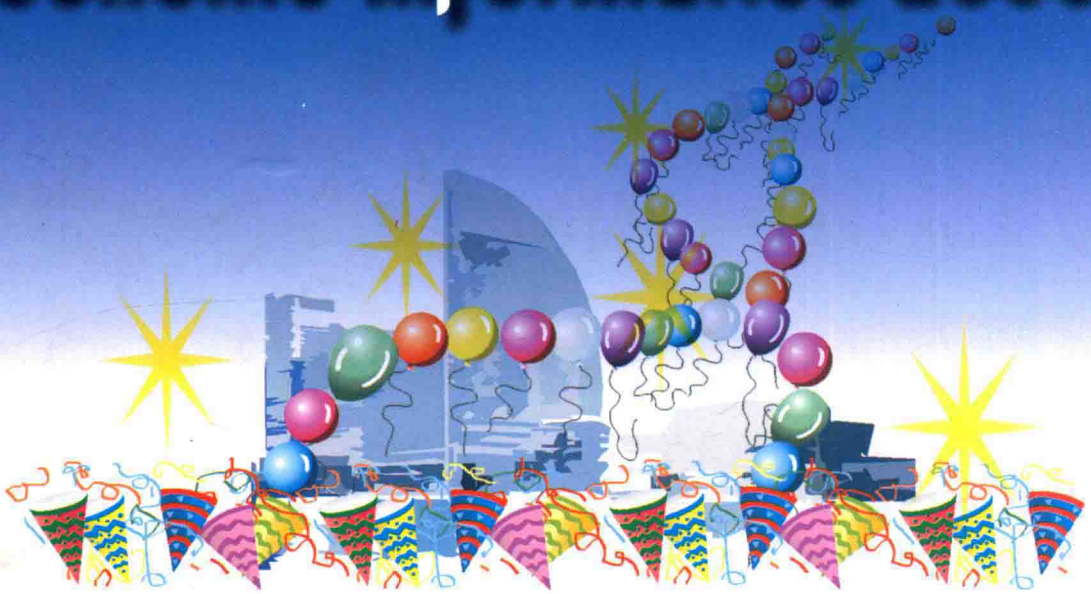
Genome Informatics Series Vol. 23 ISSN: 0919-9454

Genome Informatics 2009



Shinichi Morishita • Sang Yup Lee • Yasubumi Sakakibara
Editors

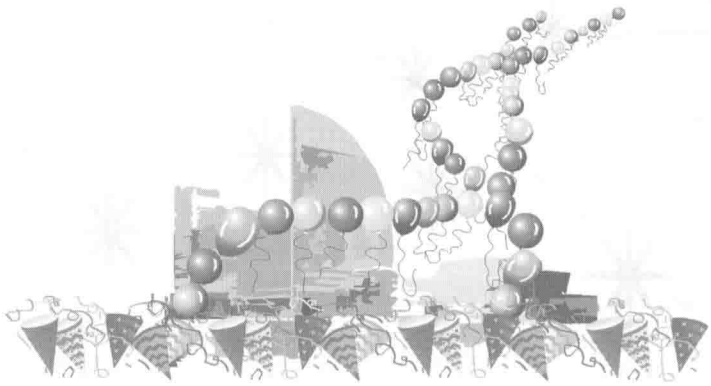
Imperial College Press

Genome Informatics Series Vol. 23 ISSN: 0919-9454

Genome Informatics 2009

Proceedings of the 20th International Conference

Pacifico Yokohama, Japan 14 – 16 December 2009



Editors

Shinichi Morishita

University of Tokyo, Japan

Sang Yup Lee

Korea Advanced Institute of Science & Technology, Korea

Yasubumi Sakakibara

Keio University, Japan

ICP

Imperial College Press

Published by

Imperial College Press
57 Shelton Street
Covent Garden
London WC2H 9HE

Distributed by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

GENOME INFORMATICS 2009
Proceedings of the 20th International Conference (GIW 2009)

ISBN-13 978-1-84816-562-5
ISBN-10 1-84816-562-5

Printed by FuIsland Offset Printing (S) Pte Ltd, Singapore

Genome Informatics 2009

GENOME INFORMATICS SERIES (GIS)

ISSN: 0919-9454

The Genome Informatics Series publishes peer-reviewed papers presented at the International Conference on Genome Informatics (GIW) and some conferences on bioinformatics. The Genome Informatics Series is indexed in MEDLINE.

No.	Title	Year	ISBN Cl./Pa.
1	Genome Informatics Workshop I	1990	(in Japanese)
2	Genome Informatics Workshop II	1991	(in Japanese)
3	Genome Informatics Workshop III	1992	(in Japanese)
4	Genome Informatics Workshop IV	1993	4-946443-20-7
5	Genome Informatics Workshop 1994	1994	4-946443-24-X
6	Genome Informatics Workship 1995	1995	4-946443-33-9
7	Genome Informatics 1996	1996	4-946443-37-1
8	Genome Informatics 1997	1997	4-946443-47-9
9	Genome Informatics 1998	1998	4-946443-52-5
10	Genome Informatics 1999	1999	4-946443-59-2
11	Genome Informatics 2000	2000	4-946443-65-7
12	Genome Informatics 2001	2001	4-946443-72-X
13	Genome Informatics 2002	2002	4-946443-79-7
14	Genome Informatics 2003	2003	4-946443-82-7
15	Genome Informatics 2004 Vol. 15, No. 1	2004	4-946443-88-6
16	Genome Informatics 2004 Vol. 15, No. 2	2004	4-946443-91-6
17	Genome Informatics 2005 Vol. 16, No. 1	2005	4-946443-93-2
18	Genome Informatics 2005 Vol. 16, No. 2	2005	4-946443-96-7
19	Genome Informatics 2006 Vol. 17, No. 1	2006	4-946443-97-5
20	Genome Informatics 2006 Vol. 17, No. 2	2006	4-946443-99-1
21	Genome Informatics 2007 Vol. 18	2007	978-1-86094-991-3
22	Genome Informatics 2007 Vol. 19	2007	978-1-86094-984-5
23	Genome Informatics 2008 Vol. 20	2008	978-1-84816-299-0
24	Genome Informatics 2008 Vol. 21	2008	978-1-84816-331-7
25	Genome Informatics 2009 Vol. 22	2009	978-1-84816-569-4
26	Genome Informatics 2009 Vol. 23	2009	978-1-84816-562-5

PREFACE

This issue of Genome Informatics contains papers presented at the Twentieth International Conference on Genome Informatics (GIW 2009) held in Yokohama, Japan from December 14th to 16th, 2009.

The first Genome Informatics Workshop (GIW) was held in Tokyo in 1990, the dawn of human genome sequencing. Remarkably, the invited talks by Akiyoshi Wada, Ross Overbeek, and Yoshiyuki Sakaki that year were all on the subject of the computational support for genome sequencing. Since then, GIW has provided unique opportunities to encourage bioinformatics and create bridges between theory and experiments, academia and industry, and East and West. GIW is the longest running international bioinformatics conference.

The 20th International Conference on Genome Informatics (GIW 2009) was held at PACIFICO Yokohama Convention Center, Japan, on December 14-16, 2009. We accepted 18 papers from the 39 submissions. The two best papers were:

- C. Nelson Hayes, Diego Diez, Nicolas Joannin, Minoru Kanehisa, Mats Wahlgren, Craig E. Wheelock, and Susumu Goto. "Tools for investigating mechanisms of antigenic variation: new extensions to varDB."
- Kouichi Kimura and Asako Koike. "Localized suffix array and its application to genome mapping problems for paired-end short reads."

In addition, this book contains abstracts from the five invited speakers: Sean Eddy, HHMI's Janelia Farm (USA), Minoru Kanehisa, Kyoto University (Japan), Sang Yup Lee, KAIST (Korea), Hideyuki Okano, Keio University, (Japan), and Mark Ragan, University of Queensland (Australia).

The electronic versions of all these papers in this issue are also freely available from the website of the Japanese Society for Bioinformatics (JSBi) (`http://www.jsbi.org/journal.html`).

Shinichi Morishita
Sang Yup Lee
GIW 2009 Program Committee Co-Chairs

Yasubumi Sakakibara
GIW 2009 Conference Chair

ACKNOWLEDGMENTS

First of all, we would like to thank the authors for their efforts in preparing their manuscripts. We also appreciate the great efforts made by the program committee members and the external reviewers in the reviewing process. We further acknowledge the assistance from the local organizing committee members for arranging the conference venue. Special thanks are due to the conference editorial staff for their excellent work, especially, Emi Ikeda, Ayumu Saito, and Asako Suzuki. Finally, GIW 2009 was sponsored by the Human Genome Center of the Institute of Medical Science at the University of Tokyo, and the Japanese Society for Bioinformatics.

PROGRAM COMMITTEE

Sang Yup Lee	– KAIST, Korea, PC Co-Chair
Shin-ichi Morishita	– University of Tokyo, Japan, PC Co-Chair
Cathy Abbott	– Flinders University, Australia
Jonathan Arthur	– University of Sydney, Australia
Vladimir Bajic	– SANBI, South Africa
Christopher Baker	– Institute for Infocomm Research, Singapore
Guillaume Bourque	– Genome Institute of Singapore, Singapore
Jung-Hsien Chiang	– National Cheng Kung University, Taiwan
Francis Chin	– The University of Hong Kong, Hong Kong
Peter Clote	– Boston College, USA
Aaron Darling	– University of California, Davis, USA
Bhaskar DasGupta	– University of Illinois, USA
Colin Dewey	– University of Wisconsin, USA
Chris Ding	– University of Texas at Arlington, USA
Wen-Lian Hsu	– Academia Sinica, Taiwan
Seiya Imoto	– University of Tokyo, Japan
Minoru Kanehisa	– Kyoto University, Japan
Uri Keich	– Cornell University, USA
Daisuke Kihara	– Purdue University, USA
Dong-Yup Lee	– Bioprocessing Institute & National University of Singapore, Singapore
Ming Li	– University of Waterloo, Canada
Frederique Lisacek	– Swiss Institute of Bioinformatics, Switzerland
Hiroshi Mamitsuka	– Kyoto University, Japan
Aleksandar Milosavljevic	– Baylor College of Medicine, USA
Satoru Miyano	– University of Tokyo, Japan
Bernard Moret	– Swiss Federal Institute of Technology, Switzerland
Akihiro Nakaya	– University of Tokyo, Japan
See-Kiong Ng	– Institute for Infocomm Research, Singapore
William Noble	– University of Washington, USA
Laxmi Parida	– IBM T.J. Watson Research Center, USA
Ron Pinter	– Technion, Israel
Shoba Ranganathan	– Macquarie University, Australia
Rintaro Saito	– Keio University, Japan
Christian Schoenbach	– Nanyang Technological University, Singapore

Tetsuo Shibuya	– University of Tokyo, Japan
Mona Singh	– Princeton University, USA
Wing-Kin Sung	– National University of Singapore, Singapore
Koji Tsuda	– Computational Biology Research Center, Japan
Gabriel Valiente	– Technical University of Catalonia, Spain
Lusheng Wang	– The City University of Hong Kong, Hong Kong
Gwan-Su Yi	– Information and Communication University, Korea
Mohammed Zaki	– Rensselaer Polytechnic Institute, USA

CO-REVIEWERS

Ai Muto	Akitsugu Suga	Andre Fujita
Caster Chen	Hsin-Nan Lin	Itai Sharon
Kazushi Hiranuka	Masaaki Kotera	Masahiro Hattori
Noa Tzunz-Henig	Qiu Long	Rashmi Hegde
Sheila Reynolds	Toshiaki Tokimatsu	Victor Tong Joo Chuan
Wataru Honda	Yi-Wen Yang	Yosuke Nishimura
Yugo Shimizu	Yuki Moriya	Zeyar Aung

JSBI SPECIAL SESSION COMMITTEE

Tsuyoshi Shirai	– Nagahama Institute of BioScience and Technology, Japan
Masanori Arita	– University of Tokyo, Japan

ORGANIZING COMMITTEE

Yasubumi Sakakibara	– Keio University, Japan, Chair
Osamu Gotoh	– Kyoto University, Japan
Emi Ikeda	– University of Tokyo, Japan
Hideo Matsuda	– Osaka University, Japan
Satoru Miyano	– University of Tokyo, Japan
Ayumu Saito	– University of Tokyo, Japan
Asako Suzuki	– University of Tokyo, Japan

STEERING COMMITTEE

Minoru Kanehisa	– Kyoto University, Japan
Satoru Miyano	– University of Tokyo, Japan
Mark Ragan	– University of Queensland, Australia
Toshihisa Takagi	– University of Tokyo, Japan
Limsoon Wong	– National University of Singapore, Singapore

CONFERENCE CHAIR

Yasubumi Sakakibara – Keio University, Japan

CONTENTS

PART A

Full Papers

PREDICTING PROTEIN-PROTEIN RELATIONSHIPS FROM LITERATURE USING LATENT TOPICS

TATSUYA ASO[1] KOJI EGUCHI[1]
dango-r@cs25.scitec.kobe-u.ac.jp eguchi@port.kobe-u.ac.jp

[1] *Department of Computer Science and Systems Engineering, Kobe University, 1-1 Rokkoudai, Nada-ku, Kobe, 657-8501, Japan*

This paper investigates applying statistical topic models to extract and predict relationships between biological entities, especially protein mentions. A statistical topic model, Latent Dirichlet Allocation (LDA) is promising; however, it has not been investigated for such a task. In this paper, we apply the state-of-the-art Collapsed Variational Bayesian Inference and Gibbs Sampling inference to estimating the LDA model. We also apply probabilistic Latent Semantic Analysis (pLSA) as a baseline for comparison, and compare them from the viewpoints of log-likelihood, classification accuracy and retrieval effectiveness. We demonstrate through experiments that the Collapsed Variational LDA gives better results than the others, especially in terms of classification accuracy and retrieval effectiveness in the task of the protein-protein relationship prediction.

Keywords: Biomedical text mining; probabilistic topic models.

1. Introduction

There have been increasing demands for organizing knowledge accumulated in documents and then generating potential hypotheses in biomedical fields. This paper focuses on the task to predict relationships between biological entities. Research trends on the biomedical relationship extraction can be categorized into: (1) methods using manually or automatically generated templates, (2) methods based on natural language processing, and (3) statistical co-occurrence-based methods [1, 2]. This paper focuses on the third approaches targeting a specific type of biomedical entities, proteins. While the natural language processing-based approaches usually extract entity relationships within a document, statistical methods are based on co-occurrence of biomedical entities or their related statements in a set of documents to extract relationships between the entities. Statistical topic models are promising for this objective.

Statistical topic models (e.g., [3, 4]) are based on the idea that documents are mixtures of topics, where a topic is a probability distribution over words, in order to capture semantics or to achieve dimensionality reduction. "Probabilistic Latent Semantic Analysis" (pLSA) [5], proposed by Hoffman, can model underlying topics for given documents; however, it cannot model the topics for *unseen* documents that were not used for parameter estimation. Blei et al. [3] proposed one of the

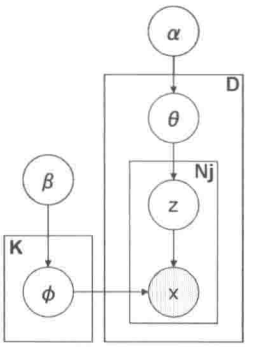


Fig. 1. The graphical model of LDA.

topic models called "Latent Dirichlet Allocation" (LDA) in an extension of pLSA, introducing a Dirichlet prior on a multinomial distribution over topics for each document. This makes the model applicable to unseen documents. The LDA model has been accepted in various fields; however, it has not been investigated for predicting biological entity relationships, to our knowledge. In this paper, we investigate applying the LDA model to extract and predict protein-protein relationships from biomedical literature. In the statistical topic modeling, a set of topics are usually assumed to be unobserved in a document collection, and so we need to infer such unknown distributions from the documents. To estimate the LDA model, "Collapsed Gibbs Sampling inference"[a] method can be used [4]. "Collapsed Variational Bayesian inference" (CVB) [6] is alternative approach to estimate the LDA model.

The focus of this paper is to investigate how to apply the LDA model to the task of protein-protein relationship prediction from biomedical literature, and to evaluate, in an extrinsic manner, the effectiveness over different model estimation methods.

2. LDA and Estimation Algorithms

2.1. *Generative Process of LDA*

Figure 1 shows the graphical model of LDA. We formally describe generative process of LDA [3], as follows,

(1) For all j documents sample $\theta_j \sim Dir(\alpha)$
(2) For all k topic sample $\phi_k \sim Dir(\beta)$
(3) For each of the N_j words x_i in document d_j
  (a) Sample a topic $z_i \sim Mult(\theta_j)$
  (b) Sample a word $x_i \sim Mult(\phi_{z_i})$

[a]It is sometimes simply called "Gibbs Sampling inference" [4].