft         | 1.9              | 1.0                  | 0.4                  | 0.3                  |
| Hmalign       | 2.2              | 4.4                  | 1.9                  | 4.6                  |
| Clustalo      | –                | –                    | 16.5                 | 26.3                 |
| Clustaw       | 75.8             | 143.0                | 27.0                 | 48.7                 |
| Pagan/guided  | 73.6             | 137.0                | 26.2                 | 46.0                 |
| Pagan/free    | 186.0            | 200.0                | 47.2                 | 57.9                 |
| PapaRa        | 144.0            | 284.0                | –                    | –                    |
| Mafft         | 103              | 103                  | 103                  | 103                  |
| Hmalign       | 198              | 198                  | 198                  | 198                  |
| Clustalo      | –                | –                    | 569                  | 518                  |
| Clustaw       | 31               | 31                   | 16                   | 16                   |
| Pagan/guided  | 298              | 334                  | 133                  | 155                  |
| Pagan/free    | 356              | 413                  | 153                  | 175                  |
| PapaRa        | 84               | 85                   | –                    | –                    |

Time (s), Memory (Mb)

nt, nucleotide; aa, amino acids.

Fig. 5. The accuracy of PAGAN and Hmalign in the extension of reference alignments of 500 DNA sequences with 5000 query fragments. For the easy set (M4 dataset from Mirarab et al. 2012; open symbols), both methods align >96% sites correctly; for the moderate (M3; crossed) and hard (M2; solid) sets, the accuracy of PAGAN is high (circles) but the fast heuristics fails to place half of the queries. Hmalign aligns all the queries but its accuracy for M2 and M3 is low (squares). For the fragments aligned by both methods, the alignments by Hmalign are less accurate (diamonds).

3.4 Impact of reference alignment

As shown by the previous analysis, PAGAN has the potential to scale up to large alignment extension tasks while its approach for modelling of insertions/deletions and uncertainty is especially well-suited for analyses of short and noisy NGS data. Hmalign has been used for the extension of reference alignments with new sequences in metagenomic analyses (Matsen et al. 2010; Stark et al. 2014) and recently Mirarab et al. (2015) developed an approach that applies Hmalign on subsets of the reference alignment. The latter mainly focused on computation time, however, and did not systematically study the strategies for choosing the optimal RA for alignment extension analyses, nor its effect on homology inference.

To understand the factors affecting alignment extension with phylogenetic and profile-based methods, we tested PAGAN and Hmalign with idealized data mimicking an RNA-seq study (see Section 2 Fig. 5b, c). Our set-up lets us assess the effects of (i) the phylogenetic position of the query species; (ii) the number of closely related reference species; (iii) the evolutionary divergence of the reference; and (iv) the inclusion of more-distantly related reference species. To assess the impact of sequence divergence on the ancestor reconstruction, we constrained the placement of the QR with the guide phylogeny. Incorrect topologies, or correct topologies but with incorrect root position, could cause alignment/placement errors. A full investigation of this, beyond typical use cases with known reference trees or inferred trees as studied here, is beyond the scope of this article but will be considered in future work.

Although the magnitude of difference and the relative performance of alternative approaches varies, the phylogenetic approach of PAGAN with full data (ingroup “large”) consistently produces the most accurate alignments (Table 4). With ‘close’ sets, the removal of sequences does not affect PAGAN’s reconstruction of the target ancestor and its performance on full and reduced
Although both methods lose accuracy, PAGAN is more consistent. The accuracy of extending reference alignments with new sequences varies with the evolutionary divergence of the RA (ingroup ‘large’ versus ‘small’) is nearly identical. When the query sequence branches out deeper in the tree (‘intermediate’ and ‘distant’), PAGAN shows the benefit from the phylogenetic information provided by denser sequence sampling: the alignments on the full RA are more accurate than those on the reduced ones, the difference growing with the increasing evolutionary divergence.

Similarly, it is understandable that hmmalign’s best performance is in the alignment of reads from a closely related query sequence against a profile based only on the two sequences from the sister subtree (‘clade’, ‘small’); the information from more distant subgroups only brings noise and the noise-to-signal ratio is at its worst when the central subgroup is represented by two sequences only (‘full’, ‘small’). The position of the query sequence is crucial, however, and the approach giving the best result for the closely related query sequence gives clearly the least correct alignments when the query sequence is deep and has a long history of its own. Although the profiles built from the full RA (‘full’, ‘large’) give marginally better results in the alignment of the most difficult cases (Table 2, bottom row), the inclusion of large numbers of sequences in the profile is generally not the best policy. Crucially, this conflicts with the requirements of real-life studies such as phylogenetic placement where the RA should be maximally representative.

As expected, decreasing similarity between the query and the RA makes the alignment more difficult (Table 2, depths 0.45 and 0.60). Although both methods lose accuracy, PAGAN is more consistent in its performance and the improvement over hmmalign grows with evolutionary divergence. As a demonstration of its efficient use of phylogenetic information, the relative improvement of PAGAN with the full RA over any other approach is greatest on the analyses of most diverged datasets (Table 2, bottom row). The correct use of phylogenetic information is important for real-life analyses: de novo alignment of distantly related sequences is error-prone and good aligners produce much better reference alignments if long sequences give clearly the least correct alignments when the query sequence branches out deeper in the tree. Although PAGAN is a truly phylogenetic method and, while it efficiently uses the information available in related sequences, our analyses show that PAGAN’s phylogenetic approach clearly outperforms most alternative methods, the improvement being especially striking in the alignment of short sequence fragments. We were impressed by the good performance of hmmalign but also noticed that its performance is highly dependent on the reference alignment used for the construction of the profile HMM.

With a carefully chosen reference alignment, hmmalign’s accuracy was in some cases comparable to that of PAGAN. In real-life analyses, one cannot typically maximize both the sensitivity of the profile HMM and the breadth of sequences included, and the widely used practice of including all the diversity available heavily penalizes hmmalign’s performance. In contrast, PAGAN is a truly phylogenetic method and, while it efficiently uses the information from closely related sequences, it is not affected by the inclusion of more distantly related ones in the reference set. We tested PAGAN in the extension of large alignments and found the initial results very promising. We believe that improvements in the assignment of queries to target nodes and speed-ups in the algorithm will make PAGAN also a competitive method for large-scale metagenomic analyses. Unlike alternative methods for the task, PAGAN aligns also insertion sites and includes full-length sequences in the resulting multiple alignment. The latter may not be crucial in phylogenetic placement of query fragments relative to the reference sequences but it will provide additional information to resolve the relations between the newly added sequences and will allow connecting related fragments to longer contigs.

Alignment extension and phylogenetic placement has interesting applications in the analyses of increasingly abundant sequencing

Table 2. The accuracy of extending reference alignments with new sequences

| Simulation Depth | Query | Ingroup Large | Ingroup Small | Ingroup Large | Ingroup Small |
|------------------|-------|---------------|---------------|---------------|---------------|
|                  | PAGAN |               |               | hmmalign/full |               |
|                  |       |               |               |               |               |
|                  |       |               |               |               |               |

1690
data produced by NGS technologies and we have implemented extensive support for such data. Failing costs per base will allow the sequencing of full genomes for many new non-model species and transcriptomes for even more species. We believe that there remains a need for comparative methods like ours, e.g. in the integration of transcriptome datasets into large reference alignments of closely related species, including the precise differentiation of close paralogues. We also envision extending our approach to the alignment of graphs produced by the de novo assemblers and using phylogenetic information to disambiguate these graphs into the correct separate sequences.

In addition to their alignment, we want to emphasize the advantages of graphs in the representation of the input sequence data. Graphs have direct applications in the modeling of NGS data from specific sequencing platforms but they can also be used to represent other features, such as repeat structures, that affect sequences’ evolution and thus have an impact on their alignment. The evolutionary process varies across sequence sites and many features related to this may one day be inferred by sophisticated alignment methods from the input sequences along with their alignment. It seems easier, however, to start with separate tools for the annotation of sequences, such as detection of low-complexity repeats, and pass this information to a generic alignment method. We believe that partial-order graphs are ideal for carrying that information.

ACKNOWLEDGEMENTS

We thank Gregory Jordan, Tim Massingham, Javier Herrero and Paul Flieck for their help, insightful comments and support during the course of this project.

Funding: Wellcome Trust [GR078968 to N.G.]; Biocenter Finland. A part of the computation was performed at CSC – IT Center for Science Ltd.

Conflict of Interest: none declared.

REFERENCES

Berger,S. and Stamatakis,A. (2011) Aligning short reads to reference alignments and trees. Bioinformatics, 27, 2068–2075.
Berger,S. et al. (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. Syst. Biol., 60, 291–302.
Deivos,M.G. and Gil.M. (2010) Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome. Biol., 11, R37.
Eddy,S. (2011) Accelerated Profile HMM Searches. PLoS Comp. Biol., 7, e1002195. doi:10.1371/journal.pcbi.1002195
Fluates,W. and Yang,Z. (2009) INDELible: a flexible simulator of biological sequence evolution. Mol. Biol. Evol., 26, 1879–1888.
Fluates,W. and Yang,Z. (2010) The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol. Biol. Evol., 27, 2257–2267.
Grishin,O. (1982) An improved algorithm for matching biological sequences. J. Mol. Biol., 162, 705–708.
Hein,J. (1989) A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. Mol. Biol. Evol., 6, 649–666.
Jordan,G. and Goldman,N. (2012) The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol. Biol. Evol., 29, 1125–1139.
Kato,R. et al. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res., 30, 3059–3066.
Kruskal,J. and Sankoff,D. (1983) An anthology of algorithms and concepts for sequence comparison. In Sankoff,D. and Kruskal,J. (eds). Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison, Addison-Wesley Reading, MA, pp. 265–310.
Laskin,M. et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics, 23, 2947–2948.
Lee,C. et al. (2002) Multiple sequence alignment using partial order graphs. Bioinformatics, 18, 452–461.
Löytynoja,A. and Goldman,N. (2005) An algorithm for progressive multiple alignment of sequences with insertions. Proc. Natl. Acad. Sci. USA, 102, 10557–10562.
Löytynoja,A. and Goldman,N. (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science, 320, 1632–1635.
Löytynoja,A. and Goldman,N. (2009) Uniting alignments and trees. Science, 324, 1528–1529.
Markova-Rama,P. and Petrov,D. (2011) High sensitivity to alignment and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. Genome Res., 21, 863–874.
Massingham,T. and Goldman,N. (2012) simNGS and simLibrary – software for simulating next-gen sequencing data. Available at http://www.ceb.ubc.ca/goldman-srv/simNGS/.
Matsen,F. et al. (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinform., 11, 336.
Mirarab,S. et al. (2012) SEPP: SAGE-enabled phylogenetic placement. Proc. Pac. Symp. Biocomput., 17, 247–258.
Paten,B. et al. (2008) Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res., 18, 1829–1843.
Sankoff,D. (1975) Minimal mutation trees of sequences. SIAM J. Appl. Math., 28, 35–42.
Sievers,F. et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol. Syst. Biol., 7, 539.
Slater,G. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinform., 6, 91.
Stamatakis,A. (2006) RAXML-VI HPC: maximum likelihood-based phylogentic analyses with thousands of taxa and mixed models. Bioinformatics, 22, 2686–2690.
Stamatakis,A. et al. (2010) MLTreeMap–accurate maximum likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. BMC Genomics, 11, 461.
Yang,Z. et al. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics, 155, 431–449.