A SIMPLE SHORTCUT TO UNSUPERVISED ALIGNMENT-FREE PHYLOGENETIC GENOME GROUPINGS, EVEN FROM UNASSEMBLED SEQUENCING READS

SEBASTIAN MAURER-STROH[1]
Bioinformatics Institute (BII),
*Agency for Science Technology and Research (A*STAR),*
30 Biopolis Street,
#07-01, Matrix,
Singapore 138671
and
School of Biological Sciences (SBS),
Nanyang Technological University (NTU),
60 Nanyang Drive,
Singapore 637551
sebastianms@bii.a-star.edu.sg

VITHIAGARAN GUNALAN[1]
Bioinformatics Institute (BII),
*Agency for Science Technology and Research (A*STAR),*
30 Biopolis Street,
#07-01, Matrix,
Singapore 138671
vithiagarang@bii.a-star.edu.sg

WING CHEONG WONG
Bioinformatics Institute (BII),
*Agency for Science Technology and Research (A*STAR),*
30 Biopolis Street,
#07-01, Matrix,
Singapore 138671
wongwc@bii.a-star.edu.sg

FRANK EISENHABER
Bioinformatics Institute (BII),
*Agency for Science Technology and Research (A*STAR),*
30 Biopolis Street,
#07-01, Matrix,
Singapore 138671
and
School of Computer Engineering (SCE),
Nanyang Technological University (NTU),
50 Nanyang Drive,
Singapore 637553
and
Department of Biological Sciences (DBS),
National University of Singapore (NUS),
8 Medical Drive 4,
Singapore 117597
franke@bii.a-star.edu.sg

[1] Corresponding Author.

Abstract:
We propose an extension to alignment-free approaches that can produce reasonably accurate phylogenetic groupings starting from unaligned genomes, for example, as fast as 1 minute on a standard desktop computer for 25 bacterial genomes. A 6-fold speed-up and 11-fold reduction in memory requirements compared to previous alignment-free methods is achieved by reducing the comparison space to a representative sample of k-mers of optimal length and with specific tag motifs. This approach was applied to the test case of fitting the enterohemorrhagic O104:H4 *E.coli* strain from the 2011 outbreak in Germany into the phylogenetic network of previously known *E.coli*-related strains and extend the method to allow assigning any new strain to the correct phylogenetic group even directly from unassembled short sequence reads from next generation sequencing data. Hence, this approach is also useful to quickly identify the most suitable reference genome for subsequent assembly steps.

Keywords: genome phylogeny; alignment free; unsupervised; Enterohemorrhagic *E.coli*; next generation sequencing.

1. Introduction

Next generation sequencing technologies allow for unprecedented speed and ease of deriving complete genomes. In the case of the enterohemorrhagic O104:H4 *E.coli* strain, which caused a widely publicized outbreak in Germany in spring of 2011,[1] the genome sequence was rapidly made available by multiple sources, while subsequent analyses of the genome were lagging behind.[2–4] One of the many aspects of interest for characterizing bacterial strains is the phylogenetic grouping. Analysing phylogenetic relationships at the whole genome level is a complicated and tedious task but the effort can provide valuable insights into biology and evolution of the analysed taxa. Such analysis has recently been completed for a broad selection of bacterial strains related to *E.coli*.[5] By carefully analysing oligonucleotide occurrences within these genomes, we propose a simple shortcut to unsupervised genome-based phylogenetic groupings using the German O104:H4 strain as example.

2. Method

Determination of distance measure
Using the 21 genomes from the previous detailed *E.coli* phylogenetic study, the number of possible different oligonucleotides for lengths from 1 to 36 was calculated and plotted (Figure 1). The theoretical number n of different sequences of length l consisting of any mixture of 4 characters (ACGT) should grow exponentially following $n=4^{l}$. However, it can be seen that the actual observed number of oligonucleotides quickly deviates from the exponential curve and rather represents a sigmoidal shape (Figure 1).

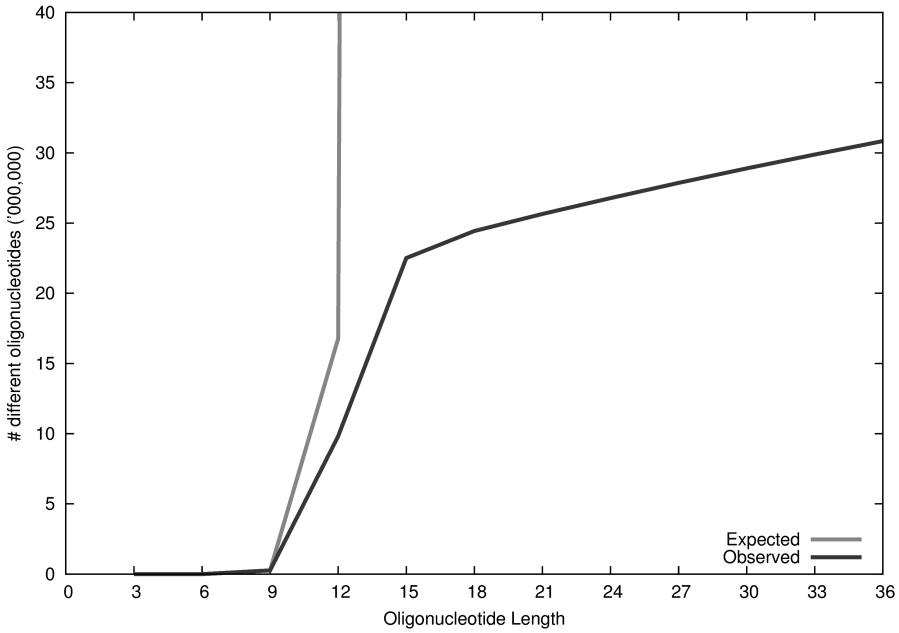

Figure 1. Comparison of number of oligonucleotides as a function of oligonucleotide length, between expected (light grey) and observed (dark grey).

The likely reason for this observation is that the naturally observed oligonucleotides in these genomes are largely undersampled compared to the combinatorically possible sequence space. The break of the sigmoidal curve appears to happen when oligonucleotides above a critical length become more or less specific for their genome. This notion is not new and has already found different applications in genome biology, for example in DNA barcode based indexing of species.[6] However, the aim here was to circumvent the use of fully species-specific oligonucleotides and find a 'sweet-spot' where oligonucleotides are still shared within their close phylogenetic group but not common to all species under comparison. To this end, theoretical aspects of k-mer statistics for biological sequence comparison have been studied in detail before.[7–10] Previous work has shown a similar approach of studying optimal length of peptides shared among close homologues for ultra-fast protein searches.[11] As also shown previously, simple oligonucleotide or peptide overlap measures between two genomes can be indicative of their phylogenetic distance.[12–14] An apparent immediate advantage of this approach compared to classical methods is the independence from sequence alignments which can be problematic and extremely slow at genome-wide scales.[15] Using such alignment-free approach, an intuitive pairwise genome similarity is the fraction of the number of overlapping oligonucleotides amongst the 2 genomes ($n_{overlap}$) divided by the total number of oligonucleotides between these genomes (n_{total}), which can simply be written as distance:

$$d = 1 - \frac{n_{overlap}}{n_{total}} \quad (1)$$

Determination of optimal oligonucleotide length

A critical part of the proposed shortcut would involve determining the optimal oligonucleotide length, since too short would mean being unspecific and too long would mean becoming too genome-specific to be informative for phylogenetic grouping.[16,17] Representative pairs of closely and remotely related genomes were chosen and the number of overlapping oligonucleotides at different lengths calculated (Figure 2). As expected, the closely related pair has a much larger number of shared oligonucleotides compared to remotely related pairs. However, all of the pairs share the same maximal oligonucleotide overlap around 15 bases which is also the beginning of the plateau of the sigmoidal curve in Figure 1. Interestingly, a progressive reduction was observed in this maximal oligonucleotide overlap between pairs of genomes, suggesting that the optimal oligonucleotide length might be a function of the pairwise distance between genomes and therefore will need to be calculated for each specific situation in which this approach is employed.

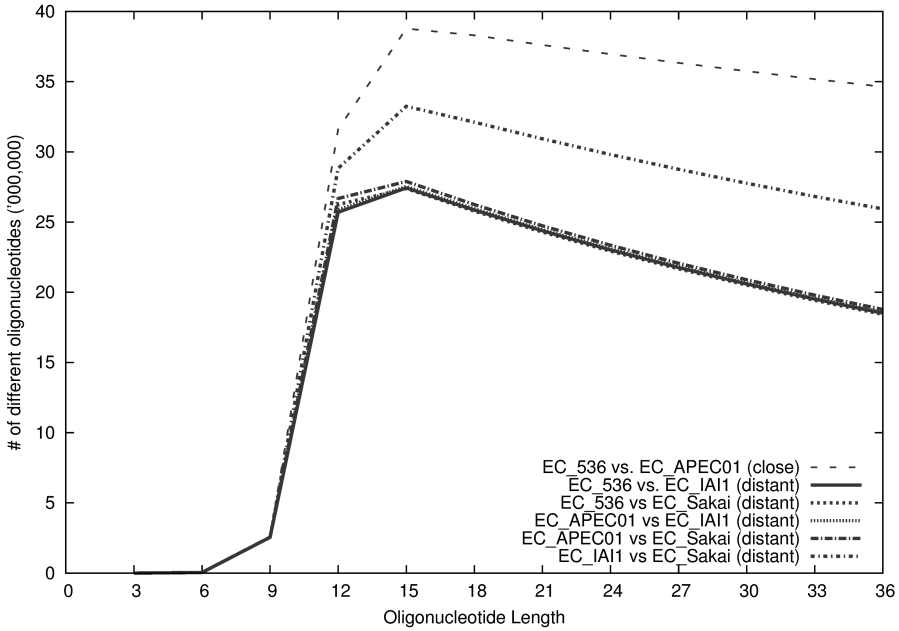

Figure 2. Comparison of shared oligonucleotides obtained as a function of oligonucleotide length for closely related strains (dark grey) and distant strains (light grey) from the EC_536 strain of enterohemorrhagic bacteria.

Representative subsamples using conserved sequence tags

At this optimal length of 15 nucleotides, the *E.coli* related genomes analysed here have approximately 48 million different oligonucleotides per genome and calculating overlaps and intersections is a feasible but still time- and compute-intensive task. It was therefore sought to further reduce the number of required calculations by using representative subsamples. In the case of the well-studied *E.coli* enterobacteria, such subsamples could be derived from oligonucleotides from known marker genes, for example ones used for multi-locus sequence typing.[18] This would however depend on species-specific prior knowledge – a fully unsupervised approach would likely

increase the utility of this method for other applications. Straight-forward random selection of oligonucleotides for subsamples is statistically unlikely to give enough overlap matches between genomes, mandating the use of some kind of sequence anchors but without the complicated step of genome alignments. Given our earlier observation of the undersampled sequence space for oligonucleotides above a critical length, a possible solution could be to use universally conserved signature tags to fish out oligonucleotides that should then be more likely to be related between genomes compared to random selection. There are several naturally recurring consensus sequence motifs within genomes typically related to gene structure or transcription or translation regulation. Several of these may be species- or kingdom-specific, such as the Kozak and Shine-Dalgarno consensus motifs in Eukaryotes and Bacteria, respectively.[19] In order to identify a species-independent anchor tag suitable for oligonucleotide selection, the importance and role of anchor tags were investigated by examination of Pearson's R^2 correlation with the full 15-mer count, as well as the compute time required for an all-against-all comparison of the set of bacterial genomes (Table 1).

Table 1. Pearson's R^2 correlation with full oligo count, median oligo count and compute time for candidate sequence anchor tags.

Motif Description	Regular Expression	R^2	#oligos (median)	time(s)
12 + stop codon	[ACGT]{12}TAG	0.989	19458	40.9
start codon + A + 11	ATGA[ACGT]{11}	0.986	15071	35.2
11 + A + stop codon	**[ACGT]{11}ATAG**	**0.984**	**7541**	**25.5**
11 + AAAA	[ACGT]{11}AAAA	0.984	17551	36.6
start codon + 12	ATG[ACGT]{12}	0.984	43909	73.3
10 + GA + stop codon	[ACGT]{10}GATAG	0.983	2410	20.9
stop codon + A + 11	ATAG[ACGT]{11}	0.983	7546	25.7
11 + CCCC	[ACGT]{11}CCCC	0.978	6068	22.5
9 + AGA + stop codon	[ACGT]{9}AGATAG	0.969	562	18.8
Shine-Dalgarno (Bac) + 9	AGGAGG[ACGT]{9}	0.924	289	17.2
11 + C + stop codon	[ACGT]{11}CTAG	0.842	764	20.4
Kozak (Euk) + 5	GCC[AG]CCATGG[ACGT]{5}	0.464	4	18.5

The classical stop codon “TAG” alone has the best correlation although due to its abundance it includes many oligonucleotides whose comparison between genomes also takes more time. This is a general trend and the even more frequent start codon “ATG” as anchor indeed requires the longest calculations. Other biologically meaningful motifs we tested were the Shine-Dalgarno and Kozak motifs. The latter is of course more specific for Eukaryotes and its low performance on the bacterial genomes is expected due to its measurable low occurrence. Surprisingly, simple repeat motifs like “AAAA” also showed reasonable correlation but again were more abundant and reduced speed. While the biological soundness of an anchor seems to help improving the performance, also many alternative motifs can be used in principle with similar performance. In order to reduce the number of oligos and increase the speed, we gradually added nucleotide restrictions to the well correlating

stop codon “TAG” motif and found that the simple “ATAG” motif is a good compromise for both accuracy and speed (Table 1, Figure 3A). The correlation is clearly motif dependent as exemplified by the comparison of “ATAG” (Figure 3A) and “CTAG” (Figure 3B) motifs. Using "ATAG" as a signature anchor tag, approximately 7500 matching 15-mer oligonucleotides were retrieved per genome instead of the 48 million from the complete sequence pool while a Pearson's R^2 correlation of 0.984 with the full oligo count was maintained (Figure 3A).

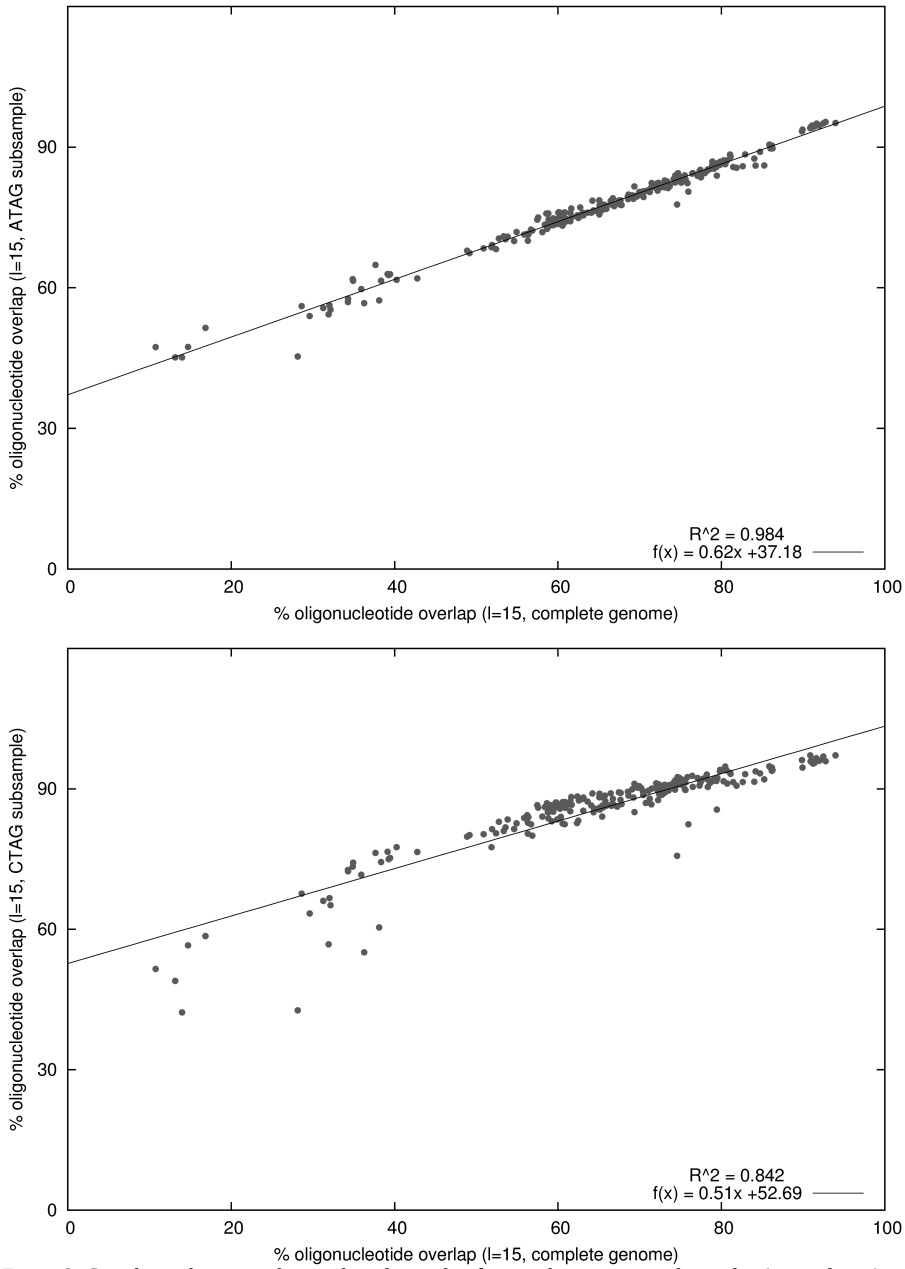

Figure 3. Correlation between oligonucleotide overlap fraction between tag subsamples (vertical axis) and whole genome (horizontal axis) is motif-dependent. (A)’ATAG’ subsample, (B) ‘CTAG’ subsample

3. Results

Inference of phylogeny

To show that the proposed simple genome similarity measure can indeed be used to infer phylogenetic groupings, all pairwise genome distances were calculated as described above for the 21 E.coli related strains from the previous detailed phylogenetic study, along with 4 available genomes from the new O104:H4 outbreak strain. Implemented in Perl, the fully automated all against-all comparison of the 25 complete genomes and production of a distance matrix in Nexus format took less than 1 minute on a standard desktop computer (Intel i7 @ 3.33Ghz) requiring only 26MB of RAM. As some genome assemblies could be the reverse complement relative to others, oligonucleotides for all genomes were also read in forward and reverse complement direction allowing for these oligonucleotides to be from either the plus or minus strand, although this effectively doubles the number of calculations. The running time for the earlier 25-genome example was then increased to about 2 minutes and the maximum memory needed is about 38MB. For comparison, the same calculation without using the novel "ATAG" tag motif filter took ~6 times longer and required ~11 times more RAM showing significant improvement of speed and memory efficiency with the tag approach.

Since closely related bacteria can represent complex mosaics of shared genomic regions, pairwise distances were transformed into a phylogenetic network which allows a richer representation of complex evolutionary scenarios compared to typical trees.[20] Using the Neighbor-Net algorithm within the SplitsTree 4 software (www.splitstree.org), the phylogenetic network (Figure 4) using the Nexus input matrix of distances was produced instantly (<1 second) and indeed closely resembled the tree as shown in the earlier detailed study.[5] The latter is based on TreePuzzle assemblies of maximum likelihood trees for ~2000 individually aligned gene families which have been identified by tedious orthologue assignments as representing the *E.coli* core gene repertoire.[5] Although our simplified approach took only around 1 minute starting from the unaligned genomes, a clear clustering of strains was observed within their correct phylogenetic groups A, B1, B2, D, E and S3 (groups are color-coded in Figure 4) and the relative positioning of the groups to each other also appeared accurate. The different *Shigella* strains are also embedded exactly as seen in the established tree. The more remotely related *E. fergusoni* was clearly distinguishable as an outlier and could be used to root the estimated phylogeny. Interestingly, while two members of group D were split in the alignment-based maximum likelihood tree,[5] our simple approach allowed for reasonably clustering of these group D strains together. Finally, the 4 genomes of the O104:H4 outbreak strain were observed to be grouped together, most closely related to strain *E.coli* 55989 within phylogenetic subgroup B1 which is in agreement with current knowledge about the outbreak strain.[2–4] Other genome-wide phylogeny studies that have been applied to the *E.coli* and *Shigella* families, including also alignment-free methods, arrive at similar conclusions with different levels of detail of the groupings.[21–24]

Rapid identification of reference genomes for read assemblies

A recent surge in next-generation sequencing methods has resulted in large gains in speed and cost savings for obtaining sequence reads. However, the accurate assembly of these reads requires the identification of a reference genome. The phylogenetic inference approach using 15-mers presented in this study could potentially be employed in rapid identification of the necessary reference genome for assembly. The characteristic differences of sequences in FASTQ read files compared to FASTA assembly files are the differing sequence lengths as well as the redundancy of the sequences in the short reads. Using the O104:H4 sequencing data provided by Pacific Biosystems (Corrected Reads in FASTQ format, http://www.pacbiodevnet.com/Share/Datasets/E-coli-Outbreak), it was observed that the phylogenetic signal to noise ratio improves with increasing the minimum read length to >100, >500, >2000 and >5000 bases (data not shown). However, most other next generation sequencing methods typically produce shorter read lengths (~100 bases), essentially eliminating this avenue of improving signal-to-noise ratio. As an alternative filter, redundancy among the reads was used, based on the idea that 15-mers occurring more often should be more reliable. Experimentation with varying parameters and thresholds suggested that an acceptable strategy for unassembled reads to become comparable with assembled contigs was to limit the used 15-mer oligonucleotides from the short reads to a similar total number per genome as used for the assembled data by only retaining those that appeared at least 5 times within the redundant short reads. This value obviously depends on the genome characteristics, the sequencing method, coverage and anchor tag employed, but it can be derived automatically using the rule of thumb of requiring a similar number of representative 15-mer oligonucleotides per genome. To further unbias from set size differences depending on tag-selection and sequencing method, the distance measure can be modified as such:

$$d = 1 - \frac{n_{overlap}}{n_{\min(set1,set2)}} \quad (2)$$

which serves to normalize by the set size of the smaller of the two compared oligonucleotide sets rather than their intersection. Indeed, after filtering the unassembled sequence reads from the Pacific Biosystems platform as well as Ion Torrent reads made available by the Beijing Genome Institute (BGI, ftp://ftp.genomics.org.cn/pub/Ecoli_TY-2482), the O104:H4 strains appear in their expected phylogenetic groups (Figure 4). Therefore, the broad usefulness of this approach is shown by the fact that heterogeneous input data such as long contigs from whole genome assemblies can be analysed interchangeably with short read files directly from different sequencing platforms in order to produce reasonable phylogenetic groupings.

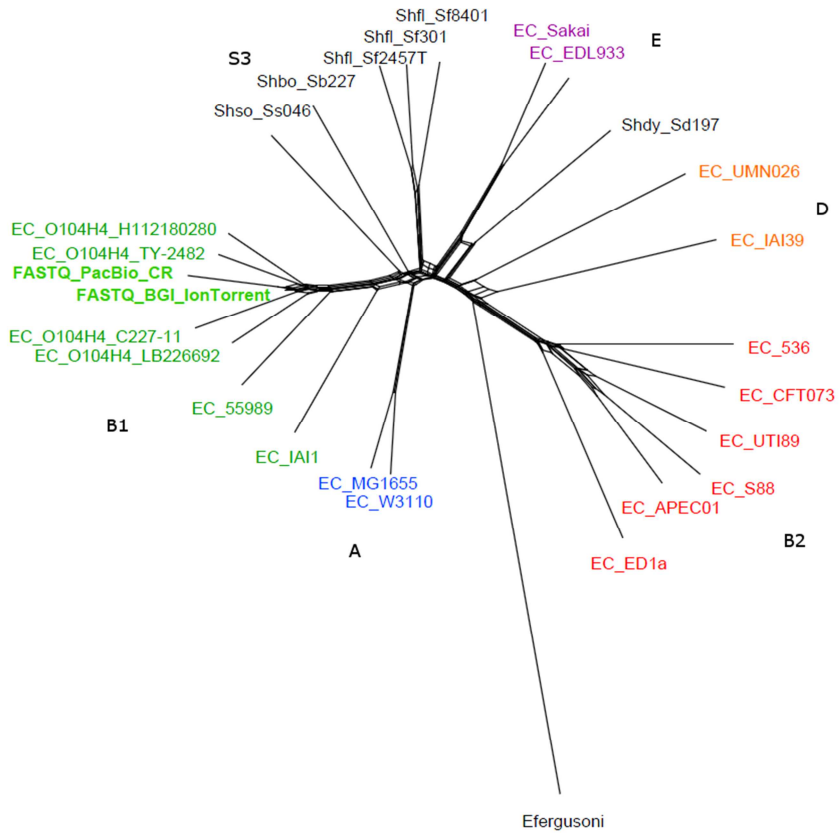

Figure 4. Phylogenetic network showing relationship between 25 complete *Escherichia spp.* genomes based on pairwise genome distances calculated as described in the text. Neighbour-Net algorithm within SplitsTree 4 software package was used to produce the tree. Phylogenetic groups are colour coded: A (blue), B1 (green), B2 (red), D (orange), E (purple), S3 (black).

4. Discussion and Conclusion

While the produced phylogenetic network appears to reflect the currently known *E.coli* genome phylogeny well, the approach needs to be tested and confirmed further on different datasets with varying parameters such as taxonomic ranges and genome sizes. Initial tests on mitochondrial genomes of mammals proved promising (future work). Realistically, the enormous speed increase through simplified genome and distance representation must come with a drop in accuracy at some levels of detail. We can see this by further reducing the number of needed calculations through longer, more restrictive, anchor tags which increase the speed but can result in too few oligonucleotides to calculate overlaps (e.g. <500) which seems to produce a gradual loss of the phylogenetic signal and mixing of the phylogenetic groups. Consequently, we do not believe that this approach will generally work well for single gene families or limited subsets of genes instead of whole genomes. In any case, it is imperative to select the proper oligonucleotide length, which, as was shown here, can also be done unsupervised and alignment-free by determining the length of maximum overlap between pairs. Additionally, reliability of tree or network structure could be estimated through bootstrapping. Furthermore, this approach in its fastest

implementation does not consider differing evolutionary rates, molecular clock or explicit substitution models. Last but not least, it is a distance-based method which is expected to be fast but for some scenarios, maximum likelihood or Bayesian approaches may be preferred.

To conclude, the speed and reasonable accuracy of deriving the phylogenetic network for the E.coli strains shows that the proposed unsupervised alignment-free approach appears suitable for fast initial assessment of phylogenetic groupings of newly sequenced genomes, as exemplified by the correct B1 subgroup assignment of the enterohemorrhagic O104:H4 E.coli strain. The measured 6-fold speed-up and 11-fold memory requirement reduction compared to previous alignment-free methods is achieved by reducing the comparison space to a representative sample of k-mers of optimal length and with specific tag motifs. It was further shown that unassembled sequence reads can be used directly as input for our method. This simplified approach might prove to be an especially useful addition to existing detailed phylogenetic methods as pre-screen tool in the light of the fast and cheap availability of complete genome data through next-generation sequencing.

5. References

1. Frank, C. *et al*, Large and ongoing outbreak of haemolytic uraemic syndrome, Germany, May 2011. *Euro Surveill* **16,** 19878, 2011.
2. Mellmann, A. *et al*, Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. *PLoS ONE* **6,** e22751, 2011.
3. Rasko, D. A. *et al*, Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N. Engl. J. Med.* **365,** 709–717, 2011.
4. Rohde, H. *et al*, Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. *N. Engl. J. Med.* **365,** 718–724, 2011.
5. Touchon, M. *et al*, Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet.* **5,** e1000344, 2009.
6. Goldstein, P. Z. & DeSalle, R, Integrating DNA barcode data and taxonomic practice: determination, discovery, and description. *Bioessays* **33,** 135–147, 2011.
7. Lippert, R. A., Huang, H. & Waterman, M. S, Distributional regimes for the number of k-word matches between two random sequences. *Proc. Natl. Acad. Sci. U.S.A.* **99,** 13980–13989, 2002.
8. Forêt, S., Wilson, S. R. & Burden, C. J, Characterizing the D2 statistic: word matches in biological sequences. *Stat Appl Genet Mol Biol* **8,** Article 43, 2009.
9. Reinert, G., Chew, D., Sun, F. & Waterman, M. S, Alignment-free sequence comparison (I): statistics and power. *J. Comput. Biol.* **16,** 1615–1634, 2009.
10. Wan, L., Reinert, G., Sun, F. & Waterman, M. S, Alignment-free sequence comparison (II): theoretical power of comparison statistics. *J. Comput. Biol.* **17,** 1467–1490, 2010.
11. Tan, J. *et al.* Tachyon search speeds up retrieval of similar sequences by several orders of magnitude. *Bioinformatics* **28,** 1645–1646, 2012.
12. Katoh, K. & Toh, H, PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences. *Bioinformatics* **23,** 372–374, 2007.
13. Sims, G. E., Jun, S.-R., Wu, G. A. & Kim, S.-H, Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 2677–2682, 2009.
14. Jun, S.-R., Sims, G. E., Wu, G. A. & Kim, S.-H, Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 133–138, 2010.
15. Vinga, S. & Almeida, J, Alignment-free sequence comparison-a review. *Bioinformatics* **19,** 513–523, 2003.

16. Forêt, S., Kantorovitz, M. R. & Burden, C. J, Asymptotic behaviour and optimal word size for exact and approximate word matches between random sequences. *BMC Bioinformatics* **7 Suppl 5,** S21, 2006.
17. Sims, G. E., Jun, S.-R., Wu, G. A. & Kim, S.-H, Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 2677–2682, 2009.
18. Urwin, R. & Maiden, M. C. J, Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol.* **11,** 479–487, 2003.
19. Osada, Y., Saito, R. & Tomita, M, Comparative analysis of base correlations in 5' untranslated regions of various species. *Gene* **375,** 80–86, 2006.
20. Huson, D. H. & Bryant, D, Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23,** 254–267, 2006.
21. Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W, Comparison of 61 sequenced Escherichia coli genomes. *Microb. Ecol.* **60,** 708–720, 2010.
22. Sims, G. E. & Kim, S.-H, Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs). *Proc. Natl. Acad. Sci. U.S.A.* **108,** 8329–8334, 2011.
23. Cheung, M. K., Li, L., Nong, W. & Kwan, H. S, 2011 German Escherichia coli O104:H4 outbreak: whole-genome phylogeny without alignment. *BMC Res Notes* **4,** 533, 2011.
24. Zhang, Y. & Lin, K, A phylogenomic analysis of Escherichia coli / Shigella group: implications of genomic features associated with pathogenicity and ecological adaptation. *BMC Evol. Biol.* **12,** 174, 2012.