УДК 573.22 + 51-76

Combinatorial Properties of Restriction Sites

Elizaveta A. Taranenko*
Institute of Fundamental Biology and Biotechnology
Siberian Federal University
Svobodny, 79, Krasnoyarsk, 660041
Russia

Received 10.03.2018, received in revised form 20.04.2018, accepted 30.05.2018

Restriction sites are the strings in DNA recognised by their specific enzymes. Recognition of the sites is provided by their combinatoric pattern. This study presents some results obtained from 6-nucleotide long restriction sites analysis using duplet strength and editing distance methods.

Keywords: frequency, palindrome, string, editing distance, duplet strength.
DOI: 10.17516/1997-1397-2018-11-3-329-330.

Introduction

Restriction enzymes (REs) are chemical compounds structurally presented as dimers and activated through binding magnesium ion to the active site. Restriction sites (RSs) are four to twenty nucleotides long sequences in DNA that are recognised specifically by their respective REs. The restriction-modification complex was first discovered in bacteria and act as immune system against other bacteria or bacteriophages. The palindromic nature of the sites allows recognition despite direction of the enzyme, or strand. This work was based on the duplet' strength method and editing distance (ED) algorithms such as Hemming distance (dH), Levenshtein distance (dL) and Damerau-Levenshtein distance (dDL) [1]. The restriction endonucleases and their specific sites were obtained from REBASE database (http://rebase.neb.com/rebase/rebase.html) provided by New English Biolabs. The idea is to determine combinatoric properties of the sequences representing restriction sites, as well as track down evolutionary modified restriction sites to those original sequences.

1. Methods

The duplet method is based on the degree of nucleotides $d(\mathsf{N})$ with Cytosine $d(\mathsf{C})$ having the highest (= 4), $d(\mathsf{G}) = 3$, $d(\mathsf{T/U}) = 2$, and $d(\mathsf{A}) = 1$. Let $V^T = [C^{(4)}\ G^{(3)}\ T/U^{(2)}\ A^{(2)}]$ be the vector ordered descending; then Cartesian product yields matrix which elements are the sum of constituent the degrees $d(\mathsf{AC}) = d(\mathsf{A}) + d(\mathsf{C})$ [2]. The duplet strength figures are shown in Fig. 1. It fails to evaluate the six-nucleotide RSs reliably, thus editing distance (ED) method was applied [1]. Pairing of RSs was done using phylogenetic analysis of the REs of respective RSs. Four phylogenetic trees were constructed using two multiple sequence alignment programs (MUSCLE and CLUSTALW) and two

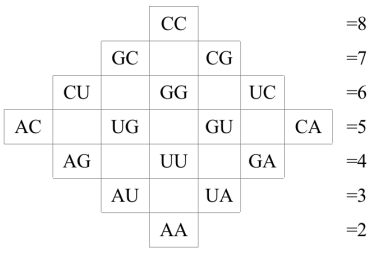

Fig. 1. Duplet strength draft

*taranenko.el@gmail.com

clustering algorithms (Maximum Likelihood and Minimal Evolution). For ED analysis substitution rates were calculated according to bacterial substitutional data derived from ten genomes for *E.coli*. The substitutions of the same nucleotides but in opposite directions ($A \Rightarrow T$ vs. $T \Rightarrow A$) gave different scores thus it was possible to determine the probable direction of the conversion, if the conversion took place at all.

2. Results and discussion

103 REs and their specific palindromic restriction sites were taken into account with 55 prototypes and 25 REs with 19 prototypes specific to non-palindromic sites. Accuracy of the method decreased since it could not resolve six-nucleotide sequences providing identical values for different sites.

Phylogenetic data provided us with up to 19 possible pairings of the most related REs for each of four applied phylogenetic pipelines. Respective RSs were used to calculate three types of editing distances (see Tab. 1). Default values for editing distances were out of the question

Table 1. Comparison of ED of two RSs determined in opposite directions

Enzyme 1	Enzyme 2	R. site 1	R. site 2	dH	dL	dDL
*Ase*I	*Bsp*HI	ATTAAT	TCATGA	6,171	4,962	5,059
*Bsp*HI	*Ase*I	TCATGA	ATTAAT	6,042	5,091	4,928

thus ten genomes for *E.coli* were used to obtain appropriate substitution rates. Obtained data correlates with available and published data on human haemoglobin substitutional rates. New substitutional rates that bear different rates for same nucleotide substitutions of opposite direction ($A \Rightarrow T$ vs. $T \Rightarrow A$) allow to calculate minimal editing distance in alternative ways of modification, thus allow to determine possible direction of the conversion. Study of RSs revealed that the most important part of the RS is central duplet while least important are terminal nucleotides. Palindromic sequences comprised of many weak nucleotides often do not act as RSs. Sequences with weak terminal nucleotides are poorly represented. However, sequences with weak nucleotides in the central duplet are abundant.

This study was funded by the research grant No. 14.Y26.31.0004 from the Government of the Russian Federation.

References

[1] V.I.Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, *Soviet Physics Doklady*, **10**(1966), no. 8, 707–710.

[2] D.Duplij, S.Duplij, Determinative degree and nucleotide content of DNA strands, *Biophysical Bull.*, Kharkov Univ., **497**(2000), 1–7.

Комбинаторные свойства сайтов рестрикции

Елизавета А. Тараненко
Сибирский федеральный университет
Свободный, 79, Красноярск, 660041
Россия

Рассмотрены комбинаторные свойства сайтов рестрикции, обеспечивающие механизм узнавания этих сайтов специальными ферментами.

Ключевые слова: частота, палиндром, цепочка, редакционное расстояние, сила дуплета.