Oscar H. Ibarra
Hsu-Chun Yen (Eds.)

Implementation and Application of Automata

LNCS 4094

11th International Conference, CIAA 2006
Taipei, Taiwan, August 2006
Proceedings

Springer

Oscar H. Ibarra Hsu-Chun Yen (Eds.)

Implementation and Application of Automata

11th International Conference, CIAA 2006
Taipei, Taiwan, August 21-23, 2006
Proceedings

Springer

Volume Editors

Oscar H. Ibarra
University of California
Department of Computer Science
Santa Barbara, CA 93106, USA
E-mail: ibarra@cs.ucsb.edu

Hsu-Chun Yen
National Taiwan University
Department of Electrical Engineering
Taipei, Taiwan 106, ROC
E-mail: yen@cc.ee.ntu.edu.tw

Library of Congress Control Number: 2006930260

CR Subject Classification (1998): F.1.1-2, F.4.2-3, F.2

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

ISSN 0302-9743
ISBN-10 3-540-37213-X Springer Berlin Heidelberg New York
ISBN-13 978-3-540-37213-4 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11812128 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 4094

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

Editorial Board

Preface

The 11th International Conference on Implementation and Application of Automata (CIAA 2006) was held at the National Taiwan University, Taiwan, August 21–23, 2006.

This volume of *Lecture Notes in Computer Science* contains the papers that were presented at CIAA 2006, as well as the abstracts of the poster papers that were displayed during the conference. The volume also includes the abstracts and extended abstracts of three invited lectures presented by Ming Li, Grzegorz Rozenberg, and Sheng Yu.

The 22 regular papers were selected from 76 submissions covering various topics in the theory, implementation, and applications of automata and related structures. Each submitted paper was reviewed by at least three Program Committee members, with the assistance of referees. The authors of the papers presented here come from the following countries: Austria, Canada, China, Cyprus, Czech Republic, Finland, France, Germany, Hungary, India, Ireland, Italy, The Netherlands, Poland, Spain, Sweden, Taiwan, UK, and USA.

We wish to thank all who have made this meeting possible: the authors for submitting papers, the Program Committee members and external referees (listed in the proceedings) for their excellent work, and our three invited speakers. Finally, we wish to express our sincere appreciation to the sponsors, local organizers, proceedings Chair, the editors of the *Lecture Notes in Computer Science* series and Springer, in particular Alfred Hofmann, for their help in publishing this volume.

August 2006

Oscar H. Ibarra
Hsu-Chun Yen

Organization

Program Committee

Marie-Pierre Béal	University of Marne-la-Vallee, France
Cristian Calude	University of Auckland, New Zealand
Jean-Marc Champarnaud	Université de Rouen, France
Erzsébet Csuhaj-Varjú	Hungarian Academy of Sciences, Hungary
Jürgen Dassow	University of Magdeburg, Germany
Jacques Farré	Université de Nice - Sophia Antipolis, France
Rudolf Freund	Vienna University of Technology, Austria
Jozef Gruska	Masaryk University, Czech Republic
Tero Harju	University of Turku, Finland
Jan Holub	Czech Technical University, Czech Republic
Markus Holzer	Technische Universität München, Germany
Oscar H. Ibarra (Co-chair)	University of California, Santa Barbara, USA
Masami Ito	Kyoto Sangyo University, Japan
Kazuo Iwama	Kyoto University, Japan
Juhani Karhumäki	University of Turku, Finland
Hsueh-I Lu	National Taiwan University, Taiwan, ROC
Denis Maurel	Université de Tours, France
Mehryar Mohri	New York University, USA
Andrei Paun	Louisiana Tech, USA
Bala Ravikumar	Sonoma State University, USA
Wojciech Rytter	Warsaw University, Poland, and NJIT, USA
Jacques Sakarovitch	CNRS/ENST, Paris, France
Kai Salomaa	Queen's University, Canada
Pierluigi San Pietro	Politecnico di Milano, Italy
Giora Slutzki	Iowa State University, USA
Bow-Yaw Wang	Academia Sinica, Taiwan, ROC
Farn Wang	National Taiwan University, Taiwan, ROC
Bruce W. Watson	University of Pretoria, South Africa
Hsu-Chun Yen (Co-chair)	National Taiwan University, Taiwan, ROC
Sheng Yu	University of Western Ontario, Canada

Additional Reviewers

Cyril Allauzen
Manuel Baclet
Miroslav Balík
Kamel Barkaoui
Samik Basu
Nicolas Bedon
Jean Berstel
Franziska Biegler
Renaud Blanch
Henning Bordihn
Ján Bouda
Bohuš Brim
Luboš Brim
Lin-Zan Cai
Cezar Campeanu
Chih-Hong Cheng
Loek Cleophas
Mark Daley
Akim Demaille
Michael Domaratzki
Ming-Chang Dong
Rémi Forax
Nathalie Friburger
Hermann Gruber
Vesa Halava
Geza Horvath
Geng-Dian Huang
Jarkko Kari
Tomi Kärki
Martin Kutrib
Eric Laporte
Martin Leucker
Chih-Wei Lin
Li-Ping Lin
Sylvain Lombardy
Ian McQuillan
Alexander Meduna
Angelo Morzenti
Gonzalo Navarro
Ernest Ketcha Ngassam
Harumichi Nishimura
Alexander Okhotin
Maciej Pilichowski
Libor Polak
Gheorge Rahonis
Ashish Rastogi
Chantal Reynaud
Matteo Rossi
Nicolae Santean
Sylvain Schmitz
Luc Segoufin
Jason R. Smith
Paola Spoletini
Alain Terlutte
Ming-Hsien Tsai
Ladislav Vagner
Thomas Worsch
Shigeru Yamashita
Artur Zaroda

Organizing Committee

Tsan-Sheng Hsu (Co-chair)	Academia Sinica, Taiwan, ROC
Oscar H. Ibarra	University of California, Santa Barbara, USA
Hsueh-I Lu	National Taiwan University, Taiwan, ROC
Bow-Yaw Wang	Academia Sinica, Taiwan, ROC
Farn Wang	National Taiwan University, Taiwan, ROC
Hsu-Chun Yen (Co-chair)	National Taiwan University, Taiwan, ROC

Proceedings Committee

Oscar H. Ibarra	University of California, Santa Barbara, USA
Hsu-Chun Yen	National Taiwan University, Taiwan, ROC

Publicity Committee

Hsueh-I Lu	National Taiwan University, Taiwan, ROC
Farn Wang	National Taiwan University, Taiwan, ROC

Steering Committee

Jean-Marc Champarnaud	Université de Rouen, France
Oscar H. Ibarra	University of California, Santa Barbara, USA
Denis Maurel	Université de Tours, France
Derick Wood	Hong Kong Univ. of Science and Technology, Hong Kong, China
Sheng Yu (Chair)	University of Western Ontario, Canada

Sponsoring Institutions

National Taiwan University, Taiwan, ROC
Center for Information and Electronic Technologies, NTU, Taiwan, ROC
National Science Council, Taiwan, ROC
Ministry of Education, Taiwan, ROC
Academia Sinica, Taiwan, ROC
University of California, Santa Barbara, USA

Lecture Notes in Computer Science

For information about Vols. 1–4007

please contact your bookseller or Springer

Vol. 4127: E. Damiani, P. Liu (Eds.), Data and Applications Security XX. X, 319 pages. 2006.

Vol. 4121: A. Biere, C.P. Gomes (Eds.), Theory and Applications of Satisfiability Testing - SAT 2006. XII, 438 pages. 2006.

Vol. 4112: D.Z. Chen, D. T. Lee (Eds.), Computing and Combinatorics. XIV, 528 pages. 2006.

Vol. 4108: J.M. Borwein, W.M. Farmer (Eds.), Mathematical Knowledge Management. VIII, 295 pages. 2006. (Sublibrary LNAI).

Vol. 4106: T.R. Roth-Berghofer, M.H. Göker, H. A. Güvenir (Eds.), Advances in Case-Based Reasoning. XIV, 566 pages. 2006. (Sublibrary LNAI).

Vol. 4099: Q. Yang, G. Webb (Eds.), PRICAI 2006: Trends in Artificial Intelligence. XXVIII, 1263 pages. 2006. (Sublibrary LNAI).

Vol. 4098: F. Pfenning (Ed.), Term Rewriting and Applications. XIII, 415 pages. 2006.

Vol. 4097: X. Zhou, O. Sokolsky, L. Yan, E.-S. Jung, Z. Shao, Y. Mu, D.C. Lee, D. Kim, Y.-S. Jeong, C.-Z. Xu (Eds.), Emerging Directions in Embedded and Ubiquitous Computing. XXVII, 1034 pages. 2006.

Vol. 4096: E. Sha, S.-K. Han, C.-Z. Xu, M.H. Kim, L.T. Yang, B. Xiao (Eds.), Embedded and Ubiquitous Computing. XXIV, 1170 pages. 2006.

Vol. 4094: O. H. Ibarra, H.-C. Yen (Eds.), Implementation and Application of Automata. XIII, 291 pages. 2006.

Vol. 4092: J. Lang, F. Lin, J. Wang (Eds.), Knowledge Science, Engineering and Management. XV, 664 pages. 2006. (Sublibrary LNAI).

Vol. 4090: S. Spaccapietra, K. Aberer, P. Cudré-Mauroux (Eds.), Journal on Data Semantics VI. XI, 211 pages. 2006.

Vol. 4088: Z.-Z. Shi, R. Sadananda (Eds.), Agent Computing and Multi-Agent Systems. XVII, 827 pages. 2006. (Sublibrary LNAI).

Vol. 4079: S. Etalle, M. Truszczyński (Eds.), Logic Programming. XIV, 474 pages. 2006.

Vol. 4077: M.-S. Kim, K. Shimada (Eds.), Advances in Geometric Modeling and Processing. XVI, 696 pages. 2006.

Vol. 4076: F. Hess, S. Pauli, M. Pohst (Eds.), Algorithmic Number Theory. X, 599 pages. 2006.

Vol. 4075: U. Leser, F. Naumann, B. Eckman (Eds.), Data Integration in the Life Sciences. XI, 298 pages. 2006. (Sublibrary LNBI).

Vol. 4074: M. Burmester, A. Yasinsac (Eds.), Secure Mobile Ad-hoc Networks and Sensors. X, 193 pages. 2006.

Vol. 4073: A. Butz, B. Fisher, A. Krüger, P. Olivier (Eds.), Smart Graphics. XI, 263 pages. 2006.

Vol. 4072: M. Harders, G. Székely (Eds.), Biomedical Simulation. XI, 216 pages. 2006.

Vol. 4071: H. Sundaram, M. Naphade, J.R. Smith, Y. Rui (Eds.), Image and Video Retrieval. XII, 547 pages. 2006.

Vol. 4070: C. Priami, X. Hu, Y. Pan, T.Y. Lin (Eds.), Transactions on Computational Systems Biology V. IX, 129 pages. 2006. (Sublibrary LNBI).

Vol. 4069: F.J. Perales, R.B. Fisher (Eds.), Articulated Motion and Deformable Objects. XV, 526 pages. 2006.

Vol. 4068: H. Schärfe, P. Hitzler, P. Øhrstrøm (Eds.), Conceptual Structures: Inspiration and Application. XI, 455 pages. 2006. (Sublibrary LNAI).

Vol. 4067: D. Thomas (Ed.), ECOOP 2006 – Object-Oriented Programming. XIV, 527 pages. 2006.

Vol. 4066: A. Rensink, J. Warmer (Eds.), Model Driven Architecture – Foundations and Applications. XII, 392 pages. 2006.

Vol. 4065: P. Perner (Ed.), Advances in Data Mining. XI, 592 pages. 2006. (Sublibrary LNAI).

Vol. 4064: R. Büschkes, P. Laskov (Eds.), Detection of Intrusions and Malware & Vulnerability Assessment. X, 195 pages. 2006.

Vol. 4063: I. Gorton, G.T. Heineman, I. Crnkovic, H.W. Schmidt, J.A. Stafford, C.A. Szyperski, K. Wallnau (Eds.), Component-Based Software Engineering. XI, 394 pages. 2006.

Vol. 4062: G. Wang, J.F. Peters, A. Skowron, Y. Yao (Eds.), Rough Sets and Knowledge Technology. XX, 810 pages. 2006. (Sublibrary LNAI).

Vol. 4061: K. Miesenberger, J. Klaus, W. Zagler, A. Karshmer (Eds.), Computers Helping People with Special Needs. XXIX, 1356 pages. 2006.

Vol. 4060: K. Futatsugi, J.-P. Jouannaud, J. Meseguer (Eds.), Algebra, Meaning and Computation. XXXVIII, 643 pages. 2006.

Vol. 4059: L. Arge, R. Freivalds (Eds.), Algorithm Theory – SWAT 2006. XII, 436 pages. 2006.

Vol. 4058: L.M. Batten, R. Safavi-Naini (Eds.), Information Security and Privacy. XII, 446 pages. 2006.

Vol. 4057: J.P. W. Pluim, B. Likar, F.A. Gerritsen (Eds.), Biomedical Image Registration. XII, 324 pages. 2006.

Vol. 4056: P. Flocchini, L. Gąsieniec (Eds.), Structural Information and Communication Complexity. X, 357 pages. 2006.

Vol. 4055: J. Lee, J. Shim, S.-g. Lee, C. Bussler, S. Shim (Eds.), Data Engineering Issues in E-Commerce and Services. IX, 290 pages. 2006.

Vol. 4054: A. Horváth, M. Telek (Eds.), Formal Methods and Stochastic Models for Performance Evaluation. VIII, 239 pages. 2006.

Vol. 4053: M. Ikeda, K.D. Ashley, T.-W. Chan (Eds.), Intelligent Tutoring Systems. XXVI, 821 pages. 2006.

Vol. 4052: M. Bugliesi, B. Preneel, V. Sassone, I. Wegener (Eds.), Automata, Languages and Programming, Part II. XXIV, 603 pages. 2006.

Vol. 4051: M. Bugliesi, B. Preneel, V. Sassone, I. Wegener (Eds.), Automata, Languages and Programming, Part I. XXIII, 729 pages. 2006.

Vol. 4049: S. Parsons, N. Maudet, P. Moraitis, I. Rahwan (Eds.), Argumentation in Multi-Agent Systems. XIV, 313 pages. 2006. (Sublibrary LNAI).

Vol. 4048: L. Goble, J.-J.C.. Meyer (Eds.), Deontic Logic and Artificial Normative Systems. X, 273 pages. 2006. (Sublibrary LNAI).

Vol. 4047: M. Robshaw (Ed.), Fast Software Encryption. XI, 434 pages. 2006.

Vol. 4046: S.M. Astley, M. Brady, C. Rose, R. Zwiggelaar (Eds.), Digital Mammography. XVI, 654 pages. 2006.

Vol. 4045: D. Barker-Plummer, R. Cox, N. Swoboda (Eds.), Diagrammatic Representation and Inference. XII, 301 pages. 2006. (Sublibrary LNAI).

Vol. 4044: P. Abrahamsson, M. Marchesi, G. Succi (Eds.), Extreme Programming and Agile Processes in Software Engineering. XII, 230 pages. 2006.

Vol. 4043: A.S. Atzeni, A. Lioy (Eds.), Public Key Infrastructure. XI, 261 pages. 2006.

Vol. 4042: D. Bell, J. Hong (Eds.), Flexible and Efficient Information Handling. XVI, 296 pages. 2006.

Vol. 4041: S.-W. Cheng, C.K. Poon (Eds.), Algorithmic Aspects in Information and Management. XI, 395 pages. 2006.

Vol. 4040: R. Reulke, U. Eckardt, B. Flach, U. Knauer, K. Polthier (Eds.), Combinatorial Image Analysis. XII, 482 pages. 2006.

Vol. 4039: M. Morisio (Ed.), Reuse of Off-the-Shelf Components. XIII, 444 pages. 2006.

Vol. 4038: P. Ciancarini, H. Wiklicky (Eds.), Coordination Models and Languages. VIII, 299 pages. 2006.

Vol. 4037: R. Gorrieri, H. Wehrheim (Eds.), Formal Methods for Open Object-Based Distributed Systems. XVII, 474 pages. 2006.

Vol. 4036: O. H. Ibarra, Z. Dang (Eds.), Developments in Language Theory. XII, 456 pages. 2006.

Vol. 4035: T. Nishita, Q. Peng, H.-P. Seidel (Eds.), Advances in Computer Graphics. XX, 771 pages. 2006.

Vol. 4034: J. Münch, M. Vierimaa (Eds.), Product-Focused Software Process Improvement. XVII, 474 pages. 2006.

Vol. 4033: B. Stiller, P. Reichl, B. Tuffin (Eds.), Performability Has its Price. X, 103 pages. 2006.

Vol. 4032: O. Etzion, T. Kuflik, A. Motro (Eds.), Next Generation Information Technologies and Systems. XIII, 365 pages. 2006.

Vol. 4031: M. Ali, R. Dapoigny (Eds.), Advances in Applied Artificial Intelligence. XXIII, 1353 pages. 2006. (Sublibrary LNAI).

Vol. 4029: L. Rutkowski, R. Tadeusiewicz, L.A. Zadeh, J. Zurada (Eds.), Artificial Intelligence and Soft Computing – ICAISC 2006. XXI, 1235 pages. 2006. (Sublibrary LNAI).

Vol. 4028: J. Kohlas, B. Meyer, A. Schiper (Eds.), Dependable Systems: Software, Computing, Networks. XII, 295 pages. 2006.

Vol. 4027: H.L. Larsen, G. Pasi, D. Ortiz-Arroyo, T. Andreasen, H. Christiansen (Eds.), Flexible Query Answering Systems. XVIII, 714 pages. 2006. (Sublibrary LNAI).

Vol. 4026: P.B. Gibbons, T. Abdelzaher, J. Aspnes, R. Rao (Eds.), Distributed Computing in Sensor Systems. XIV, 566 pages. 2006.

Vol. 4025: F. Eliassen, A. Montresor (Eds.), Distributed Applications and Interoperable Systems. XI, 355 pages. 2006.

Vol. 4024: S. Donatelli, P. S. Thiagarajan (Eds.), Petri Nets and Other Models of Concurrency - ICATPN 2006. XI, 441 pages. 2006.

Vol. 4021: E. André, L. Dybkjær, W. Minker, H. Neumann, M. Weber (Eds.), Perception and Interactive Technologies. XI, 217 pages. 2006. (Sublibrary LNAI).

Vol. 4020: A. Bredenfeld, A. Jacoff, I. Noda, Y. Takahashi (Eds.), RoboCup 2005: Robot Soccer World Cup IX. XVII, 727 pages. 2006. (Sublibrary LNAI).

Vol. 4019: M. Johnson, V. Vene (Eds.), Algebraic Methodology and Software Technology. XI, 389 pages. 2006.

Vol. 4018: V. Wade, H. Ashman, B. Smyth (Eds.), Adaptive Hypermedia and Adaptive Web-Based Systems. XVI, 474 pages. 2006.

Vol. 4017: S. Vassiliadis, S. Wong, T.D. Hämäläinen (Eds.), Embedded Computer Systems: Architectures, Modeling, and Simulation. XV, 492 pages. 2006.

Vol. 4016: J.X. Yu, M. Kitsuregawa, H.V. Leong (Eds.), Advances in Web-Age Information Management. XVII, 606 pages. 2006.

Vol. 4014: T. Uustalu (Ed.), Mathematics of Program Construction. X, 455 pages. 2006.

Vol. 4013: L. Lamontagne, M. Marchand (Eds.), Advances in Artificial Intelligence. XIII, 564 pages. 2006. (Sublibrary LNAI).

Vol. 4012: T. Washio, A. Sakurai, K. Nakajima, H. Takeda, S. Tojo, M. Yokoo (Eds.), New Frontiers in Artificial Intelligence. XIII, 484 pages. 2006. (Sublibrary LNAI).

Vol. 4011: Y. Sure, J. Domingue (Eds.), The Semantic Web: Research and Applications. XIX, 726 pages. 2006.

Vol. 4010: S. Dunne, B. Stoddart (Eds.), Unifying Theories of Programming. VIII, 257 pages. 2006.

Vol. 4009: M. Lewenstein, G. Valiente (Eds.), Combinatorial Pattern Matching. XII, 414 pages. 2006.

Vol. 4008: J.C. Augusto, C.D. Nugent (Eds.), Designing Smart Homes. XI, 183 pages. 2006. (Sublibrary LNAI).

Table of Contents

Invited Lectures

Technical Contributions

Poster Abstracts

Information Distance and Its Applications

Ming Li

School of Computer Science, University of Waterloo, Waterloo, Ont. N2L 3G1, Canada
mli@uwaterloo.ca
http://www.cs.uwaterloo.ca/~mli

Abstract. We summarize the recent developments of a general theory of information distance and its applications in whole genome phylogeny, document comparison, internet query-answer systems, and many other data mining tasks. We also solve an open problem regarding the universality of the normalized information distance.

1 Introduction

We live in an information society. Internet has created the cyber, or information, space. In the classical Newton world, we know how to measure physical distances. Have you thought about the equally fundamental question of how to measure the "information distance" between two objects: two documents, two letters, two emails, two music scores, two languages, two programs, two pictures, two systems, or two genomes? Such a measurement should not be application dependent. Just like in the classical world, we do not measure distances sometimes by the amount of time a bird flies and sometimes by the number of pebbles lining up on the Santa Barbara beach.

A good information distance metric should not only be application-independent but also universally minorize all other "reasonable" definitions.

The task of a universal definition of information distance is illusive. Traditional distances such as the Euclidean distance or the Hamming distance obviously fail for even trivial examples. For instance, we (human) perceive a positive photo to be similar to its negative print, while their Hamming distance is the largest. In fact, for any computable distance, we can always find such counterexamples. Furthermore, when we wish to adopt a metric to be the universal standard of information distance, we must justify it. It should not be out of thin air. It should not be from a specific application. It should not require amendaments for different applications. It should be as good as any definition for any application, in some sense.

From a simple and accepted assumption in thermodynamics, we have derived such a universal information distance [2,18,19] and a general method to measure similarities between two sequences [18,19]. The theory has been initially applied to alignment free whole genome phylogeny [18], chain letter history [3], language history [4,19], plagiarism detection [5], and more recently to music classification [9], parameter-free data mining paradigm [13], internet knowledge discovery [8], among many recent applications.

O.H. Ibarra and H.-C. Yen (Eds.): CIAA 2006, LNCS 4094, pp. 1–9, 2006.

2 A Theory of Information Distance

Given a binary string x, the Kolmogorov complexity of x condition on y, $K(x|y)$, is the length of the shortest program that outputs x with input y. When $y = \epsilon$, we write $K(x|\epsilon)$ as $K(x)$. For formal definitions and a comprehensive study of Kolmogorov complexity, see [20]. What would be a good departure point for defining "information distance" between two sequences? What should be the properties it must satisfy? The second question is easy to answer. We can use our common sense of a metric: (a) It must be symmetric; (b) It should satisfy the triangle inequality; (c) The distance of any sequence x to itself is 0, and positive otherwise.

To answer the first question, in early 1990's, we have studied the energy cost of convertion between two strings x and y. Over half a century ago, John von Neumann hypothesized that performing 1 bit of information processing costs $1KT$ of energy, where K is the Boltzmann's constant and T is the room temperature. Observing that reversable computations can be done for free, in early 1960's Rolf Landauer revised von Neumann's proposal to hold only for irreversible computations. We thought about using the minimum energy needed to convert between x and y to define their distance, as it is an objective measure. Thus, if you have x and wish to erase it, then you can reversibly convert it to x^*, x's shortest effective description, then erase $|x^*|$. Only the process of erasing $|x^*|$ bits is irreversible computation. Carrying on from this line of thinking, we [2] have defined the energy to convert between x and y to be the length of shortest program converting x to y and vice versa. That is, with respect to a universal Turing machine U, the cost of converting between x and y is:

$$E(x, y) = \min\{|p| : U(x, p) = y,\ U(y, p) = x\} \tag{1}$$

A natural upper bound for $E(x, y)$ is $K(x|y) + K(y|x)$. Using this (and other reasons), we have defined the sum distance in [2]:

$$d_{\text{sum}}(x, y) = K(x|y) + K(y|x).$$

However, the following theorem proved in [2] was a surprise.

Theorem 1. $E(x, y) = \max\{K(x|y), K(y|x)\}$.

Thus, we have defined the max distance:

$$d_{\max}(x, y) = \max\{K(x|y), K(y|x)\}.$$

Both distances are shown to satisfy the basic distance requirements such as positivity, symmetricity, triangle inequality, in [2]. We have further shown that $d_{\max}$ and d_{sum} minorizes all other distances that are computable and satisfies some reasonable density condition that within distance k to any string x, there are at most 2^k other strings. Formally, a distance D is admissible if

$$\sum_y 2^{-D(x,y)} \leq 1. \tag{2}$$

Then we proved that for any admissible computable distance D, there is a constant c, for all x, y, $d_{\max}(x,y) \leq D(x,y) + c$. Put it bluntly, if any other distance recovers some regularity between two sequences, so will $d_{\max}$.

The remaining question is to demonstrate that such distances are useful. However when we [18] tried to use our information distances, d_{sum} or $d_{\max}$, to measure similarity between genomes in 1998, we were in trouble. *E. coli* and *H. influenza* are sister species but their genome lengths defer greatly. The *E. coli* genome is about 5 megabases whereas the *H. influenza* genome is only 1.8 megabase long. $d_{\max}$ or d_{sum} between the two genomes are predominated by genome length difference rather than the amount of information they share. Such a measure trivially classifies *H. influenza* to be closer to a more remote species of similar genome length such as *A. fulgidus* (2.18 megabases) than to *E. coli*.

In order to solve this problem, we introduced "shared information distance" in [18]:

$$d_{\text{share}}(x,y) = 1 - \frac{K(x) - K(x|y)}{K(xy)}.$$

where $K(x) - K(x|y)$ is mutual information between sequences x and y [20]. We proved the basic distance metric requirements such as symmetry and triangle inequality, and have demonstrated its successful application in whole genome phylogeny in [18]. It turns out that d_{share} is equivalent to

$$\frac{K(x|y) + K(y|x)}{K(xy)}.$$

Thus, it can be viewed as the normalized sum distance. Hence, it becomes natural to normalize the optimal max distance in [19]:

$$d(x,y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}} \tag{3}$$

We have called $d(x,y)$ the "normalized information distance" proved metricity properties similar to that of normalized sum distance.

However, a key issue of universality of the normalized information distance, all versions, has remained unsolved. The similar proof for $d_{\max}$ and d_{sum} does not work any more for normalized distances d_{share} and d. In [18], in order to prove the universality statement, we were only able to prove a very weak statement: for any computable distance D, there is a constance $c \leq 2$ such that, with probability 1, for all sequences x and y, $d(x,y) \leq cD(x,y)$. This seemingly innocent statement is actually begging the question: the random sequences have probability 1, whereas it is non-random sequences we are interested in measuring and this statement says nothing about them.

In our second paper [19], we have tried to avoid this problem by rescaling the density conditions changing from

$$|\{y : |y| = n \text{ and } D(x,y) \leq d \leq 1\}| \leq 2^{dn} \tag{4}$$

in [18] to

$$|\{y : |y| = n \text{ and } D(x,y) \leq d \leq 1\}| \leq 2^{dK(x)} \tag{5}$$

However it turns out that Formula (5) is so restrictive that no reasonable distances can satisfy such requirement except our own normalized information distance. Thus the universality statement is again meaningless. Cilibrasi and Vitanyi have tried to further change the definition of normalized admissible distances [9].

3 Fixing the Theory

We did not need to change the definition after all. Using the the original defition of [18], Formula (4), we now prove the full universality theorem, removing the "with probability 1" condition.

Theorem 2. *For any computable distance D, satisfying density requirement (4), for all sequence x and y, $d(x,y) \leq D(x,y) + O(\log n / \max\{K(x), K(y)\})$.*

Proof. For any binary sequence x of length n, Muchnik [21] proved that there exists a (shortest) program x^*, such that $|x^*| = K(x)$, $K(x|x^*) = O(\log n)$ and $K(x^*|x) = O(\log n)$. That is, x^* is a shortest program for x and it does not contain too much extra information unrelated to x.

For any sequences x and y of length up to n, there are x^* and y^* satisfying Muchnick's theorem. Given y, we can compute y^* using $O(\log n)$ information. Then using $K(x^*|y^*)$ information, we can compute x^*, which in turn gives x with $O(\log n)$ information. We have proved:

$$K(x|y) \leq K(x^*|y^*) + O(\log n). \tag{6}$$

The equality actually holds. More general exploration of this is in [12].

Applying Inequality (6),

$$\begin{aligned} d(x,y) &= \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}} \\ &\leq \frac{\max\{K(x^*|y^*), K(y^*|x^*)\} + O(\log n)}{\max\{K(x), K(y)\}} \end{aligned} \tag{7}$$

Given D, using the density property of Formula (4) and the computability of D, we know $K(x^*|y^*) \leq D(x,y)|x^*|$ and $K(y^*|x^*) \leq D(y,x)|y^*|$. Thus, from Formula (7) and symmetry of D, we have,

$$\begin{aligned} d(x,y) &\leq \frac{\max\{K(x^*|y^*), K(y^*|x^*)\} + O(\log n)}{\max\{K(x), K(y)\}} \\ &\leq \frac{\max\{D(x,y)|x^*|, D(x,y)|y^*|\} + O(\log n)}{\max\{|x^*|, |y^*|\}} \\ &\leq D(x,y) + O(\log n / \max\{K(x), K(y)\}). \end{aligned}$$

Similar proof gives the universality statement for the normalized sum distance (d_{share}), defined in [18].