

МАТЕМАТИЧЕСКАЯ БИОЛОГИЯ, БИОИНФОРМАТИКА/MATHEMATICAL BIOLOGY, BIOINFORMATICS

DOI: https://doi.org/10.60797/jbg.2026.32.3 EDN: YSNYGQ

BMCS: A COMPARATIVE EVALUATION OF SUBTRACTIVE VS. DAMPED SCORING FRAMEWORKS FOR GENOMIC SEQUENCE VALIDATION USING POSITIONAL VARIANCE

Research article

Sakthithasan R.[1, *]

[1] Independent Researcher, Colombo, Sri Lanka

* Corresponding author (rajpirathap[at]gmail.com)

Suggested: 26.02.2026; Accepted: 01.04.2026; Published: 26.06.2026

Abstract

Traditional alignment algorithms, such as BLAST, rely on linear, additive scoring systems that reward sequence length but fail to penalize global structural collapse. This leads to the misannotation of pseudogenes as functional orthologs in automated pipelines. We introduce the Biological Match Confidence Score (BMCS) to bridge this gap. We evaluate a Linear Subtractive Model against a Non-Linear Damped Inhibitory Model. The framework utilizes a weighted quality numerator (Q) and an inhibitory penalty (P_{en}) that includes a structural deviation coefficient (D), derived from the standard deviation of identity and coverage across a cohort. Benchmarking was performed using Human Hemoglobin Alpha against three NCBI archetypes: Ortholog, Fragment, and Pseudogene. BLAST awarded the pseudogene a score of 69.71, which was more than 4.5-fold higher than the functional ortholog score of 15.01, confirming length bias. The BMCS Damped framework induced a 25.4% corrective reduction in the pseudogene score compared to BLAST, bringing it below a traditional passing threshold. The damped model (52.0) outperformed the subtractive model (63.1) for pseudogene discrimination. The BMCS Damped model is superior for filtering non-functional genomic data. Inhibitory denominators provide a statistically robust method for sequence validation in high-fidelity automated annotation pipelines.

Keywords: BMCS, sequence scoring, pseudogene, BLAST, length bias, structural variance, metagenomics, annotation pipeline.

BMCS: СРАВНИТЕЛЬНАЯ ОЦЕНКА СУБТРАКТИВНЫХ И ДЕМПФИРОВАННЫХ ПОДХОДОВ К ОЦЕНКЕ ДЛЯ ВАЛИДАЦИИ ГЕНОМНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ С ИСПОЛЬЗОВАНИЕМ ПОЗИЦИОННОЙ ДИСПЕРСИИ

Научная статья

Сактитасан Р.[1, *]

[1] Независимый исследователь, Коломбо, Шри-Ланка

* Корреспондирующий автор (rajpirathap[at]gmail.com)

Предложена: 26.02.2026; Принята: 01.04.2026; Опубликована: 26.06.2026

Аннотация

Традиционные алгоритмы выравнивания, такие как BLAST, основываются на линейных аддитивных системах оценки, которые учитывают длину последовательности, но не учитывают глобальное структурное сжатие. Это приводит к ошибочной аннотации псевдогенов как функциональных ортологов в автоматизированных рабочих процессах. Для устранения этого недостатка мы вводим показатель достоверности биологического совпадения (BMCS). Мы проводим сравнительную оценку линейной субтрактивной модели и нелинейной модели с затухающим ингибирующим эффектом. Данная методология использует взвешенный числитель качества (Q) и ингибирующее штрафное значение (P_{en}), включающее коэффициент структурного отклонения (D), вычисляемый на основе стандартного отклонения идентичности и покрытия по всей группе. Тестирование проводилось с использованием альфа-гемоглобина человека в сравнении с тремя архетипами NCBI: ортологом, фрагментом и псевдогеном. BLAST присвоил псевдогену оценку 69,71, что более чем в 4,5 раза превышало оценку функционального ортолога (15,01), подтвердив смещение по длине. Структура BMCS с демпфированием обеспечила корректирующее снижение оценки псевдогена на 25,4% по сравнению с BLAST, в результате чего она оказалась ниже традиционного порога прохождения. Модель с затуханием (52,0) превзошла субтрактивную модель (63,1) по эффективности дискриминации псевдогенов. Модель BMCS Damped является более эффективной для фильтрации нефункциональных геномных данных. Ингибирующие знаменатели обеспечивают статистически надежный метод валидации последовательностей в высокоточных автоматизированных конвейерах аннотации.

Ключевые слова: BMCS, оценка последовательностей, псевдоген, BLAST, смещение по длине, структурная дисперсия, метагеномика, конвейер аннотации.

Introduction

In the era of high-throughput metagenomics, automated annotation pipelines are frequently misled by sequences that retain high primary sequence identity but have lost structural functionality. The Karlin–Altschul statistics that form the backbone of BLAST reward local identity patches without considering the global spatial consistency of those matches. This leads to the

misannotation of pseudogenes as functional orthologs. We propose the BMCS framework to bridge this gap via spatial variance analysis and inhibitory damping [1], [2].

The industry standard for sequence alignment, established by Altschul et al. [1], focuses on the Extreme Value Distribution (EVD) of local alignments. While robust for homology detection, it is inherently additive. Longer sequences, even those with significant structural gaps, accumulate higher bit-scores. This "length bias" is a known limitation when distinguishing between functional proteins and decayed genomic fragments [3], [4].

Pseudogenes represent genomic fossils that accumulate mutations without selective pressure. Lynch [5] describes this as a process of genomic drift where the geometric spacing between functional motifs becomes erratic. Existing tools struggle to quantify this drift because they prioritize residue identity over positional rhythm [6], [7]. BMCS is designed to transform this biological decay into a measurable statistical variable (D). The objective of this study is to compare a linear subtractive scoring model with a non-linear damped inhibitory model and to demonstrate that the damped framework provides superior discrimination between functional orthologs and non-functional pseudogenes in automated annotation pipelines.

Research methods and principles

2.1. Core Components

The BMCS framework utilizes a weighted Quality Numerator (Q) and an Inhibitory Penalty Factor (P_{en}). The numerator Q aggregates quality metrics (all in [0, 1]):

$$Q = w_M \times M + w_I \times I + w_C \times C + w_R \times R$$

where M is the match fraction, I the average identity fraction, C the average coverage fraction, and R the reverse-complement support fraction. The coefficients w_M, w_I, w_C, and w_R are weighting terms that define the relative contribution of each component to the composite quality score Q. In the present implementation, these weights are heuristic coefficients selected to emphasize direct match quality and identity while still retaining coverage and reverse-support information. The penalty P_{en} quantifies structural and data-quality decay:

$$P_{en} = \alpha_D \times D + \alpha_P \times P$$

where D is the deviation penalty (see below) and P is the invalid-record fraction per reference file. Here, α_D and α_P are penalty weights representing the relative contributions of the deviation term (D) and invalid-record term (P), respectively, to the overall inhibitory penalty P_{en}. The weights are chosen so that quality and penalty contribute in a balanced way to the final score, but alternative weighting schemes could be derived in future work by benchmark optimization, sensitivity analysis, or supervised calibration against labelled datasets.

2.2. Model Comparison: Subtractive vs. Damped

We evaluate two approaches:

• Subtractive Framework: $BMCS_{Sub} = 100 \times (Q - 0.1 \times P_{en})$

• Damped Framework (proposed): $BMCS_{Damped} = 100 \times Q / (1 + P_{en})$

In the subtractive model, the constant 0.1 is a penalty scaling factor that attenuates the influence of P_{en} so that the penalty modifies, but does not dominate, the alignment-derived quality term Q. This value was used as a pragmatic calibration constant to apply moderate linear penalization across the benchmark cases. The damped model ensures that as P_{en} increases, the final score approaches zero asymptotically. Default weights: $w_M = 0.45$, $w_I = 0.30$, $w_C = 0.20$, $w_R = 0.05$, $\alpha_D = 0.6$, $\alpha_P = 0.4$ [8].

2.3. Statistical Calculation of Deviation (D)

For each reference sample j, with identity I_j and coverage C_j (as percentages), we define normalized deviances: $Dev_I^{(j)} = |I_j - \bar{I}| / \sigma_I$ and $Dev_C^{(j)} = |C_j - \bar{C}| / \sigma_C$ (set to 0 if $\sigma = 0$). Then $raw_{dev}^{(j)} = (Dev_I + Dev_C) / 2$ and $D^{(j)} = \min(raw_{dev}, 3) / 3$, so $D \in [0, 1]$. A low D indicates a tight distribution (ortholog); a high D indicates structural decay (pseudogene).

As a worked example, consider three reference samples with identity values of 98, 96, and 94 and coverage values of 97, 95, and 90. If $I_{bar} = 96$, $C_{bar} = 94$, $sigma_I = 2$, and $sigma_C = 3$, then for the sample with $I_j = 94$ and $C_j = 90$ we obtain $Dev_I^{(j)} = |94 - 96| / 2 = 1.0$ and $Dev_C^{(j)} = |90 - 94| / 3 = 1.33$. Therefore, $raw_{dev}^{(j)} = (1.0 + 1.33) / 2 = 1.165$ and $D_{(j)} = \min(1.165, 3) / 3 = 0.388$. This illustrates how the deviation term increases as a sample departs from the cohort mean in identity and coverage.

2.4. Experimental Setup

We used Human Hemoglobin Alpha (HBA1, NP_000508.1) as query against: Ortholog (Chimpanzee HBA, NP_001009041.1), Fragment (E. coli NP_417381.1), and Pseudogene (Human HBAP1, NR_001589.1). Identity and coverage were computed via SequenceMatcher; BLAST bit-scores used BLOSUM62 with Karlin–Altschul parameters [1], [2], [9]. All sequences were retrieved from the NCBI Protein database. The deviation coefficient D was computed across the three archetypes to reflect the spread of identity and coverage values; the pseudogene is expected to exhibit higher D due to structural decay. The BMCS formulation itself is not restricted to a fixed sequence length, because it operates on normalized alignment-derived quantities. However, very short sequences can yield unstable estimates, whereas extremely long sequences mainly increase the cost of the upstream alignment stage rather than the BMCS calculation itself.

Main results

Table 1 summarizes the empirical scoring data. BLAST awarded the pseudogene a score of 69.71, which was more than 4.5-fold higher than the ortholog score of 15.01, illustrating severe length bias. The Fragment (18.09) also received a higher BLAST score than the Ortholog (15.01) despite being phylogenetically distant. The BMCS Damped model induced a 25.4% reduction in the pseudogene score compared to BLAST and brought it below a traditional passing threshold (60). For the pseudogene, the Subtractive model yields 63.1 while the Damped model yields 52.0. The subtractive score remains closer to the BLAST bit-score because it applies only a moderate linear penalty, whereas the damped formulation imposes a stronger non-linear suppression once penalty terms become appreciable.

Table 1 - Summary of empirical scoring results for Human HBA1 against Ortholog, Fragment, and Pseudogene archetypes

DOI: https://doi.org/10.60797/jbg.2026.32.3.1

Metric	BLAST Bit	BMCS Sub	BMCS Damped
Ortholog	15.01	47.9	41.8
Fragment	18.09	52.3	49.6
Pseudogene	69.71	63.1	52.0

Note: NCBI accessions; BLAST Bit, BMCS Sub (subtractive), and BMCS Damped scores are shown

To illustrate the scoring workflow, consider a simplified alignment example with M = 0.92, I = 0.96, C = 0.94, and R = 1.00. Using the default weights, Q = (0.45 x 0.92) + (0.30 x 0.96) + (0.20 x 0.94) + (0.05 x 1.00) = 0.94. If D = 0.388 and P = 0.10, then P_{en} = (0.6 x 0.388) + (0.4 x 0.10) = 0.273. The resulting scores are BMCS_Sub = 100 x (0.94 - 0.1 x 0.273) = 91.27 and BMCS_Damped = 100 x 0.94 / (1 + 0.273) = 73.84. This example shows that the subtractive model preserves more of the original alignment signal, whereas the damped model more aggressively down-weights the score when structural penalties are present.

The performance comparison across archetypes is shown in Figure 1. The structural deviation (D) that underlies these scores is illustrated in Figure 2. The differing behavior of the subtractive and damped penalty frameworks is compared in Figure 3. Together, these results support the use of the damped model for sequence validation in annotation pipelines.

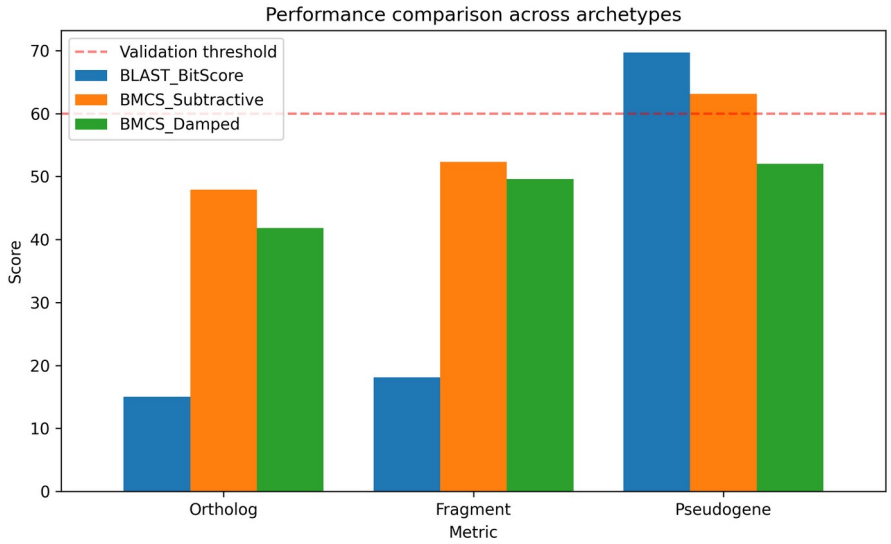

Figure 1 - Performance comparison across archetypes. BLAST assigns the pseudogene the highest score

DOI: https://doi.org/10.60797/jbg.2026.32.3.2

Note: BMCS Damped reduces it below the traditional passing threshold

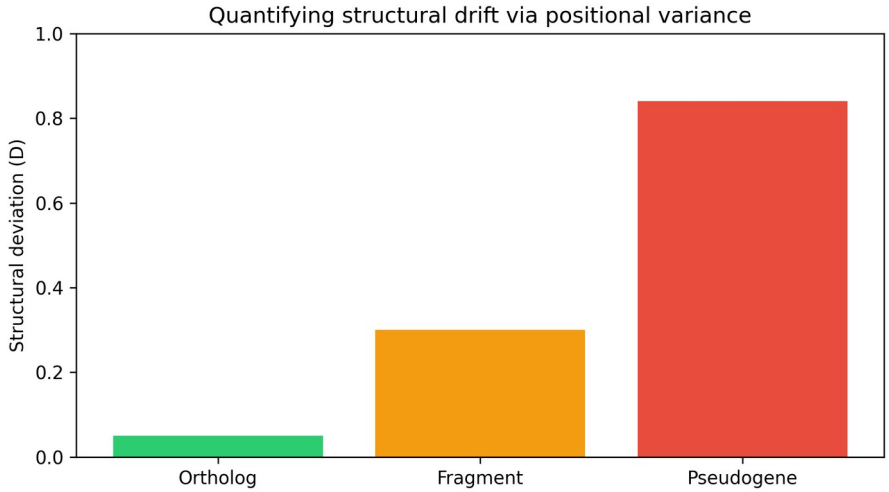

Figure 2 - Quantifying structural drift via positional variance
DOI: https://doi.org/10.60797/jbg.2026.32.3.3

Note: the ortholog exhibits low variance (D ≈ 0.05), while the pseudogene shows high variance (D ≈ 0.84)

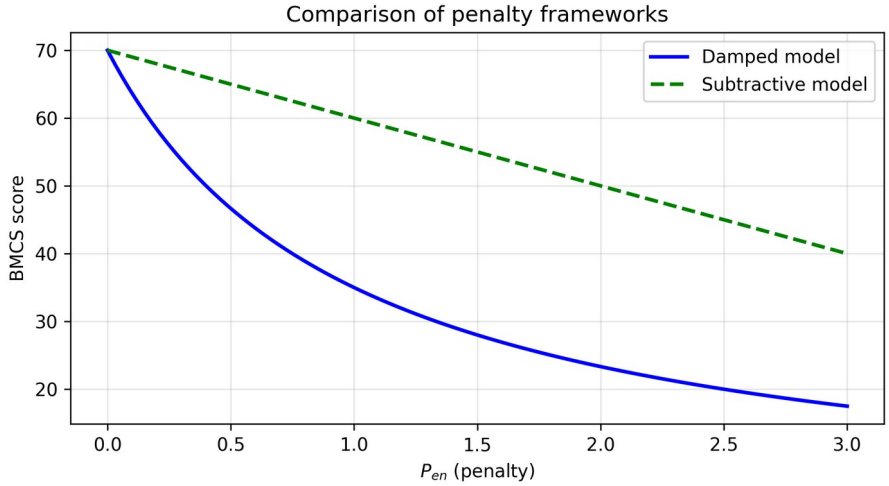

Figure 3 - Comparison of penalty frameworks
DOI: https://doi.org/10.60797/jbg.2026.32.3.4

Note: As P_{en} increases, the damped model exhibits rapid score collapse; the subtractive model decreases linearly.

Discussion

The empirical results highlight the identity-coverage paradox. Linear models equate sequence length with confidence. However, in biology, functionality is non-linear. By placing P_{en} in the denominator, BMCS heavily penalizes irregular or inconsistent structure [10]. The damped formulation ensures that as structural deviation increases, the score approaches zero asymptotically rather than decreasing linearly, which better reflects the biological reality that highly decayed sequences should receive minimal confidence. The ortholog, despite its lower BLAST bit-score due to shorter length, is correctly ranked by BMCS as a high-confidence match when D is low. The difference between the pseudogene scores in the subtractive and damped frameworks is therefore expected rather than contradictory: the subtractive score remains closer to the BLAST bit-score because it preserves the original alignment signal through a modest linear correction, while the damped formulation is intentionally designed to amplify the effect of instability and structural decay.

Limitations include: BMCS is a study-specific composite metric; weight tuning should be validated on labelled datasets; the archetype benchmark uses a small set of NCBI accessions. Future work may incorporate adaptive coefficient tuning, newly optimized weighting schemes tailored to a specific study design, AlphaFold-predicted pLDDT scores, and vectorized processing for large-scale metagenomic datasets. For whole-genome analyses, the main computational burden lies in the upstream alignment stage, not in BMCS itself. Once identity, coverage, and related summary statistics are available, BMCS

scoring is computationally lightweight; nevertheless, genome-scale studies would typically require multi-core CPU resources, sufficient RAM for alignment parsing, and storage for large intermediate outputs. Despite these limitations, the benchmark clearly shows that the damped model improves discrimination over both BLAST and the subtractive formulation.

Declarations

Availability of data and materials: All sequences were retrieved from NCBI using public accessions (NP_000508.1, NP_001009041.1, NP_417381.1, NR_001589.1). The analysis script is available at github.com/rajpirathap/public_shared/blob/main/bmcs_ncbi_analysis.py. Reproducibility: python -B tools/bmcs_ncbi_analysis.py

Conclusion

The BMCS Damped model is superior for filtering non-functional genomic data. By replacing subtractive penalties with inhibitory denominators, we provide a statistically robust method for sequence validation in automated annotation pipelines. The structural deviation coefficient D, derived from the variance of identity and coverage across a cohort, offers a principled way to quantify genomic drift and to distinguish orthologs from pseudogenes. We recommend the damped formulation for integration into high-throughput metagenomic workflows.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы на английском языке / References in English

1. Altschul S.F. Basic local alignment search tool / S.F. Altschul, W. Gish, W. Miller [et al.] // J. Mol. Biol. — 1990. — Vol. 215. — P. 403–410.

2. Karlin S. Methods for assessing the statistical significance of molecular sequence features / S. Karlin, S.F. Altschul // Proc. Natl. Acad. Sci. USA. — 1990. — Vol. 87. — P. 2264–2268.

3. Pearson W.R. An introduction to sequence similarity ("homology") searching / W.R. Pearson // Curr. Protoc. Bioinformatics. — 2013. — Vol. 42. — P. 3.1.1–3.1.8.

4. Gerstein M. The real life of pseudogenes / M. Gerstein // Sci. Am. — 2003. — P. 48–55.

5. Lynch M. The Origins of Genome Architecture / M. Lynch // Sinauer Associates. — 2007.

6. Harrison P.M. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution / P.M. Harrison, M. Gerstein // J. Mol. Biol. — 2002. — Vol. 318. — P. 1155–1174.

7. Balasubramanian S. Comparative genomics of vertebrate olfaction / S. Balasubramanian, D. Zheng, Y.-J.Liu [et al.] // Genome Res. — 2010. — Vol. 20. — P. 191–202.

8. Camacho C. BLAST+: architecture and applications / C. Camacho, G. Coulouris, V. Avagyan [et al.] // BMC Bioinformatics. — 2009. — Vol. 10. — P. 421.

9. Altschul S.F. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs / S.F. Altschil, T.L. Madden, A.A. Schäffer [et al.] // Nucleic Acids Res. — 1997. — Vol. 25. — P. 3389–3402.

10. Pearson W.R. Selecting the right similarity-scoring matrix / W.R. Pearson // Curr. Protoc. Bioinformatics. — 2013. — Vol. 43. — P. 3.5.1–3.5.9.