Alberto Apostolico
Maxime Crochemore
Kunsoo Park (Eds.)

LNCS 3537

# Combinatorial Pattern Matching

16th Annual Symposium, CPM 2005
Jeju Island, Korea, June 2005
Proceedings

Springer

Alberto Apostolico Maxime Crochemore
Kunsoo Park (Eds.)

# Combinatorial Pattern Matching

16th Annual Symposium, CPM 2005
Jeju Island, Korea, June 19-22, 2005
Proceedings



Springer

Volume Editors

Alberto Apostolico
University of Padova, Department of Information Engineering
Via Gradenigo 6a, 35131 Padova, Italy
E-mail: axa@dei.unipd.it

Maxime Crochemore
King's College London
and University of Marne-la-Vallée, Institut Gaspard-Monge
77454 Marne-la-Vallée, France
E-mail: maxime.crochemore@univ-mlv.fr

Kunsoo Park
Seoul National University
School of Computer Science and Engineering
Seoul 151-744, Korea
E-mail: kpark@snu.ac.kr

Library of Congress Control Number: 2005927143

CR Subject Classification (1998): F.2.2, I.5.4, I.5.0, I.7.3, H.3.3, J.3, E.4, G.2.1, E.1

| | |
|---|---|
| ISSN | 0302-9743 |
| ISBN-10 | 3-540-26201-6 Springer Berlin Heidelberg New York |
| ISBN-13 | 978-3-540-26201-5 Springer Berlin Heidelberg New York |



Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN: 11496656 06/3142 5 4 3 2 1 0

# Lecture Notes in Computer Science 3537

*Commenced Publication in 1973*
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

## Editorial Board

# Lecture Notes in Computer Science

For information about Vols. 1–3431

please contact your bookseller or Springer

Vol. 3537: A. Apostolico, M. Crochemore, K. Park (Eds.), Combinatorial Pattern Matching. XI, 444 pages. 2005.

Vol. 3535: M. Steffen, G. Zavattaro (Eds.), Formal Methods for Open Object-Based Distributed Systems. X, 323 pages. 2005.

Vol. 3532: A. Gómez-Pérez, J. Euzenat (Eds.), The Semantic Web: Research and Applications. XV, 728 pages. 2005.

Vol. 3531: J. Ioannidis, A. Keromytis, M. Yung (Eds.), Applied Cryptography and Network Security. XI, 530 pages. 2005.

Vol. 3528: P.S. Szczepaniak, J. Kacprzyk, A. Niewiadomski (Eds.), Advances in Web Intelligence. XVII, 513 pages. 2005. (Subseries LNAI).

Vol. 3526: S.B. Cooper, B. Löwe, L. Torenvliet (Eds.), New Computational Paradigms. XVII, 574 pages. 2005.

Vol. 3525: A.E. Abdallah, C.B. Jones, J.W. Sanders (Eds.), Communicating Sequential Processes. XIV, 321 pages. 2005.

Vol. 3524: R. Barták, M. Milano (Eds.), Integration of AI and OR Techniques in Constraint Programming for Combinatorial Problems. XI, 320 pages. 2005.

Vol. 3523: J.S. Marques, N.P. de la Blanca, P. Pina (Eds.), Pattern Recognition and Image Analysis, Part II. XXVI, 733 pages. 2005.

Vol. 3522: J.S. Marques, N.P. de la Blanca, P. Pina (Eds.), Pattern Recognition and Image Analysis, Part I. XXVI, 703 pages. 2005.

Vol. 3521: N. Megiddo, Y. Xu, B. Zhu (Eds.), Algorithmic Applications in Management. XIII, 484 pages. 2005.

Vol. 3520: O. Pastor, J. Falcão e Cunha (Eds.), Advanced Information Systems Engineering. XVI, 584 pages. 2005.

Vol. 3518: T.B. Ho, D. Cheung, H. Li (Eds.), Advances in Knowledge Discovery and Data Mining. XXI, 864 pages. 2005. (Subseries LNAI).

Vol. 3517: H.S. Baird, D.P. Lopresti (Eds.), Human Interactive Proofs. IX, 143 pages. 2005.

Vol. 3516: V.S. Sunderam, G.D. van Albada, P.M.A. Sloot, J.J. Dongarra (Eds.), Computational Science – ICCS 2005, Part III. LXIII, 1143 pages. 2005.

Vol. 3515: V.S. Sunderam, G.D. van Albada, P.M.A. Sloot, J.J. Dongarra (Eds.), Computational Science – ICCS 2005, Part II. LXIII, 1101 pages. 2005.

Vol. 3514: V.S. Sunderam, G.D. van Albada, P.M.A. Sloot, J.J. Dongarra (Eds.), Computational Science – ICCS 2005, Part I. LXIII, 1089 pages. 2005.

Vol. 3513: A. Montoyo, R. Mu\~noz, E. Métais (Eds.), Natural Language Processing and Information Systems. XII, 408 pages. 2005.

Vol. 3510: T. Braun, G. Carle, Y. Koucheryavy, V. Tsaoussidis (Eds.), Wired/Wireless Internet Communications. XIV, 366 pages. 2005.

Vol. 3509: M. Jünger, V. Kaibel (Eds.), Integer Programming and Combinatorial Optimization. XI, 484 pages. 2005.

Vol. 3508: P. Bresciani, P. Giorgini, B. Henderson-Sellers, G. Low, M. Winikoff (Eds.), Agent-Oriented Information Systems II. X, 227 pages. 2005. (Subseries LNAI).

Vol. 3507: F. Crestani, I. Ruthven (Eds.), Information Context: Nature, Impact, and Role. XIII, 253 pages. 2005.

Vol. 3506: C. Park, S. Chee (Eds.), Information Security and Cryptology – ICISC 2004. XIV, 490 pages. 2005.

Vol. 3505: V. Gorodetsky, J. Liu, V. A. Skormin (Eds.), Autonomous Intelligent Systems: Agents and Data Mining. XIII, 303 pages. 2005. (Subseries LNAI).

Vol. 3503: S.E. Nikoletseas (Ed.), Experimental and Efficient Algorithms. XV, 624 pages. 2005.

Vol. 3502: F. Khendek, R. Dssouli (Eds.), Testing of Communicating Systems. X, 381 pages. 2005.

Vol. 3501: B. Kégl, G. Lapalme (Eds.), Advances in Artificial Intelligence. XV, 458 pages. 2005. (Subseries LNAI).

Vol. 3500: S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 632 pages. 2005. (Subseries LNBI).

Vol. 3499: A. Pelc, M. Raynal (Eds.), Structural Information and Communication Complexity. X, 323 pages. 2005.

Vol. 3498: J. Wang, X. Liao, Z. Yi (Eds.), Advances in Neural Networks – ISNN 2005, Part III. L, 1077 pages. 2005.

Vol. 3497: J. Wang, X. Liao, Z. Yi (Eds.), Advances in Neural Networks – ISNN 2005, Part II. L, 947 pages. 2005.

Vol. 3496: J. Wang, X. Liao, Z. Yi (Eds.), Advances in Neural Networks – ISNN 2005, Part II. L, 1055 pages. 2005.

Vol. 3495: P. Kantor, G. Muresan, F. Roberts, D.D. Zeng, F.-Y. Wang, H. Chen, R.C. Merkle (Eds.), Intelligence and Security Informatics. XVIII, 674 pages. 2005.

Vol. 3494: R. Cramer (Ed.), Advances in Cryptology – EUROCRYPT 2005. XIV, 576 pages. 2005.

Vol. 3493: N. Fuhr, M. Lalmas, S. Malik, Z. Szlávik (Eds.), Advances in XML Information Retrieval. XI, 438 pages. 2005.

Vol. 3492: P. Blache, E. Stabler, J. Busquets, R. Moot (Eds.), Logical Aspects of Computational Linguistics. X, 363 pages. 2005. (Subseries LNAI).

Vol. 3489: G.T. Heineman, I. Crnkovic, H.W. Schmidt, J.A. Stafford, C. Szyperski, K. Wallnau (Eds.), Component-Based Software Engineering. XI, 358 pages. 2005.

Vol. 3488: M.-S. Hacid, N.V. Murray, Z.W. Raś, S. Tsumoto (Eds.), Foundations of Intelligent Systems. XIII, 700 pages. 2005. (Subseries LNAI).

Vol. 3486: T. Helleseth, D. Sarwate, H.-Y. Song, K. Yang (Eds.), Sequences and Their Applications - SETA 2004. XII, 451 pages. 2005.

Vol. 3483: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications – ICCSA 2005, Part IV. XXVII, 1362 pages. 2005.

Vol. 3482: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications – ICCSA 2005, Part III. LXVI, 1340 pages. 2005.

Vol. 3481: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications – ICCSA 2005, Part II. LXIV, 1316 pages. 2005.

Vol. 3480: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications – ICCSA 2005, Part I. LXV, 1234 pages. 2005.

Vol. 3479: T. Strang, C. Linnhoff-Popien (Eds.), Location- and Context-Awareness. XII, 378 pages. 2005.

Vol. 3478: C. Jermann, A. Neumaier, D. Sam (Eds.), Global Optimization and Constraint Satisfaction. XIII, 193 pages. 2005.

Vol. 3477: P. Herrmann, V. Issarny, S. Shiu (Eds.), Trust Management. XII, 426 pages. 2005.

Vol. 3475: N. Guelfi (Ed.), Rapid Integration of Software Engineering Techniques. X, 145 pages. 2005.

Vol. 3474: C. Grelck, F. Huch, G.J. Michaelson, P. Trinder (Eds.), Implementation and Application of Functional Languages. X, 227 pages. 2005.

Vol. 3468: H.W. Gellersen, R. Want, A. Schmidt (Eds.), Pervasive Computing. XIII, 347 pages. 2005.

Vol. 3467: J. Giesl (Ed.), Term Rewriting and Applications. XIII, 517 pages. 2005.

Vol. 3465: M. Bernardo, A. Bogliolo (Eds.), Formal Methods for Mobile Computing. VII, 271 pages. 2005.

Vol. 3464: S.A. Brueckner, G.D.M. Serugendo, A. Karageorgos, R. Nagpal (Eds.), Engineering Self-Organising Systems. XIII, 299 pages. 2005. (Subseries LNAI).

Vol. 3463: M. Dal Cin, M. Kaâniche, A. Pataricza (Eds.), Dependable Computing - EDCC 2005. XVI, 472 pages. 2005.

Vol. 3462: R. Boutaba, K.C. Almeroth, R. Puigjaner, S. Shen, J.P. Black (Eds.), NETWORKING 2005. XXX, 1483 pages. 2005.

Vol. 3461: P. Urzyczyn (Ed.), Typed Lambda Calculi and Applications. XI, 433 pages. 2005.

Vol. 3460: Ö. Babaoglu, M. Jelasity, A. Montresor, C. Fetzer, S. Leonardi, A. van Moorsel, M. van Steen (Eds.), Self-star Properties in Complex Information Systems. IX, 447 pages. 2005.

Vol. 3459: R. Kimmel, N.A. Sochen, J. Weickert (Eds.), Scale Space and PDE Methods in Computer Vision. XI, 634 pages. 2005.

Vol. 3458: P. Herrero, M.S. Pérez, V. Robles (Eds.), Scientific Applications of Grid Computing. X, 208 pages. 2005.

Vol. 3456: H. Rust, Operational Semantics for Timed Systems. XII, 223 pages. 2005.

Vol. 3455: H. Treharne, S. King, M. Henson, S. Schneider (Eds.), ZB 2005: Formal Specification and Development in Z and B. XV, 493 pages. 2005.

Vol. 3454: J.-M. Jacquet, G.P. Picco (Eds.), Coordination Models and Languages. X, 299 pages. 2005.

Vol. 3453: L. Zhou, B.C. Ooi, X. Meng (Eds.), Database Systems for Advanced Applications. XXVII, 929 pages. 2005.

Vol. 3452: F. Baader, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XI, 562 pages. 2005. (Subseries LNAI).

Vol. 3450: D. Hutter, M. Ullmann (Eds.), Security in Pervasive Computing. XI, 239 pages. 2005.

Vol. 3449: F. Rothlauf, J. Branke, S. Cagnoni, D.W. Corne, R. Drechsler, Y. Jin, P. Machado, E. Marchiori, J. Romero, G.D. Smith, G. Squillero (Eds.), Applications of Evolutionary Computing. XX, 631 pages. 2005.

Vol. 3448: G.R. Raidl, J. Gottlieb (Eds.), Evolutionary Computation in Combinatorial Optimization. XI, 271 pages. 2005.

Vol. 3447: M. Keijzer, A. Tettamanzi, P. Collet, J.v. Hemert, M. Tomassini (Eds.), Genetic Programming. XIII, 382 pages. 2005.

Vol. 3444: M. Sagiv (Ed.), Programming Languages and Systems. XIII, 439 pages. 2005.

Vol. 3443: R. Bodik (Ed.), Compiler Construction. XI, 305 pages. 2005.

Vol. 3442: M. Cerioli (Ed.), Fundamental Approaches to Software Engineering. XIII, 373 pages. 2005.

Vol. 3441: V. Sassone (Ed.), Foundations of Software Science and Computational Structures. XVIII, 521 pages. 2005.

Vol. 3440: N. Halbwachs, L.D. Zuck (Eds.), Tools and Algorithms for the Construction and Analysis of Systems. XVII, 588 pages. 2005.

Vol. 3439: R.H. Deng, F. Bao, H. Pang, J. Zhou (Eds.), Information Security Practice and Experience. XII, 424 pages. 2005.

Vol. 3438: H. Christiansen, P.R. Skadhauge, J. Villadsen (Eds.), Constraint Solving and Language Processing. VIII, 205 pages. 2005. (Subseries LNAI).

Vol. 3437: T. Gschwind, C. Mascolo (Eds.), Software Engineering and Middleware. X, 245 pages. 2005.

Vol. 3436: B. Bouyssounouse, J. Sifakis (Eds.), Embedded Systems Design. XV, 492 pages. 2005.

Vol. 3434: L. Brun, M. Vento (Eds.), Graph-Based Representations in Pattern Recognition. XII, 384 pages. 2005.

Vol. 3433: S. Bhalla (Ed.), Databases in Networked Information Systems. VII, 319 pages. 2005.

Vol. 3432: M. Beigl, P. Lukowicz (Eds.), Systems Aspects in Organic and Pervasive Computing - ARCS 2005. X, 265 pages. 2005.

¥566.40元

# Preface

The 16th Annual Symposium on Combinatorial Pattern Matching was held on Jeju Island, Korea on June 19–22, 2005. Previous meetings were held in Paris, London, Tucson, Padova, Asilomar, Helsinki, Laguna Beach, Aarhus, Piscataway, Warwick, Montreal, Jerusalem, Fukuoka, Morelia, and Istanbul over the years 1990–2004.

In response to the call for papers, CPM 2005 received a record number of 129 papers. Each submission was reviewed by at least three Program Committee members with the assistance of external referees. Since there were many high-quality papers, the Program Committee's task was extremely difficult. Through an extensive discussion the Program Committee accepted 37 of the submissions to be presented at the conference. They constitute original research contributions in combinatorial pattern matching and its applications.

In addition to the selected papers, CPM 2005 had three invited presentations, by Esko Ukkonen from the University of Helsinki, Ming Li from the University of Waterloo, and Naftali Tishby from The Hebrew University of Jerusalem.

We would like to thank all Program Committee members and external referees for their excellent work, especially given the demanding time constraints; they gave the conference its distinctive character. We also thank all who submitted papers for consideration; they all contributed to the high quality of the conference.

Finally, we thank the Organizing Committee members and the graduate students who worked hard to put in place the logistical arrangements of the conference. It is their dedicated contribution that made the conference possible and enjoyable.

June 2005 Alberto Apostolico, Maxime Crochemore, and Kunsoo Park

# Organization

## Program Committee

| | |
|---|---|
| Tatsuya Akutsu | Kyoto University |
| *Alberto Apostolico* | University of Padova and Purdue University |
| Setsuo Arikawa | Kyushu University |
| *Maxime Crochemore* | University of Marne la Vallée |
| Martin Farach-Colton | Rutgers University |
| Costas S. Iliopoulos | King's College London |
| Juha Karkkainen | University of Helsinki |
| Dong Kyue Kim | Pusan National University |
| Tak-Wah Lam | University of Hong Kong |
| Moshe Lewenstein | Bar Ilan University |
| Bin Ma | University of Western Ontario |
| Giovanni Manzini | University of Piemonte Orientale |
| S. Muthukrishnan | Rutgers University |
| *Kunsoo Park* | Seoul National University |
| Mathieu Raffinot | CNRS |
| Rajeev Raman | University of Leicester |
| Cenk Sahinalp | Simon Fraser University |
| Andrew C. Yao | Tsinghua University |
| Frances F. Yao | City University of Hong Kong |
| Nivio Ziviani | Federal University of Minas Gerais |

## Steering Committee

| | |
|---|---|
| Alberto Apostolico | University of Padova and Purdue University |
| Maxime Crochemore | University of Marne la Vallée |
| Zvi Galil | Columbia University |

## Organizing Committee

| | |
|---|---|
| Dong Kyue Kim | Pusan National University |
| Yoo-Jin Chung | Hankuk University of Foreign Studies |
| Sung-Ryul Kim | Konkuk University |
| Heejin Park | Hanyang University |
| Jeong Seop Sim | Inha University |
| Jin Wook Kim | Seoul National University |

## Sponsoring Institutions

Ministry of Science and Technology, Korea
Seoul National University
SIGTCS of the Korea Information Science Society

## External Referees

Julien Allali
Marie-Pierre Béal
Anne Bergeron
Marshall Bern
Vincent Berry
Mathieu Blanchette
Dan Brown
Julien Clément
Fabien Coulon
Thierson Couto Rosa
Fabiano Cupertino Botelho
Artur Czumaj
Davi de Castro Reis
Fabien de Montgolfier
Funda Ergun
Paolo Ferragina
Guillaume Fertin
Dora Giammarresi
Sylvie Hamel
Tzvika Hartman
Patrice Koehl
Tsvi Kopelowitz
Gregory Kucherov
Thierry Lecroq
Veli Mäkinen
Wagner Meira, Jr.
Nadia Pisanti
Sven Rahman
Isidore Rigoutsos
Eric Rivals
Romeo Rizzi
Dominique Rossin
Kunihiko Sadakane
Jens Stoye
Jorma Tarhio
Stéphane Vialette
Hao-Chi Wong

# Table of Contents

# Sharper Upper and Lower Bounds for an Approximation Scheme for CONSENSUS-PATTERN

Broňa Brejová, Daniel G. Brown, Ian M. Harrower, Alejandro López-Ortiz, and Tomáš Vinař

School of Computer Science, University of Waterloo
{bbrejova,browndg,imharrow,alopez-o,tvinar}@cs.uwaterloo.ca

**Abstract.** We present sharper upper and lower bounds for a known polynomial-time approximation scheme due to Li, Ma and Wang [7] for the CONSENSUS-PATTERN problem. This NP-hard problem is an abstraction of motif finding, a common bioinformatics discovery task. The PTAS due to Li *et al.* is simple, and a preliminary implementation [8] gave reasonable results in practice. However, the previously known bounds on its performance are useless when runtimes are actually manageable. Here, we present much sharper lower and upper bounds on the performance of this algorithm that partially explain why its behavior is so much better in practice than what was previously predicted in theory. We also give specific examples of instances of the problem for which the PTAS performs poorly in practice, and show that the asymptotic performance bound given in the original proof matches the behaviour of a simple variant of the algorithm on a particularly bad instance of the problem.

## 1 Introduction

Bioinformaticists often find themselves with several different DNA or protein sequences that are known to share a particular function, but where the origin of the function in the sequence is unknown. For example, suppose one has the DNA sequence of the region surrounding several genes, known to be regulated by a particular transcription factor. Here, the shared regulatory behavior may be caused by a sequence element common to all, to which the transcription factor binds. Discovering this experimentally is very expensive, so computational approaches can be helpful to limit searches.

The motif discovery problem is an abstraction of this problem. In it, we are given $n$ sequences, all of length $m$, over an alphabet $\Sigma$. We seek a single motif, of length $L$ that is found approximately as a substring of all sequences. Several variants of this problem exist. One can seek to minimize the maximum Hamming distance between the motif and its instances in all strings (*e.g.* [2, 10]), maximize the information content (minimize the entropy) of the chosen motif instances (*e.g.* [1, 3, 6]), or minimize the total of the Hamming distances between the motif and its instances [7]. This latter problem can be formally defined as follows:

A. Apostolico, M. Crochemore, and K. Park (Eds.): CPM 2005, LNCS 3537, pp. 1–10, 2005.

**Definition 1** (CONSENSUS-PATTERN). *Given: $n$ sequences $s_1, \ldots, s_n$, each of length $m$ and over an alphabet of size $A$. Find a substring $t_i$ of a given length $L$ in each of the sequences and a median string $s$ of length $L$ so that the total Hamming distance $\sum_i d_H(s, t_i)$ is minimized.*

Li, Ma and Wang [7] give a very simple polynomial-time approximation scheme (PTAS) for this combinatorial motif problem. For a given value of $r$, consider all choices of $r$ substrings of length $L$ from the $n$ sequences. We note explicitly here that the sampling is made with replacement, so that the same substring may occur multiple times. For each such collection $C$ of substrings, we compute its consensus by identifying the most common letter in the first position of each chosen substring, the second position, and so on, producing a motif $M_C$. It is easy to identify for a given motif $M_C$ its closest match in each of the $n$ sequences, and thus its score. We do this for all $n^r(m-L+1)^r$ possible collections of $r$ substrings, and pick the collection with the best score. The algorithm has $O(L(nm)^{r+1})$ running time, and thus runs in polynomial time for any particular value of $r$. Li *et al.* also give an upper bound on the worst-case approximation ratio of this algorithm for $r \geq 3$:

$$1 + \frac{4A-4}{\sqrt{e}\left(\sqrt{4r+1}-3\right)}, \tag{1}$$

where $A$ is the alphabet size. For example, if $r = 3$, this approach gives an algorithm that runs in $O(L(nm)^4)$ runtime, but whose approximation guarantee for DNA sequences (where $A = 4$) is approximately 13. To achieve a reasonable approximation ratio, 2, we would have to use $r \geq 8$ for DNA sequences, or $r \geq 27$ for protein sequences ($A = 20$), giving hopelessly large running times. The high value of the proven bound would seem to suggest that the algorithm will be useless in practice.

However, many successful combinatorial motif finders do work by generalizing from small samples in this way, such as SP-STAR [10] and CONSENSUS (samples of 1) [3], COMBINE (samples of 2 to 3) [9], COPIA (samples of arbitrary size) [8]. Here, focusing on Li *et al.*'s PTAS, we show tighter bounds on its performance that are much closer to reasonable numbers for practical values of $r$. We also provide the first substantial lower bounds on the PTAS's performance, by identifying specific examples of the problem for which the algorithm performs poorly. In the general case, for a binary alphabet, we find that the variant of the algorithm that works by sampling *without* replacement performs poorly on a particular bad example, and we conjecture that our example will also be difficult for the original Li *et al.* algorithm that samples *with* replacement.

Our results are summarized in Table 1.

## 2 Basic Observations

We begin our discussion of the algorithm by noting that it is sufficient to look at the performance of the PTAS when run on the actual instances of the motif (which are sequences of length $L$), rather than on the $m$-letter input strings.

**Table 1.** Overview of the results.

| Condition | New results | | Previous upper bound |
|---|---|---|---|
| | Lower bound | Upper bound | |
| $r = 1$ | 2 | 2 | N/A |
| $r = 3$ | 1.5 | $\approx 1.528$ | $\approx 1 + 4.006 \cdot (A - 1)$ |
| general $r$ binary alphabet | $1 + \Theta(1/r^2)$ conjecture: $1 + \Theta(1/\sqrt{r})$ (proved for sampling without replacement) | | $1 + \Theta(1/\sqrt{r})$ |
| general $r$ general alphabet | $1 + \Theta(1/r^2)$ conjecture: $1 + \Theta(1/\sqrt{r})$ (proved for sampling without replacement) | | $1 + \Theta(A/\sqrt{r})$ |

**Lemma 1.** *Suppose that the PTAS of Li* et al. *achieves approximation ratio $\alpha$ for a given set $s_1 \ldots, s_n$ of input sequences, motif length $L$ and sample motif size $r$. Suppose also that the instance of the optimal motif in sequence $s_i$ is $t_i$. Then the PTAS, if run only on the sequences $t_1, \ldots, t_n$, would achieve approximation ratio at least $\alpha$.*

*Proof.* We begin by noting that if $m = L$, the actual problem is trivial: the optimal motif $s^*$ is the consensus of all of the input strings.

However, the PTAS still is well defined in this case, even though the actual optimization problem is trivial. It examines all sets $C$ of $r$ strings, including ones where the same string is chosen multiple times, and for each of them, computes its consensus $M_C$. Then, the central motif $M_{C^*}$ with smallest total Hamming distance to all $s_i$ is chosen as the motif center.

This motif center can be no better than the one found by the PTAS when run on the entire $m$-letter strings, because the set of substrings we have considered in the truncated problem is a subset of the set of substrings we would have examined in the full problem. As such, if the original algorithm would have found a solution whose approximation ratio is $\alpha$, we can only have done as well or worse in the truncated problem.

This lemma is useful because if we can show that, for given values of $L$, $n$ and $r$, and when run only on the optimal motif instances, the PTAS has approximation ratio at most $\beta$, then its approximation ratio on longer strings can still be no worse than $\beta$.

To simplify notation, we assume that the alphabet is $\{0, 1, \ldots, A-1\}$. In the special case we focus on, where $m = L$, we also always renumber the characters in each column, so the consensus for that column is 0. This causes the overall optimal motif to be $s^* = 0^L$. This transformation only works when $m = L$; it does not work when $m > L$.

Finally, we can encounter the problem of ties, that is, a situation when the consensus string $u$ of some collection $C$ is not unique. Consider for example $r = 3$ and input strings 01, 02, 10, and 20. The optimal motif is 00, with cost 4. If $C$

contains the first three strings, the consensus $M_C$ can be any of the strings 00, 01, and 02. The first of them is optimal, but the latter two have cost 5.

It is not realistic to assume that the PTAS will always guess the best of all possible consensus strings; their number can be exponential in $L$. For simplicity, we assume that the PTAS will choose the worst consensus string, and study the performance of this "unlucky" motif finding algorithm, which in our example would choose either 01 or 02.

## 3 Upper Bounds

In this section, we give better worst-case bounds on the approximation guarantee of the algorithm in the cases where $r = 1$ or $r = 3$, corresponding to algorithms with quadratic or quartic bounds on their runtime.

**Theorem 1.** *The approximation ratio of the PTAS is at most 2 for all values of $r$, including $r = 1$, and for any alphabet size $A$.*

*Proof.* Let $c$ be the cost of the optimal motif $0^L$, that is, the total number of non-zero elements in all sequences. Let $a_i$ be the number of non-zero elements in sequence $s_i$. If the PTAS chooses sequence $s_i$ as the motif (which will happen when the $r$ samples from the $n$ sequences are all of $s_i$), the cost will increase by at most $n$ for every column where $s_i$ has non-zero element. Therefore the cost will be at most $c + na_i$. The sum of this quantity over all sequences $s_i$ is $nc + n\sum_{i=1}^{n} a_i = 2nc$. Since the sum of costs for $n$ different potential motifs $s_i$ is at most $2nc$, at least one of these has cost at most $2c$, which means the approximation ratio is at most 2.

**Theorem 2.** *The approximation ratio of the PTAS for $r = 3$ is at most $(64 + 7\sqrt{7})/54 \approx 1.528$ regardless of the size of the alphabet.*

*Proof.* Let $p$ be the proportion of zeroes and $q = (1 - p)$ be the proportion of non-zeroes in the input sequences. The optimal cost is therefore $qnL$. Let $b_j$ be the number of non-zeroes in column $j$.

The algorithm will examine all possible samples consisting of 3 rows, choosing the one with the best consensus string. To get an upper bound, we will consider the expected cost of the consensus string obtained by sampling 3 rows uniformly at random.

For each column, we can estimate the expected cost of the column. The consensus in a particular column will only be non-zero if two or three of the chosen rows contain non-zero entries. If the column contains $b$ non-zero entries, there are $b^3 + 3b^2(n - b)$ such samples. Each of these samples will incur cost of at most $n$ in this column. The consensus will be zero for samples with two or three zeroes (their number is $(n - b)^3 + 3(n - b)^2 b$). Each of these samples will incur cost $b$ in this column.

Thus the expected cost $E(b)$ for a column with $b$ non-zeroes is at most $C(b)/n^3$, where $C(b)$ is the sum of costs over all triples of rows:

$$C(b) = [b^3+3b^2(n-b)]n+[(n-b)^3+3(n-b)^2b]b = 2b^4-5b^3n+3b^2n^2+bn^3. \quad (2)$$

From linearity of expectation, the expected cost over all columns is

$$E(b_1, \ldots, b_L) = \sum_{j=1}^{L} E(b_j) = \frac{1}{n^3} \cdot \sum_{j=1}^{L} C(b_j). \tag{3}$$

There must exist a sample with cost at most $E(b_1, \ldots, b_L)$. Such a sample achieves approximation ratio $E(b_1, \ldots, b_L)/qnL$.

We will prove by induction on $L$ that $E(b_1, \ldots, b_L) \leq HqnL$, where $H = (64 + 7\sqrt{7})/54$. This implies that $H \approx 1.528$ is an upper bound on the approximation ratio for $r = 3$.

For $L = 1$, the approximation ratio is

$$E(qn)/qnL = 2q^3 - 5q^2 + 3q + 1. \tag{4}$$

The maximum of this ratio, which is equal to $H$, is reached when $q = \frac{5-\sqrt{7}}{6}$.

Now, assume that the induction hypothesis is true for $L - 1$. We will prove that it is also true for $L$. The expected cost of the first column is $E(b_1)$, which can be computed with Equation 2 above. By our induction hypothesis, the expected cost of the remaining $L-1$ columns is at most $(qnL - b_1) \cdot H$. Note that $qnL - b_1$ is the optimal cost for the remaining $L - 1$ columns. Therefore:

$$\begin{aligned} E(b_1, \ldots, b_L) &\leq E(b_1) + (qnL - b_1) \cdot H \\ &= \underbrace{\frac{2b^4 - 5b^3n + 3b^2n^2 + (1 - H)bn^3}{n^3}}_{(*)} + HqnL \end{aligned} \tag{5}$$

We want to prove, that (*) is never positive for $b$ in the range $0 \leq b \leq n$. Indeed, (*) can be simplified as $(b/(108n^3)) \cdot (6b - (5 + 2\sqrt{7})n) \cdot (6b - (5 - \sqrt{7})n)^2$. The first and third factors are always non-negative, and the second factor is non-positive for all $b < n$. Therefore the whole term (*) is never positive on the interval.

It is, in fact, possible to easily characterize the "worst-case" scenario that maximizes $E(b_1, \ldots, b_L)$: this is achieved when the non-zero elements are distributed equally among a subset of the columns as follows.

**Lemma 2.** *For a given $q$, $n$, and $L$, $E(b_1, \ldots, b_L)$ is maximized, when for some $k \leq L$, $b_1, \ldots, b_k = 0$, and $b_{k+1} = b_{k+2} = \ldots = b_L \leq n$ (if we allow $b_1, \ldots, b_L$ to be non-integral).*

*Proof.* (by induction on $L$). For $L = 1$, the hypothesis holds trivially.

Let us assume that the hypothesis holds for all $L' < L$. Without loss of generality, we assume that the columns are sorted by $b_j$. If $b_1 = 0$, the hypothesis holds trivially from the induction hypothesis. Let $b_1 > 0$. Then, by the induction hypothesis, all the rest of the columns must by distributed equally (there are no columns with $b_i = 0$, since $b_1$ is the smallest). The cost will be therefore:

$$C(b_1) + (L - 1) \cdot C\left(\frac{qnL - b_1}{L - 1}\right), \tag{6}$$