

Сравнительный анализ алгоритмов, вычисляющих расстояния последовательностей ДНК и некоторые связанные проблемы

Авторы: Б.Ф. Мельников, С.В.
Пивнева,
М.А. Трифонов

Задача сравнения схожести строковых последовательностей ДНК

ДНК№

ATCGCGTCGAAACGCGCGTCGAACGCGCGTCGAACGCGTCGAA

....

ДНК№

ATCGCGTCGAAACGCGCGTCGAACGCGCGTCGAACGCGTCGAA

....

- За последние годы были описаны различные подходы к определению схожести последовательностей ДНК. Каждый из таких подходов определяет множество значений которое необходимо нуждается в качественной оценке.....

Качественная оценка алгоритмов

- В настоящей статье предлагается новый подход к решению этой задачи, причём алгоритмы для его реализации выполнены на основе ранее разработанного нами мультиэвристического подхода к задачам дискретной оптимизации. Однако основным предметом данной статьи является описание нашего оригинального подхода к сравнению качества определяемых метрик на множестве последовательностей ДНК. Последний подход основан на том, что тройки расстояний между геномами в идеале должны образовывать равнобедренные остроугольные **треугольники**.

Алгоритмы для качественного сравнения

- 1) Мультиэвритический алгоритм
- 2) Расстояние Джаро-Винклера
- 3) Расстояние Хэмминга
- 4) Расстояние Дамерау — Левенштейна
- 5) метрика Смита-Вотермана

Исходные данные

The screenshot shows the NCBI website interface. At the top, there's a navigation bar with 'NCBI Resources' and 'How To' links, and a 'Sign in to NCBI' button. Below this is a search bar with a dropdown menu set to 'All Databases'. The left sidebar contains a list of resources, with 'DNA & RNA' highlighted. The main content area is titled 'DNA & RNA' and includes tabs for 'All', 'Databases', 'Downloads', 'Submissions', 'Tools', and 'How To'. The 'Databases' tab is active, showing links to 'Assembly', 'BioProject (formerly Genome Project)', 'BioSample', 'Consensus CDS (CCDS)', 'Database of Expressed Sequence Tags (dbEST)', and 'Database of Genome Survey Sequences (dbGSS)'. Each link is followed by a brief description of the database. On the right, there's a 'Quick Links' section with links to 'BioProject (formerly Genome Project)', 'Database of Short Genetic Variations (dbSNP)', 'GenBank', 'Nucleotide Database', 'PopSet', 'RefSeqGene', 'Reference Sequence (RefSeq)', 'Sequence Read Archive (SRA)', 'Trace Archive', 'UniGene', 'BLAST (Stand-alone)', 'GenBank: BankIt', 'GenBank: Sequin', 'GenBank: tbl2asn', 'Basic Local Alignment Search Tool (BLAST)', and 'E-Utilities'.

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

All Databases Search

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

DNA & RNA

All Databases Downloads Submissions Tools How To

Databases

[Assembly](#)
A database providing information on the structure of assembled genomes, assembly names and other meta-data, statistical reports, and links to genomic sequence data.

[BioProject \(formerly Genome Project\)](#)
A collection of genomics, functional genomics, and genetics studies and links to their resulting datasets. This resource describes project scope, material, and objectives and provides a mechanism to retrieve datasets that are often difficult to find due to inconsistent annotation, multiple independent submissions, and the varied nature of diverse data types which are often stored in different databases.

[BioSample](#)
The BioSample database contains descriptions of biological source materials used in experimental assays.

[Consensus CDS \(CCDS\)](#)
A collaborative effort to identify a core set of human and mouse protein coding regions that are consistently annotated and of high quality.

[Database of Expressed Sequence Tags \(dbEST\)](#)
A division of GenBank that contains short single-pass reads of cDNA (transcript) sequences. dbEST can be searched directly through the Nucleotide EST Database.

[Database of Genome Survey Sequences \(dbGSS\)](#)
A division of GenBank that contains short single-pass reads of genomic DNA. dbGSS can be searched directly through the Nucleotide GSS Database.

Quick Links
[BioProject \(formerly Genome Project\)](#)
[Database of Short Genetic Variations \(dbSNP\)](#)
[GenBank](#)
[Nucleotide Database](#)
[PopSet](#)
[RefSeqGene](#)
[Reference Sequence \(RefSeq\)](#)
[Sequence Read Archive \(SRA\)](#)
[Trace Archive](#)
[UniGene](#)
[BLAST \(Stand-alone\)](#)
[GenBank: BankIt](#)
[GenBank: Sequin](#)
[GenBank: tbl2asn](#)
[Basic Local Alignment Search Tool \(BLAST\)](#)
[E-Utilities](#)

<https://www.ncbi.nlm.nih.gov/guide/dna-rna/>

Результаты вычислений.

Матрица расстояний 100X100

	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
1	41	40	43	49	43	57	43	52	43	43	42	61	42	43	43	40	2	38	43	46	41	41	43	44	43	40	46	41	42	39
2	41	39	44	89	43	68	45	63	43	43	42	47	41	43	43	42	2	40	44	60	40	41	43	44	42	39	43	40	41	38
3	69	72	61	37	62	39	61	39	62	61	70	39	61	62	62	35	4	36	67	37	69	71	61	67	67	76	37	72	71	34
4	57	57	56	37	58	39	57	39	57	58	56	39	58	58	58	33	5	34	58	36	56	58	57	59	59	57	36	57	58	32
5	66	65	62	40	63	42	63	41	63	62	66	40	62	63	63	36	4	37	80	40	63	67	63	79	75	66	38	66	68	35
6	41	40	43	45	43	53	43	49	43	43	41	60	42	43	43	36	3	36	42	43	41	41	43	43	42	41	45	41	42	36
7	71	68	62	39	62	40	61	40	63	62	66	39	61	63	63	35	3	36	67	38	66	72	62	68	67	69	38	69	73	34
8	59	58	58	37	60	38	58	39	59	59	57	38	59	60	59	33	4	34	59	36	58	59	59	60	60	60	36	59	60	32
9	42	40	44	70	44	72	44	65	43	43	42	48	41	43	43	41	1	40	44	58	40	41	43	44	43	39	43	40	42	38
10	61	60	62	40	63	41	63	42	62	62	60	39	62	63	62	35	4	35	63	39	59	62	63	65	64	61	38	61	63	33
11	76	68	62	39	62	41	61	40	62	61	67	39	61	62	62	36	4	37	67	39	65	85	62	68	67	69	38	68	78	35
12	70	72	62	36	63	38	61	38	62	62	68	38	62	63	62	33	4	34	67	36	69	71	62	67	68	75	36	73	71	32
13	38	36	39	44	38	44	38	41	38	38	40	44	36	37	38	45	1	43	40	46	37	37	38	39	38	36	43	37	37	46
14	37	36	38	38	38	40	38	38	37	37	38	38	36	37	37	40	5	39	38	39	37	37	37	39	37	35	38	36	37	40
15	42	40	44	48	43	55	44	52	44	43	42	67	43	43	43	38	1	37	43	45	41	42	44	44	43	41	46	41	42	38
16	70	73	61	36	62	38	61	38	62	62	69	38	62	62	62	34	4	35	66	36	71	70	62	67	67	88	36	73	71	33
17	76	67	61	38	61	39	60	39	61	60	66	39	60	61	61	35	4	36	66	37	65	76	61	67	66	69	37	68	78	34
18	60	59	68	41	80	42	67	42	73	72	59	40	67	81	73	36	3	36	61	41	58	61	68	62	61	60	39	60	62	34
19	62	61	70	41	84	42	70	42	76	75	61	40	70	85	76	36	3	37	63	40	60	63	70	64	63	61	39	62	64	34
20	62	61	54	37	53	39	52	38	54	54	58	39	53	54	54	35	5	35	59	37	59	62	54	58	59	63	37	61	63	34
21	71	68	63	40	63	41	62	41	63	62	67	40	61	63	63	36	3	37	67	40	65	71	62	68	68	68	38	68	72	35
22	61	59	68	40	73	41	70	41	71	71	60	40	68	73	71	36	4	36	62	40	59	61	68	63	62	60	39	60	62	35
23	62	61	70	40	92	42	70	42	75	75	61	40	70	88	76	36	3	36	63	40	60	62	70	64	64	61	39	62	63	34
24	67	60	56	39	56	41	54	41	56	56	59	40	54	56	56	36	4	35	60	39	59	72	56	60	61	62	38	61	70	35
25	66	65	63	41	63	42	63	41	62	62	66	40	61	63	62	36	4	37	92	40	63	67	63	79	75	66	38	66	68	35
26	42	39	45	80	44	68	45	63	43	43	42	47	42	43	43	42	1	40	44	61	40	41	44	45	43	39	43	40	42	38
27	66	65	62	39	63	41	62	41	63	62	66	39	62	62	63	35	4	36	79	39	63	67	62	77	74	66	38	66	67	34
28	61	60	62	40	62	42	62	41	62	61	60	39	61	62	62	35	4	36	63	39	59	62	62	64	63	60	38	61	63	33
29	69	68	60	36	61	38	60	38	60	60	66	38	60	61	60	34	4	35	65	36	65	70	60	66	65	69	36	68	71	34
30	41	39	44	81	44	68	44	64	43	43	42	47	41	43	43	42	1	40	44	60	40	41	44	44	43	39	43	40	41	38

Матрица расстояний рассматривается как метрическое пространство

Метрическое пространство

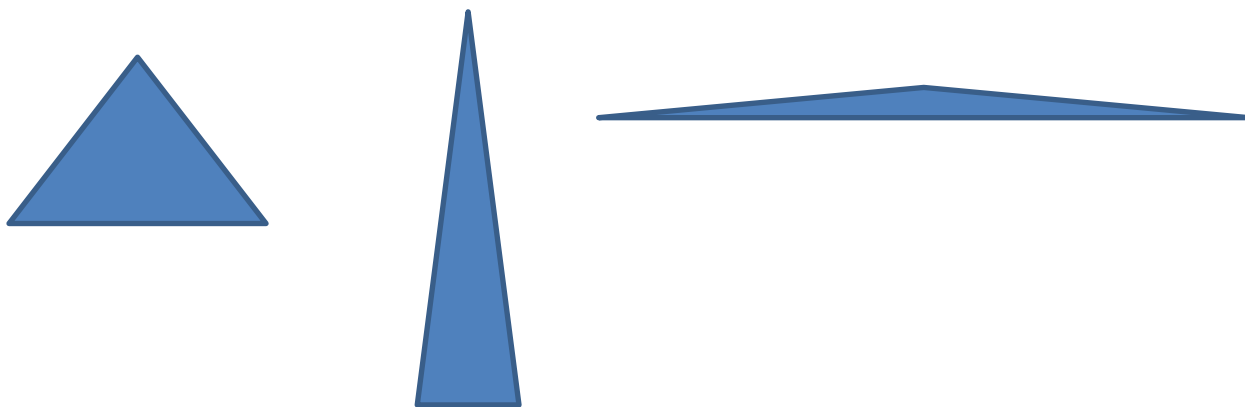
Метрическое пространство M есть множество точек с функцией расстояния (также называется **метрикой**) (где обозначает множество вещественных чисел). Для любых точек x, y, z из M эта функция должна удовлетворять следующим условиям:

$$d(x, y) \geq 0$$

$$d(x, y) = 0 \quad x = y.$$

$$d(x, y) = d(y, x) \quad (\text{симметрия})$$

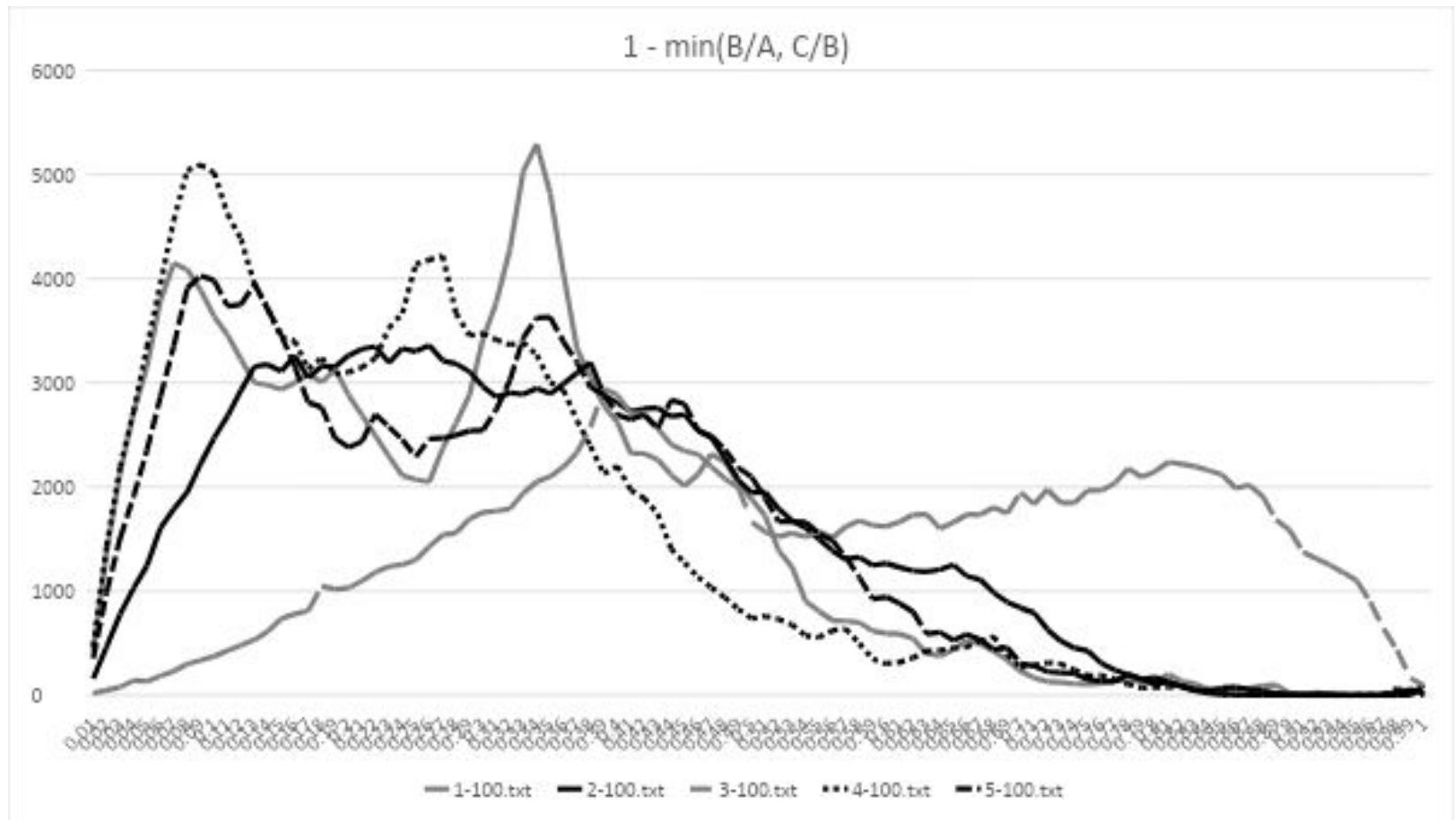
$$d(x, z) \leq d(x, y) + d(y, z) \quad (\text{неравенство треугольника}).$$



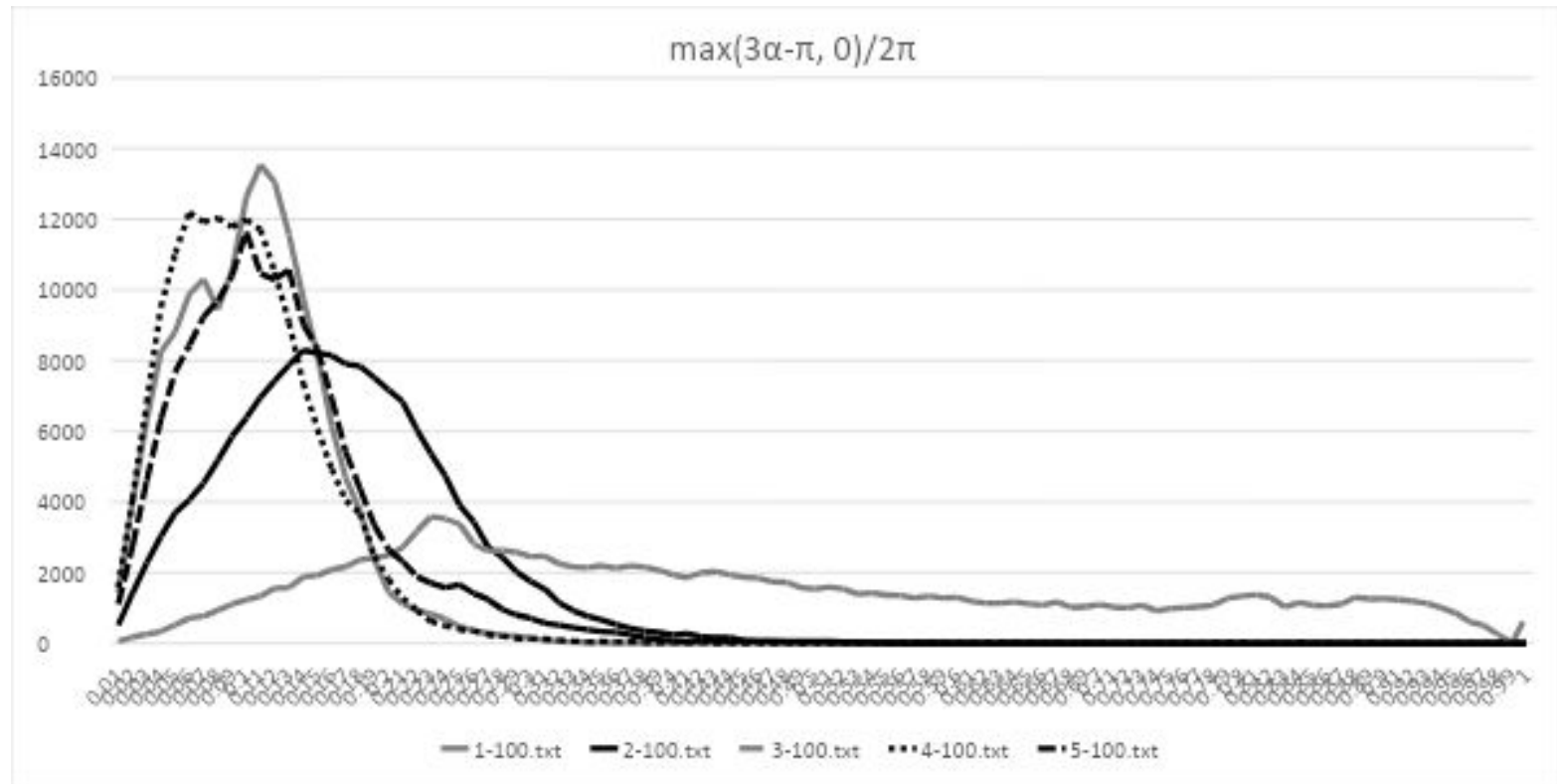
Bedness

- Итак, мы в *простых случаях* будем считать badness (норму) всей матрицы расстояний суммой, а для badness каждого треугольника будем применять один из следующих 4 вариантов. (Всюду считаем, что в рассматриваемом треугольнике стороны – a , b и c , причём $a \geq b \geq c$; углы – α , β и γ , причём $\alpha \geq \beta \geq \gamma$.)
- $(\alpha - \beta) / \pi$.
- $(\alpha - \beta) / \alpha$.
- $(a - b) / a$.
- В последней норме «нарушение равнобедренности» и «нарушение остроугольности» рассмотрим *отдельно*:
- (A) $1 - \min(b/a, c/b)$;
- (B) $\max(3\alpha - \pi, 0) / (2\pi)$;
- общий ответ – $(A+B) / 2$.
- При этом максимальные значения badness (в каждом из этих 4 случаев) для некоторого треугольника может быть равно 1. В самом же плохом случае работы алгоритмов построения метрики – т.е. при возникающем нарушении неравенства треугольника – мы полагаем это значение равным от 1 до 2 (также в зависимости от количественных характеристик этого нарушения).
- Отметим заранее, что мы иногда рассматриваем и несколько более сложные варианты, которые, однако, в настоящей статье не описаны.

Нарушение равнобедренности



Нарушение остроугольности



ИТОГИ

№	время (ч)	нарушения	badness-1, $(\alpha-\beta) / \pi$	badness-2, $(\alpha-\beta) / \alpha$	badness-3, $(a-b) / a$	badness-4, $(A+B) / 2$
1	27	0	0,0372	0,0822	0,0416	0,196
2	2.1	0	0,0954	0,197	0,0926	0,252
3	2.3	0	0,345	0,476	0,163	0,468
4	28	0.37	0,0416	0,0907	0,0469	0,176
5	28	0.38	0,0549	0,116	0,0556	0,214