Human Genome Center

Division of Medical Data Informatics
医療データ情報学分野

Professor	Tetsuo Shibuya, Ph.D.	教　授　博士(理学)	渋　谷　哲　朗
Assistant Professor	Robert Daniel Barish, Ph.D.	助　教　博士(学術)	ロバート　ダニエル　バリッシュ

The objective of Division of Medical Data Informatics is to develop fundamental data informatics technologies for medical data, including algorithm theory, big data technologies, artificial intelligence, data mining, and privacy preserving technologies. Medical data, especially genome data are increasing exponentially from basics to clinical research in medical science. Our aim is to innovate medical science with novel data informatics solutions.

1. Development of Privacy Preserving Technologies

a. Differentially Private Mechanisms using Direction-oriented Smooth Sensitivity

Akito Yamamoto[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, The Institute of Medical Science, The University of Tokyo

With the emergence of data science, protecting the privacy of individuals' sensitive information is crucial. Data collection and publication under differential privacy have gained considerable attention in a wide range of fields, such as bioinformatics and machine learning. In this context, the benefit of the concept of smooth sensitivity, which can add tailored noise to respective datasets, has been recognized substantially. However, the existing concept focuses only on absolute values regarding the added noise and does not consider its direction. By adjusting the noise scales in the positive and negative directions for each output element, more tailored and realistic perturbations can reduce overall noise. Therefore, we propose a novel concept, direction-oriented smooth sensitivity (DOSS), where the amount of noise is never larger than when using the original smooth sensitivity [1]. Further, there is a lack of theoretical analysis and discussion of the probability distributions used for noise generation; therefore, a concise general form is provided for the first time. We then propose a new-differentially private algorithm using DOSS and our general form, along with efficient computation methods. To demonstrate the effectiveness of the proposed algorithm, we applied it to genomic statistical analysis, which plays a crucial role in the development of personalized medicine.

We furthermore improve the concept of direction-oriented local sensitivity to reduce the lower bound of the noise scales, and we re-define DOSS [2]. Simultaneously, we revisit the conditions for satisfying ϵ-differential privacy and use the enhanced DOSS to construct a more sophisticated mechanism. Furthermore, we propose an unbiased mechanism where the expected value of the output equals the input value. Experimental results regarding the publication of genome statistics demonstrate that the proposed mechanisms can achieve higher utility than existing mechanisms. Overall, this study represents a significant step toward realizing mechanisms that can add realistic and reliable noise.

b. Subgraph Counting under Local Differential Privacy

Quentin Hillebrand[1], [2]Vorapong Suppakitpaisarn, Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, The Institute of Medical Science, The University of Tokyo, [2]Graduate School of Information Science and Technology, The University of Tokyo

We suggest the use of hash functions to cut down the communication costs when counting triangle and other subgraphs under edge local differential privacy [3]. While various algorithms exist for computing graph statistics, including the count of subgraphs, under the edge local differential privacy, many suffer with high communication costs, making them less efficient for large graphs. Though data compression is a typical approach in differential privacy, its application in local differential privacy requires a form of compression that every node can reproduce. In our study, we introduce linear congruence hashing. Leveraging amplification by sub-sampling, with a sampling size of s, our method can cut communication costs by a factor of s^2, albeit at the cost of increasing variance in the published graph statistic by a factor of s. The experimental results indicate that, when matched for communication costs, our method achieves a reduction in the ℓ_2-error by up to 1000 times for triangle counts and by up to 10^3 times for 4-cycles counts compared to the performance of leading algorithms.

c. Facility Location Problem under Local Differential Privacy

Quentin Hillebrand[1], [2]Vorapong Suppakitpaisarn, Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, The Institute of Medical Science, The University of Tokyo, [2]Graduate School of Information Science and Technology, The University of Tokyo

We introduce an adaptation of the facility location problem and analyze it within the framework of local differential privacy (LDP) [5]. Under this model, we ensure the privacy of client presence at specific locations. When n is the number of points, Gupta et al. established a lower bound of $\Omega(n^{0.5})$ on the approximation ratio for any differentially private algorithm applied to the original facility location problem. As a result, subsequent works have adopted the superset assumption, which may, however, compromise user privacy. We show that this lower bound does not apply to our adaptation by presenting an LDP algorithm that achieves a constant approximation ratio with a relatively small additive factor. Additionally, we provide experimental results demonstrating that our algorithm outperforms the straightforward approach on both synthetically generated and real-world datasets.

d. Cycle Counting under Local Differential Privacy for Degeneracy-bounded Graphs

Quentin Hillebrand[1], [2]Vorapong Suppakitpaisarn, Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, The Institute of Medical Science, The University of Tokyo, [2]Graduate School of Information Science and Technology, The University of Tokyo

We propose an algorithm for counting the number of cycles under local differential privacy for degeneracy-bounded input graphs [6]. Numerous studies have focused on counting the number of triangles under the privacy notion, demonstrating that the expected ℓ_2-error of these algorithms is $\Omega(n^{1.5})$, where n is the number of nodes in the graph. When parameterized by the number of cycles of length four (C_4), the best existing triangle counting algorithm has an error of $O(n^2)$. In this paper, we introduce an algorithm with an expected ℓ_2-error of $O(\delta^{1.5}n^{0.5} + \delta^{0.5}d_{\max}^{0.5}n^{0.5})$, where δ is the degeneracy and $d$max is the maximum degree of the graph. For degeneracy-bounded graphs ($\delta \in \Theta(1)$) commonly found in practical social networks, our algorithm achieves an expected ℓ_2-error of $O(n)$. Our algorithm's core idea is a precise count of triangles following a preprocessing step that approximately sorts the degree of all nodes. This approach can be extended to approximate the number of cycles of length k, maintaining a similar ℓ_2-error, namely $O(n^{(k-1)/2})$ for degeneracy-bounded graphs.

2. Development of Technologies for Sequence Analysis

a. Improved Approximation Ratios for the Shortest Common Superstring Problem with Reverse Complements

Yrosuke Yamano[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, The Institute of Medical Science, The University of Tokyo

The Shortest Common Superstring (SCS) problem asks for the shortest string that contains each of a given set of strings as a substring. Its reverse-complement variant, the Shortest Common Superstring problem with Reverse Complements (SCS-RC), naturally arises in bioinformatics applications, where for each input string, either the string itself or its reverse complement must appear as a substring of the superstring. The well-known MGREEDY algorithm for the standard SCS constructs a superstring by first computing an optimal cycle cover on the overlap graph and then concatenating the strings corresponding to the cycles, while its refined variant, TGREEDY, further improves the approximation ratio. Although the original 4- and 3-approximation bounds of these algorithms have been successively improved for the standard SCS, no such progress has been made for the reverse complement setting. A previous study extended MGREEDY to SCS-RC with a 4-approximation guarantee and briefly suggested that extending

TGREEDY to the reverse-complement setting could achieve a 3-approximation. In this work, we strengthen these results by proving that the extensions of MGREEDY and TGREEDY to the reverse-complement setting achieve 3.75- and 2.875-approximation ratios, respectively [7]. Our analysis extends the classical proofs for the standard SCS to handle the bidirectional overlaps introduced by reverse complements. These results provide the first formal improvement of approximation guarantees for SCS-RC, with the 2.875-approximate algorithm currently representing the best-known bound for this problem.

b. Linear-Space Subquadratic-Time String Alignment Algorithm for Arbitrary Scoring Matrices

Yrosuke Yamano[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, The Institute of Medical Science, The University of Tokyo

Theoretically, the fastest algorithm by Crochemore et al. for computing the alignment of two given strings of size n over a constant alphabet takes $O(n^2/\log n)$ time. The algorithm uses Lempel–Ziv parsing to divide the dynamic programming matrix into blocks and utilizes the repetitive structure. It is the only previously known subquadratic-time algorithm that can handle scoring matrices of arbitrary weights. However, this algorithm takes $O(n^2/\log n)$ space, and reducing the space while preserving the time complexity has been an open problem for more than 20 years. We propose a solution to this issue by achieving an $O(n)$ space algorithm that maintains $O(n^2/\log n)$ time [8]. The classical refinement by Hirschberg reduces the space complexity of the textbook $O(n^2)$ algorithm to $O(n)$ while preserving the quadratic time. However, applying this technique to the algorithm of Crochemore et al. has been considered challenging because their method requires $O(n^2/\log n)$ space even when computing only the alignment score. Our modification enables the application of Hirschberg's refinement, allowing traceback computation in $O(n)$ space while preserving the $O(n^2/\log n)$ overall time complexity. Our algorithm can be applied to both global and local string alignment problems.

c. Faster Algorithm for Bounded Damerau–Levenshtein Distance

Yrosuke Yamano[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, The Institute of Medical Science, The University of Tokyo

The Damerau–Levenshtein distance between two strings is the minimum number of insertions, deletions, substitutions, and adjacent transpositions required to transform one string into the other. Unlike the standard Levenshtein distance, it accounts for the common typing error of adjacent character swaps. When edits are restricted so that no substring is edited more than once, existing algorithms for Levenshtein distance can be extended with relatively minor changes to support trans-positions. However, in the unrestricted setting (i.e., edits may overlap or interact arbitrarily), the problem becomes significantly more complex, and existing techniques no longer apply directly. In this work, we show that even in the unrestricted setting, the Damerau–Levenshtein distance can be computed efficiently [9]. We present two algorithms that extend the classic $O(n + k^2)$ edit-distance frameworks of Myers and Landau and Vishkin, adapting them to accommodate unrestricted transpositions. The first algorithm runs in $O(\sigma n + k^2)$ time, where σ is the alphabet size. The second achieves $O(n + k^2 \log n)$ time for integer alphabets. Here, n is the length of the input strings and k is the distance threshold. Experimental results show that our algorithms achieve substantial speedups over k-independent methods when k is small.

3. Developing Technologies for Protein Structure Analysis

a. Protein Hinge Estimation Based on Information Criteria

Bunsho Koyano[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, The Institute of Medical Science, The University of Tokyo

Protein hinges are flexible parts connecting several rigid substructures of proteins that are crucial to determine protein function. Various methods have been developed for efficiently and accurately estimating protein hinge positions by comparing two different

conformations of the same protein for a growing number of protein structures. However, few studies have focused on accurately estimating the number of hinges, and it is required to accurately estimate both the number and positions of hinges. We propose faster and more accurate algorithms for estimating the number and positions of hinges by utilizing information criteria that run in $O(n^2)$-time, where n is the protein length [11]. Our algorithms utilize BIC (Bayesian Information Criterion) or AIC (Akaike Information Criterion) based on a newly proposed k-hinge structure generation model that models the hinge motions between two protein conformations.

b. Comparison Algorithms for Protein Conformational Ensembles

Bunsho Koyano[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, The Institute of Medical Science, The University of Tokyo

Molecular dynamics (MD) simulations yield varied results based on simulation conditions; therefore,

the results must be compared across different conditions. Previous studies have introduced measures to compare two protein conformational ensembles, each containing multiple protein structures, generated via MD simulations. However, existing brute-force algorithms for computing measures, such as the minimum root mean square deviation (RMSD) and average minimum RMSD, require $O(nNM)$-time, where n denotes the protein length and N and M are the number of structures in each ensemble. This time complexity can be prohibitively slow when comparing two conformational ensembles generated via long MD simulations. We propose three faster heuristic methods—single-direction method, dual-direction method, and all-direction method—for computing the minimum RMSD and average minimum RMSD in $O(n(N + M))$-time using the SMAWK algorithm [12].

c. Protein Structure Alignment Algorithms

Masahito Tsukahara[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, The Institute of Medical Science, The University of Tokyo

Detecting common substructures between three-dimensional (3D) protein structures is a fundamental problem in structural bioinformatics and drug discovery. This issue can be formulated as the well-known geometric Largest Common Point-set (LCP) problem under the bottleneck distance. In this study, we focus on the order-dependent alignment variant. The fastest known exact algorithm for this variant has a time complexity of $O(n^{32})$, whereas the fastest approximation algorithm with a theoretical guarantee runs in $O(n^8)$ time, where n is the size of the largest input structure. However, these algorithms suffer from impractical computational costs. Therefore, we propose two new approximation algorithms that improve efficiency and accuracy [13]. The first algorithm optimizes an existing approach, achieving a reduced time complexity of $O(n^7 \log n)$, whereas the second algorithm provides even greater efficiency under specific conditions. Experimental results on the PDB database demonstrate that our proposed methods outperform existing algorithms in both efficiency and accuracy.

4. Development of Technologies for Graph Databases

a. KEGG: biological systems database as a model of the real world

Minoru Kanehisa[1], Miho Furumichi[2], Yoko Sato[2], Yuriko Matsuura[2], Mari Ishiguro-Watanabe[3]: [1]Institute for Chemical Research, Kyoto University,[2]Pathway Solutions, [3]Division of Medical Data Informatics, The Institute of Medical Science, The University of Tokyo

KEGG is a database resource for representation and analysis of biological systems. Pathway maps are the primary dataset in KEGG representing systemic functions of the cell and the organism in terms of molecular interaction and reaction networks. The KEGG Orthology (KO) system is a mechanism for linking genes and proteins to pathway maps and other molecular networks. Each KO is a generic gene identifier and each pathway map is created as a network of KO nodes. This architecture enables KEGG pathway mapping to uncover systemic features from KO assigned genomes and metagenomes. Additional roles of KOs include characterization of conserved genes and conserved units of genes in organism groups, which can be done by taxonomy mapping. A new tool has been developed for identifying conserved gene orders in chromosomes, in which gene orders are treated as sequences of KOs [14]. Furthermore, a new dataset called VOG (virus ortholog group) is computationally generated from virus proteins and expanded to proteins of cellular organisms, allowing gene orders to be compared as VOG sequences as well. Together with these datasets and analysis tools, new types of pathway maps are being developed to present a global view of biological processes involving multiple organism groups.

b. Packing dimers to maximum occupancy under soft-core constraints

Robert Daniel Barish[1], Tetsuo Shibuya[1]: [1]Division of Medical Data Informatics, The Institute of Medical Science, The University of Tokyo

Dimer models have a long and fruitful history in fields ranging from organic chemistry to statistical mechanics and condensed matter physics, where in each case one can abstract the states or degrees of freedom of a physical system as instances of perfect matchings in a graph. Here, the enumeration of perfect matchings has often proven germane to characterizing the general behavior of physical systems, and in predicting critical phenomena such as phase transitions. In the other direction, the search for exact solutions to the dimer model has been a boon to graph theorists and computer scientists, having resulted in the development of the classic Fisher-Kasteleyn-Temperley algorithm for counting perfect matchings in planar graphs. In this work we introduce the valency model, which represents soft-core constraints for dimer models (i.e., constraints allowing multiple dimers to share an endpoint) at a maximum occupancy limit where we require each vertex in a graph to host some specified number of dimers [19]. In particular, the valency model abstracts these packings as assignments of integral weights to the edges of a graph under constraints for vertex strengths (i.e., sums over weights of incident edges).

Publications

1. Akito Yamamoto and Tetsuo Shibuya, "Direction-Oriented Smooth Sensitivity and Its Application to Genomic Statistical Analysis", *Proc. 30th Australasian Conference on Information Security and Privacy*, 2025, pp.63-83.
2. Akito Yamamoto and Tetsuo Shibuya, "Differentially Private Mechanisms Using Enhanced Direction-Oriented Smooth Sensitivity", *Proc. IEEE 16th Annual Computing and Communication Workshop and Conference*, in press.
3. Quentin Hillebrand, Vorapong Suppakitpaisarn, Tetsuo Shibuya, "Communication Cost Reduction for Subgraph Counting under Local Differential Privacy via Hash Functions", *Transactions on Machine Learning Research*, 2025.
4. Vorapong Suppakitpaisarn, Donlapark Ponnoprat, Nicha Hirankarn, Quentin Hillebrand, "Counting Graphlets of Size k under Local Differential Privacy", *Proc. the 28th International Conference on Artificial Intelligence and Statistics*, *CoRR*, vol. abs/2505.12954.
5. Kevin Pfisterer, Quentin Hillebrand, Vorapong Suppakitpaisarn, "Facility Location Problem Under Local Differential Privacy Without Super-Set Assumption", Data and Applications Security and Privacy XXXIX (DBSec 2025), Lecture Notes in Computer Science, vol. 15722, Springer, Cham, pp. 293-310.
6. Quentin Hillebrand, Vorapong Suppakitpaisarn, and Tetsuo Shibuya, "Cycle Counting under Local Differential Privacy for Degeneracy-bounded Graphs", *LIPIcs*, vol. 327, pp. 49:1–49:22, 2025.
7. Ryosuke Yamano and Tetsuo Shibuya, "Improved Approximation Ratios for the Shortest Common Superstring Problem with Reverse Complements", *Proc. Annual Symposium on Combinatorial Pattern Matching*, in press.
8. Ryosuke Yamano and Tetsuo Shibuya. "Linear-Space Subquadratic-Time String Alignment Algorithm for Arbitrary Scoring Matrices", *LIPIcs*, vol. 344, pp. 21:1-21:14, 2025.
9. Ryosuke Yamano and Tetsuo Shibuya, "Faster Algorithm for Bounded Damerau–Levenshtein Distance", *LNCS*, vol 16073. Springer, Cham; pp 291–303.
10. Cherie Au-Yeung, Yuen-Ting Cheung, Joshua Cheng, Ken Ip, Sau-Dan Lee, Victor Yang, Amy Lau, Chit Lee, Peter Chong, King Wai Lau, Jurgen van Lunenburg, Damon Zheng, Brian Ho, Crystal Tik, Kingsley Ho, Ramesh Rajaby, Chun-Hang Au, Mullin Yu, Wing-Kin Sung, "UniVar: A variant interpretation platform enhancing rare disease diagnosis through robust filtering and unified analysis of SNV, INDEL, CNV and SV", *Computers in Biology and Medicine*, vol. 185, no. 109560.
11. Bunsho Koyano, Tetsuo Shibuya, "Faster and More Accurate Estimation of Protein Hinges Based on Information Criteria", *Journal of Computational Biology*, vol. 32(5), 2025, pp.498-519.
12. Bunsho Koyano, Tetsuo Shibuya, "Fast and Accurate Comparison of Protein Conformational Ensembles", *IPSJ Transactions on Bioinformatics*, vol. 18, pp. 20-38.
13. Masahito Tsukahara and Tetsuo Shibuya, "Efficient and Accurate Approximation Algorithm for Protein Structure Alignment", *LNCS*, vol 15756, pp. 122-134.
14. Minoru Kanehisa, Miho Furumichi, Yoko Sato, Yuriko Matsuura, Mari Ishiguro-Watanabe, "KEGG: biological systems database as a model of the real world", *Nucleic Acids Research*, vol. 53(D1), 2025, Pages D672–D677.
15. Robert Barish, Tetsuo Shibuya, "Sokoban-style PushPush games with unique solutions", *Journal of Information Processing*, vol. 33, 2025, pp. 1101-1109.
16. Robert Barish, Tetsuo Shibuya, "Reconfiguring planar perfect matchings via bounded length alternating cycles", *LNCS*, vol. 16106, pp. 45-56.
17. Robert Barish, Tetsuo Shibuya, "Cubic Planar Positive 1-in-3 Satisfiability and the complexity of tiling finite simply connected regions", *Proc. CJCDCGGG*, in press.
18. Robert Barish, Tetsuo Shibuya, "Counting 2-factors of 4-regular bipartite graphs is #P-complete", JCDCG^3 2022 collection of papers, Lecture Notes in Computer Science, Springer, Vol. 14364, pp. 94-103, 2025.
19. Robert Barish, Tetsuo Shibuya, "Packing dimers to maximum occupancy under soft-core constraints", *LNCS*, vol. 15680, pp. 283-298.
20. Daisuke Tsukayama, Jun-ichi Shirakashi, Tetsuo Shibuya, Hiroshi Imai, "Enhancing computational accuracy with parallel parameter optimization in variational quantum eigensolver", *AIP Advances*, 15(1), 015226, 2025.
21. Takumi Kanezashi, Daisuke Tsukayama, Jun-ichi Shirakashi, Tetsuo Shibuya and Hiroshi Imai, "Utility of NISQ devices: optimizing experimental parameters for the fabrication of Au atomic junction using gate-based quantum computers", *Applied Physics Express*, 18(4), 047001, 2025.
22. Shanchuan Li, Daisuke Tsukayama, Jun-ichi Shirakashi, Tetsuo Shibuya and Hiroshi Imai, Quantum architecture search with neural predictor based on ZX-calculus, *EPJ Quantum Technol*. 12, 106 (2025).