# Filtering with alignment free distances for high throughput DNA reads assembly

**Maria C De Cola**[1,2]✉, **Giovanni Felici**[2], **Daniele Santoni**[2], **Emanuel Weitschek**[2,3]

[1]Department of Statistics, University La Sapienza, Rome, Italy
[2]Institute of Systems Analysis and Computer Science, National Research Council, Rome, Italy
[3]Department of Informatics and Automation, Università degli Studi Roma Tre, Rome, Italy

## Motivation and Objectives

The output of a high throughput next generation sequencing (NGS) machine is a collection of short reads, which have to be properly assembled in order to reconstruct the original DNA sequence of the analyzed organism (Metzker, 2010; Earl, 2011). The DNA sequence assembly process is based on aligning and merging these reads for effectively reconstructing the real primary structure of the DNA sample sequence or reference genome. The use of NGS machines results in much larger sets of reads to be assembled, posing new problems for computer scientists and bioinformaticians. In particular, a relevant issue is related with the trade-off between precision of the assembly process and its computational time, stating the need for faster methods that can keep pace with the speed and volume of reads that are generated with NGS. An important step in DNA assembly is the identification of a subset of read pairs that have a high probability of being aligned sequentially in the reconstruction. Such step is often referred to as filtering, and amounts in selecting a significantly smaller subset of the initial set of read pairs (whose dimension is quadratic in the number of initial reads) that can be then processed by an alignment algorithm, usually quite time consuming. The desired effect of filtering is then to quickly filter out from the candidate set of read pairs those that would not provide a good alignment in the following phase. The computation cost of filtering should then be balanced by the speed-up obtained when a smaller set of read pairs is considered for alignment.

In this work we propose and test the use of alignment free distances to evaluate the similarity between two short reads as a technique for filtering good read pairs to be assembled.

The method operates in constant time in the string length and is tested in its ability to emulate, with a proper level of precision, much more time consuming methods to evaluate the similarity between short DNA sequences, such as the established Needleman-Wunsch edit distance (Needleman, Wunsch, and Christian, 1970), often used in the final step of the assembly procedure. These preliminary experiments show the efficacy of this approach for filtering the promising read pairs - eligible candidates to successfully assemble the entire genome of a given organism. Therefore, the alignment free reads filtering may significantly accelerate the assembly process without a substantial loss in accuracy of the DNA sample sequence reconstruction.

## Methods

### ODNA sequence assembly

The DNA sequence assembly process is based on the alignment and merging of reads (stretch of sequences) in order to reconstruct the original primary structure of the DNA sample sequences. Given a set of sequences $S = \{S_1, S_2, \ldots, s_n\}$, where $s \in S$ is a fragment of the primary structure of DNA (read) (e.g. $s = \{ATTCGA...CTGACT\}$), assembly is in charge of building the longest sequence from the set S where each pair of consequent reads obey certain similarity conditions.

### DNA read pairs filtering and Alignment Free Distance

This step identifies the promising read pairs in order to reduce the amount of input data given to the real assembly algorithm. We adopted a very quick measure of the similarity between two reads, Alignment Free (AF) based distance (Vinga and Almeida, 2003). AF computes the similarity of two strings based only on the dictionary of their substrings, irrespective of their relative position. As a dictionary we considered the set of 4-mer (sequences composed of 4 different nucleotides) and then built a profile for each read composed by the relative frequencies of each 4-mer in the read. The Euclidean distance between the profiles of two reads was taken as

an inverse measure of the similarity of the two reads and thus as an indication that the two reads formed a promising pair to be considered in the assembly phase. AF filtering was then used defining a proper threshold on the AF distance and discarding all the pairs that exhibited a AF distance above the threshold. Computational complexity of AF distance is a constant linearly bounded by the number of k-mers adopted and the length of the strings to be compared.

**Comparing with other distances: Needleman-Wunsch and "Bowtie" distance**

Along with AF we considered the well-established Needleman-Wunsch edit distance (NW) and compared them in their ability to identify significant pairs. This comparison was based on the computation of a sort of perfect distance computed after an alignment over an already known sequence has been performed. Such distance, referred to as Bowtie distance (BT), was obtained as follows:

a. a large number of reads coming from a known sequence were considered;
b. hese reads were aligned over the known sequence using the standard Bowtie algorithm (Langmead et al, 2009);
c. any two reads received a maximum BT distance if their alignment did not intersect over the reference sequence, else they received a distance inversely proportional to their intersection over the sequence (e.g., they would have BT distance equal to 0 if they were aligned one on top (or inside) of the other by the Bowtie algorithm). By construction we assumed BT distance to be the reference distance, e.g., the distance that expressed the best possible alignments - being based on the knowledge of the reference sequence - and tested the correlation of AF and NW with BT; moreover, we verified the ability of AF and NW to predict that a given read pair had BT distance above or below a given threshold.

## Results and Discussion

For our test we considered the E.Coli genome and a set of reads from this genome reads obtained by Roche 454 sequencing machine. Reads have average length of ~235 nucleotides and standard deviation of approx. 10 (the large majority of them having length in the interval 225-245). Reads were aligned with the reference sequence with Bowtie and then 100,000 were sampled at random according to their alignment along the sequence. Reads were considered both forward and reversed, giving rise to a total of 200,000[2] read pairs. All 620,798 read pairs with BT distance < 1 were considered for the experiments; then, out of the remaining pairs, 233,099 were sampled at random. A total of 853,897 read pairs composed the working data set. For all these reads, NW distance and AF distance over the 4-mer were computed. AF, NW and BT distances were all normalized between 0 (maximal similarity) and 1 (maximal dissimilarity). The first interesting results was that the correla-

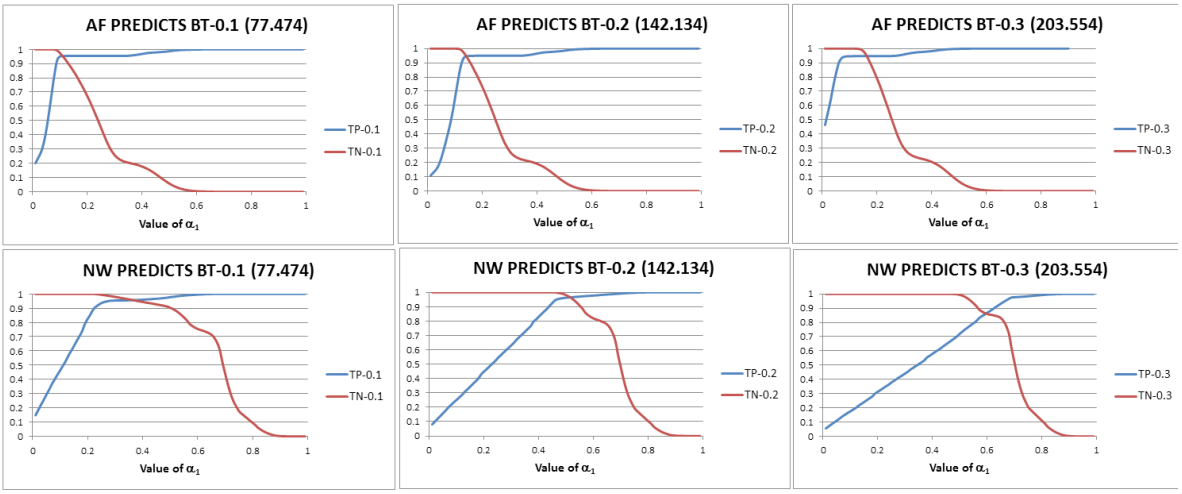


Figure 1. Error curves for predictors of BT. Error rates of threshold predictors for BT based on AF are plotted in the charts of the first row; predictors for BT based on AF are in the second row; blue lines represent True positive rates, red lines represent true negative rates.robustness can also apply to dynamic processes in development.

tion between distances showed that AF approximates BT somehow better than NW: we obtained a correlation coefficient of 0.761 for AF and BT, compared with a smaller 0.706 when NW and BT were considered (coherently, correlation between AF and NW is 0,721). The second interesting results was obtained when we compared the ability of AF and NW to predict whether BT was above or below a given threshold. We defined a threshold predictor for a given function $F_2$ based on function $F_1$ and on a given pair $\alpha_1$, $\alpha_2$ as follows: if ($F_1 < \alpha_1$) then predict ($F_2 < \alpha_2$), else predict ($F_2 \geq \alpha_2$). To a given pair ($\alpha_1$, $\alpha_2$), we associated the measure of True Positive rate (TP) (percentage of cases where ($F_1 < \alpha_1$) and ($F_2 < \alpha_2$) and of True Negative rate (TN) (percentage of cases where ($F_1 \geq \alpha_1$) and ($F_2 \geq \alpha_2$); analogously we defined False Positive rate (FP) and False Negative rate (FN).

For each ($\alpha_1$, $\alpha_2$) with both values ranging from 0 to 1, we then computed, with step 0.05, the positive and negative error rates taking AF as a predictor of BT and NW as a predictor of BT. Part of the results are summarized in the charts of Figure 1, that show for 3 different levels of $\alpha_2$ (0.1, 0.2, and 0.3) the precision of the predictors (y-axis) when the value of $\alpha_1$ is changed (x-axis), both when AF is used as a predictor of BT (charts in the first row) and when NW is used as a predictor of BT (charts in second row). Similar results are obtained also for other levels of $\alpha_2$, here omitted for brevity. The curves bring to light very clearly how AF is a very good threshold predictor for BT for the considered data; despite its light computational complexity, it appears to perform significantly better than the more complex NW edit distance when its ability to support a threshold predictor is considered.

## Acknowledgements

The authors are partially supported by the FLAGSHIP "InterOmics" project (PB.P05) funded by the Italian MIUR and CNR institutions, and by the cooperative programme 2010–2012 between the National Research Council of Italy (CNR) and the Polish Academy of Sciences (PAN).

## References

1. Earl D et al (2011); Assemblathon 1: A competitive assessment of de novo short read assembly methods; Genome Research, 21
2. Langmead B, Trapnell C, et al (2009); Ultrafast and memory-efficient alignment of short DNA sequences to the human genome; Genome Biology, 10:R25
3. Metzker ML (2010); Sequencing technologies — the next generation; Nat Rev Genet., 11(1)
4. Needleman SB, Wunsch CD. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48(3): 443–53
5. Vinga S, Almeida J (2003); Alignment-free sequence comparison—a review, Bioinformatics 19 (4), 513-523