VRIJE UNIVERSITEIT BRUSSEL

UNIVERSITÉ LIBRE DE BRUXELLES

Review of 'Comparing sequences without using alignments: application to HIV/SIV subtyping'

Supervised by Prof. Mathieu Defrance (*Machine Learning Group* - *ULB*)
& Prof. Wim Vranken (*Bioengineering Sciences Department* - *VUB*)

Master's in Computer Science

Tomás GUIJA VALIENTE
Benjamin OBERTHÜR

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Contents

Notations

Notation	Meaning
$[\![\cdot, \cdot]\!] : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$	Integer interval, $[\![a, b]\!] = \{a, a+1, \dots, b\}$

I. Introduction

This paper describes the problem in sequence comparison of the fact that when we want to compare a large number of sequences together, some types of sequence alteration like insertion or deletion are poorly - if ever - handled by classic sequence alignment methods. Intuitive ideas to solve this problem without alignment would be to look at the nucleotids or amino acids frequency, but this methods is not really meaningful, as sequences with similar frequencies can be a lot different. A more sophisticated - and more working - is dealing with what [Did+07] calls N-words (in nowaday's litterature, we call them k-mers). With those, we can compute dissimilarities between sequences [KL94] which can help us show evolutionary relationships between sequences.

I.1 Local decoding method of order N

The first step of this analysis presented by [Did+07] is the local decoding method of order N, or N-local decoding.

Definition I..1 (N-words)**.** In a sequence, a N-word is a **contiguous sub-sequence of size** N of the given sequence. The set of its N-words is the its sub-sequences of size N.

Example. The set of the 3-words of the sequence $AGTACGT$ is AGT, GTA, TAC, ACG, CGT.

Let $S = S_1 S_2 \dots S_i \dots S_{|S|}$ be a sequence, i a site (or index) of S. For a given $N \in \mathbb{N}^*$, we consider the set of N-words of S covering the site i.

Definition I..2 (Direct relation)**.** Two sites are said **directly related** if they have the same position in two (or more) occurences of the same N-word.

Example. (We take the example on Figure 6a in [Did+07]). Let seq1 = CATTG TCCGC **T**GGAC CACAC and seq2 = CACT**T** GGACA CATAC CATGC. We consider the site 11 in seq1 and the site 5 in seq2 (bolded in their definitions), and look at the 5-words covering this site (contained in the sites colored in red).

C	C	G	C	**T**	G	G	A	C
C	C	G	C	T				
	C	G	C	T	G			
		G	C	T	G	G		
			C	T	G	G	A	
				T	**G**	**G**	**A**	**C**

C	A	C	T	**T**	G	G	A	C
C	A	C	T	T				
	A	C	T	T	G			
		C	T	T	G	G		
			T	T	G	G	A	
				T	**G**	**G**	**A**	**C**

The 5-word TGGAC appears in both these sequences, and the site 11 in seq1 and 5 in seq2 are both in first position of the 5-word, so these two sites are **directly related**

Definition I..3 (Transitivity)**.** Let $\mathcal{R}$ be a binary relation. $\mathcal{R}$ is said to be **transitive** if it respects the following property [Cau21]:

$$a\mathcal{R}b \wedge b\mathcal{R}c \implies a\mathcal{R}c$$

Definition I..4 (Transitive closure)**.** Let $\mathcal{R}$ and $\mathcal{R}'$ be binary relations. $\mathcal{R}'$ is the **transitive closure** of $\mathcal{R}$ if [Sch77]:

$$\forall a, b;\ a\mathcal{R}b \implies a\mathcal{R}'b$$

and

$$\forall a, b, c;\ a\mathcal{R}b \wedge b\mathcal{R}c \implies a\mathcal{R}'c$$

Definition I..5. We define the (simple) relation between two sites as the transitive closure of the direct relation. Therfore, we say that two sites are related of there is a (finite) chain of direct relations linking those sites

We can divide those sites in a partition of relation classes[1]. [Did+07] calls them N-classes.

Definition I..6 (N-classes)**.** An N-class can be defined as follows (with a being the identifier of the N-class):

$$C(a) = \{x \in [\![1, |S|]\!];\ x \text{ is related with } a\}$$

[1]The proof comes from the fact that the relation we described is an equivalence relation, i.e. it is reflexive, symmetric, and transitive

II. Material & Method

II.1 Data Sets

We used two different sequence sets to run our experiments: one given from Gilles Didier, co-author of [Did+07], and another that we retrieved ourselves, both coming from the data base maintened by the 'Los Alamos National Laboratory'[2]. The first set contains 66 sequences writen prior to 2007, and the second one has been created from the query to have 47 sequences sampled in 2016 in Germany (so that we look at mutations in a restrained place and time).

III. Experiments

IV. Results

V. Conclusion

[2]https://www.hiv.lanl.gov

References

[Cau21] Augustin-Louis Cauchy. *Cours d'analyse de l'École royale polytechnique. Analyse algébrique.* .1ère partie. Imprimerie royale, 1821.

[Sch77] Ernst Schröder. *Der Operationskreis des Logikkalkiils.* Ed. by Teubner-Verlag. Leipzig, 1877.

[KL94] Samuel Karlin and Istvan Ladunga. "Comparisons of eukaryotic genomic sequences." In: *Proceedings of the National Academy of Sciences* 91.26 (1994), pp. 12832–12836.

[Did+07] Gilles Didier et al. "Comparing sequences without using alignments: application to HIV/SIV subtyping". In: *BMC Bioinformatics* 8.1 (Jan. 2007), p. 1. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-1. URL: https://doi.org/10.1186/1471-2105-8-1.