*Bioinformatics*, 37(10), 2021, 1468–1470
doi: 10.1093/bioinformatics/btaa853
Advance Access Publication Date: 25 November 2020
Applications Note

Sequence analysis

# ADACT: a tool for analysing (dis)similarity among nucleotide and protein sequences using minimal and relative absent words

**Mujtahid Akon**[1], **Muntashir Akon**[2], **Mohimenul Kabir**[1], **M. Saifur Rahman**[1] and **M. Sohel Rahman**[1,*]

[1]Department of CSE, BUET, Dhaka, Bangladesh and [2]Department of CSE, RUET, Rajshahi, Bangladesh

*To whom correspondence should be addressed.

Associate Editor: Cowen Lenore

Received on May 21, 2019; revised on September 9, 2020; editorial decision on September 19, 2020; accepted on September 21, 2020

## Abstract

**Motivation:** Researchers and practitioners use a number of popular sequence comparison tools that use many alignment-based techniques. Due to high time and space complexity and length-related restrictions, researchers often seek alignment-free tools. Recently, some interesting ideas, namely, Minimal Absent Words (MAW) and Relative Absent Words (RAW), have received much interest among the scientific community as distance measures that can give us alignment-free alternatives. This drives us to structure a framework for analysing biological sequences in an alignment-free manner.

**Results:** In this application note, we present Alignment-free Dissimilarity Analysis & Comparison Tool (ADACT), a simple web-based tool that computes the analogy among sequences using a varied number of indexes through the distance matrix, species relation list and phylogenetic tree. This tool basically combines absent word (MAW or RAW) computation, dissimilarity measures, species relationship and thus brings all required software in one platform for the ease of researchers and practitioners alike in the field of bioinformatics. We have also developed a restful API.

**Availability and implementation:** ADACT has been hosted at http://research.buet.ac.bd/ADACT/.

**Contact:** msrahman@cse.buet.ac.bd

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

## 1 Introduction

The sequence comparison problem is one of the most common classical problems with extensive applications in different branches of science and engineering. Traditionally, the concept of pairwise or multiple sequence alignment has got significant attention in the context of sequence comparison. However, this can be inappropriate, inaccurate and infeasible (Zielezinski *et al.*, 2019). High computational resource requirement makes the use of alignment-based approaches limited (even for pairwise alignments), especially for large-scale sequence data. Alignment-free sequence analysis methods provide alternatives over alignment-based approaches. An alignment-free technique uses information from the sequences rather than from the alignment thereof. One interesting concept related to such information is the *absent word*.

Formally, an absent word of a string $x$ is defined as a string $y$ that is not a substring of $x$. The number of absent words is exponential in the sequence length. So, it is meaningful to consider a subset of absent words having cardinality linear in the sequence length. String $y$ is a minimal absent word (MAW) of $x$ if and only if all of its ($y$) proper factors occur in $x$. The MAW has the capability to extract necessary information from sequences and achieve the target of sequence comparison. This is why researchers have proposed distance measures based on MAW (Chairungsee and Crochemore, 2012, Garcia and Pinho, 2011, Yang *et al.*, 2013). A number of distance measures have been studied and analysed by Rahman *et al.* (2016) for possible use as (dis)similarity measures using MAW. These indexes include Length weighted index (LWI), GC-content (GCC), Total Variation Distance and Jaccard Distance.

Relative Absent Word (RAW) is another related but relatively new idea introduced by Silva *et al.* (2015) for differential identification of sequences that are derived from a pathogen genome (i.e. EBOLA virus) but absent from its host. Relative absent words of a string $x$ with respect to a string $y$ are the minimal substrings that are not found in $x$ but are present in $y$. Rahman *et al.* (2016) considered RAW as well for LWI and GCC.

A phylogenetic tree is a diagram that expresses evolutionary relationships among organisms. Two popular algorithms for the

**Table 1.** Results of two AFproject datasets tests

| Dataset | No. of species | AFproject Rank (no. of ranks) | Time (ADACT) (s) | Time (Bowtie2) (s) |
|---|---|---|---|---|
| Fish mtDNA | 25 | 2 (9) | 18 | 579 |
| E.coli strains | 29 | 12 (18) | 9120 | – |

*Note*: '–' denotes timeout or memory exhausted event.

**Table 2.** Results of several example tests

| No. | Dataset | No. of species | Total size KBytes | Time (ADACT) (s) | Time (Bowtie2) (s) |
|---|---|---|---|---|---|
| 1 | ClassA Peptide Duffy | 65 | 26 | 2 | 260 |
| 2 | ClassA Nucleotide Adenosine | 93 | 38 | 2 | 515 |
| 3 | ClassA Peptide Neuromedin | 60 | 30 | 1 | 220 |
| 4 | ClassA Peptide Chemokine | 575 | 259 | 93 | – |
| 5 | GIN numbers in example in WiKi page | 6 | 100 | 12 | 27 |

*Note*: '–' denotes timeout or memory exhausted event.

phylogenetic tree construction are unweighted pair group method with arithmetic mean (UPGMA) by Sung (2009) and neighbour joining (NJ) by Saitou and Nei (1987).

In this application note, we are incorporating absent word computation with the calculation of (dis)similarity. We are presenting Alignment-free Dissimilarity Analysis & Comparison Tool (ADACT), an alignment-free comparison tool that computes (dis)similarity among biological sequences via a number of indexes and outputs the results through a distance matrix, species relation list and phylogenetic tree. We have also facilitated a restful API.

## 2 ADACT features

ADACT supports both nucleotide and protein sequences. It supports almost all popular gene, genome and protein databases. A user can choose either MAW or RAW for the alignment-free comparison. ADACT provides a limited storage to the users for saving the results of previous runs.

**Input:** ADACT accepts inputs in several ways, i.e. either in zip format or by entering the accession number or GI number. Users can also give input sequences by hand-typing without a file. A zip file may contain either a single FASTA-formatted file having multiple sequences or multiple FASTA files each containing a single sequence.

**Configuration parameters:** To specify a project, a user should define some configuration parameters before running the project. These include: short names, Absent word types (MAW/RAW), *k*-mer size, Reverse complement, Sequence types (Nucleotide/Protein), Dissimilarity index (according to Rahman *et al.*, 2016).

**Backend processes:** According to the input sequences and configuration parameters, ADACT computes MAW (or RAW). Then, ADACT calculates distance measures and species relations according to the definition of Rahman *et al.* (2016). After finishing one's run (project), ADACT sends an email to the user for notifying the status of the project.

**API:** We have developed a restful API (documentation available at https://github.com/mujtahid-akon/ADACT/wiki). We have developed a number of endpoints by which one can perform all functionalities of ADACT.

## 3 Experiments

We have conducted rigorous experiments to examine and analyse the usability of ADACT. Here, we briefly present the results of two important experiments (A and B). In Experiment A, we tested assembled benchmark datasets collected from AFproject (Zielezinski *et al.*, 2019) under the genome-based phylogeny category (Table 1). The rank reported by AFproject for each of the tests is reported in the table. Informatively, AFproject is a community developed free service that provides performance comparison of alignment-free sequence comparison tools on different datasets (for details, please visit http://afproject.org). In Experiment B, we compared ADACT with Bowtie 2 using protein sequences from GDS dataset (Davies *et al.*, 2007). Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences (Langmead and Salzberg, 2012). Since Bowtie 2 is a state-of-the-art aligner, it makes sense to compare the execution time thereof with ADACT on an equal footing. Table 2 reports the execution times of ADACT and Bowtie 2. We have run Bowtie 2 using its default option (see Supplementary Section S9 of Supplementary Material for details).

## 4 Conclusion

String comparison has massive applications from our routine text editor to complex biological compounds. Our alignment-free comparison tool, ADACT can efficiently analyze (dis)similarity among nucleotide and protein sequences. ADACT is user friendly; its overall UI design is simple and eye-soothing. For smooth operation, ADACT imposes some usage restrictions considering the available resources (see Supplementary Material for details). On an ending note, ADACT provides an easy-to-use portal to explore alignment-free dissimilarity measures in sequences at nucleotide and protein level, combining absent and relative absent word based methods towards fast phylogeny reconstruction.

## Data availability

All code and data can be found from: https://github.com/mujtahid-akon/ADACT/

*Financial Support*: Partially supported by an innovation fund of ICT Division, Government of the Peoples' Republic of Bangladesh.

*Conflict of Interest*: none declared.

## References

Chairungsee,S. and Crochemore,M. (2012) Using minimal absent words to build phylogeny. *Theor. Comput. Sci.*, **450**, 109–116.

Davies,M.N. *et al.* (2007) On the hierarchical classification of G protein-coupled receptors. *Bioinformatics*, **23**, 3113–3118.

Garcia,S.P. and Pinho,A.J. (2011) Minimal absent words in four human genome assemblies. *PLoS One*, **6**, e29344.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Rahman,M.S. *et al.* (2016) Absent words and the (dis) similarity analysis of DNA sequences: an experimental study. *BMC Res. Notes*, **9**, 186.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Silva,R.M. *et al.* (2015) Three minimal sequences found in Ebola virus genomes and absent from human DNA. *Bioinformatics*, **31**, 2421–2425.

Sung,W.-K. (2009) *Algorithms in Bioinformatics: A Practical Introduction. Chapman & Hall/CRC Mathematical and Computational Biology*. CRC Press, Boca Raton, Florida.

Yang,L. *et al.* (2013) Large local analysis of the unaligned genome and its application. *J. Comput. Biol.*, **20**, 19–29.

Zielezinski,A. *et al.* (2019) Benchmarking of alignment-free sequence comparison methods. *Genome Biol.*, **20**, 144.