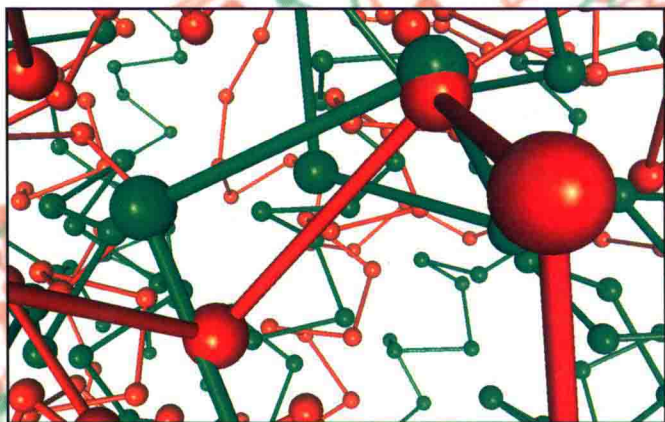


Series on Advances in Bioinformatics and Computational Biology – Volume 5

PROCEEDINGS OF THE 5TH ASIA-PACIFIC  
**BIOINFORMATICS**  
CONFERENCE

EDITORS

DAVID SANKOFF  
LUSHENG WANG  
FRANCIS CHIN



Imperial College Press

PROCEEDINGS OF THE 5TH ASIA-PACIFIC  
**BIOINFORMATICS**  
CONFERENCE

HONG KONG      15 – 17 JANUARY 2007

EDITORS

**David Sankoff**

UNIVERSITY OF OTTAWA, CANADA

**LUSHENG WANG**

CITY UNIVERSITY OF HONG KONG, HONG KONG

**FRANCIS CHIN**

THE UNIVERSITY OF HONG KONG, HONG KONG

*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.

**PROCEEDINGS OF THE 5TH ASIA-PACIFIC BIOINFORMATICS CONFERENCE**

Copyright © 2007 by Imperial College Press

*All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.*

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

ISBN-13 978-1-86094-783-4

ISBN-10 1-86094-783-2

Printed in Singapore by B & JO Enterprise

PROCEEDINGS OF THE 5TH ASIA-PACIFIC

# BIOINFORMATICS

CONFERENCE

**SERIES ON ADVANCES IN BIOINFORMATICS  
AND COMPUTATIONAL BIOLOGY**

**Series Editors:**

**ISSN: 1751-6404**

Ying XU (*University of Georgia, USA*)

Limsoon WONG (*National University of Singapore, Singapore*)

**Associate Editors:**

Ruth Nussinov (*NCI, USA*)

Rolf Apweiler (*EBI, UK*)

Ed Wingender (*BioBase, Germany*)

See-Kiong Ng (*Inst for Infocomm Res, Singapore*)

Kenta Nakai (*Univ of Tokyo, Japan*)

Mark Ragan (*Univ of Queensland, Australia*)

---

Vol. 1: Proceedings of the 3rd Asia-Pacific Bioinformatics Conference  
*Eds: Yi-Ping Phoebe Chen and Limsoon Wong*

Vol. 2: Information Processing and Living Systems  
*Eds: Vladimir B. Bajic and Tan Tin Wee*

Vol. 3: Proceedings of the 4th Asia-Pacific Bioinformatics Conference  
*Eds: Tao Jiang, Ueng-Cheng Yang, Yi-Ping Phoebe Chen  
and Limsoon Wong*

Vol. 4: Computational Systems Bioinformatics  
*Eds: Peter Markstein and Ying Xu*

Vol. 5: Proceedings of the 5th Asia-Pacific Bioinformatics Conference  
*Eds: David Sankoff, Lusheng Wang and Francis Chin*

## PREFACE

High-throughput sequencing and functional genomics technologies have given us the human genome sequence as well as those of other experimentally, medically and agriculturally important species, and have enabled large-scale genotyping and gene expression profiling of human populations. Databases containing large numbers of sequences, polymorphisms, structures and gene expression profiles of normal and diseased tissues are rapidly being generated for human and model organisms. Bioinformatics is thus rapidly growing in importance in the annotation of genomic sequences, in the understanding of the interplay among and between genes and proteins, in the analysis the genetic variability of species, in the identification of pharmacological targets and in the inference of evolutionary origins, mechanisms and relationships.

The Asia-Pacific Bioinformatics Conference series is an annual forum for exploring research, development, and novel applications of bioinformatics. It brings together researchers, professionals, and industrial practitioners for interaction and exchange of knowledge and ideas. The Fifth Asia-Pacific Bioinformatics Conference, APBC2007, was held in Hong Kong 15–17 January, 2007.

A total of 104 papers were submitted to APBC 2007. These submissions came from Bangladesh, China, Hong Kong, India, Japan, Korea, Malaysia, Singapore, Taiwan, Thailand, Australia, New Zealand, Denmark, France, Germany, Hungary, Italy, Israel, Portugal, UK, Canada, Mexico and USA. We assigned each paper to at least three members of the programme committee. Although not all members of the programme committee managed to review all the papers assigned to them, a total of 317 reviews were received, so that there were about three reviews per paper on average.

A total of 35 papers (33%) were accepted for presentation and publication in the proceedings of APBC 2007. Based on the affiliation of the authors, 1.25 of the accepted papers were from China, 1.46 were from Hong Kong, 3 were from Japan, 0.83 were from Korea, 1 were from Singapore, 2.30 were from Australia, 2 were from Denmark, 0.15 were from France, 4.5 were from Germany, 0.5 were from Italy, 1 were from Portugal, 1.66 were from UK, 4.57 were from Canada, 10.78 were from USA.

In addition to the accepted papers, the scientific programme of APBC 2006 also included three keynote talks, by Jennifer A. Marshall Graves, Joseph H. Nadeau and Pavel A. Pevzner, as well as tutorial and poster sessions. The presentations were of very high quality. Almost a third focused on evolution and phylogeny, largely at the genome level, a similar number dealt with protein structure and proteomics more generally, and a good proportion studied various aspects of pathways,

networks, transcriptomics and microarray technology. A range of other topics in bioinformatics and computational biology were also covered, ranging from motif and gene recognition, through haplotypes and population genetics, to databases and text mining. Much of this work featured techniques of sequence analysis, while many of the papers included applications to biology and medicine.

We had a great time in Hong Kong, enhancing the interactions between many researchers and practitioners, and reuniting the Asia-Pacific bioinformatics community in the context of an international conference with worldwide participation.

Finally, we wish to express our gratitude to the authors of the submitted papers, the members of the programme committee and their subreferees, the members of the organizing committee, Phoebe Chen and Limsoon Wong (our liaisons in the APBC steering committee), the keynote speakers, our generous sponsors, and supporting organizations for making APBC 2007 a great success.

David Sankoff  
Lusheng Wang  
Francis Chin

17 January 2007

# APBC2007 ORGANIZATION

## Conference Chair

Francis YL Chin, The University of Hong Kong, Hong Kong

## Organizing Committee

Francis YL Chin (Chair), The University of Hong Kong  
David Smith (Poster Session), The University of Hong Kong  
H.F. Ting (Finance and Registration), The University of Hong Kong  
Lusheng Wang (Publication), The City University of Hong Kong  
Xiaowen Liu (Publication), The City University of Hong Kong  
S.M. Yiu (Local Arrangement and Tutorial), The University of Hong Kong  
Daniel Hung (Webmaster), The University of Hong Kong  
Samson Sin (Webmaster), The University of Hong Kong

## Steering Committee

Phoebe Chen (Chair), Deakin University, Australia  
Sang Yup Lee, KAIST, Korea  
Satoru Miyano, University of Tokyo, Japan  
Mark Ragan, University of Queensland, Australia  
Limsoon Wong, National University of Singapore



## Program Committee

David Sankoff (Chair), The University of Ottawa  
 Lusheng Wang (Chair), The City University of Hong Kong  
 Tatsuya Akutsu, Kyoto University  
 Miguel Andrade, Ottawa Health Research Institute  
 Stephane Aris-Brosou, University of Ottawa  
 Joel Bader, Johns Hopkins University  
 Serafim Batzoglou, Stanford University  
 David Bryant, University of Auckland  
 Jeremy Buhler, Washington University in St. Louis  
 Peter Donnelly, University of Oxford  
 Dannie Durand, Carnegie Mellon University  
 Nadia El-Mabrouk, University of Montreal  
 Robert Giegerich, Bielefeld University  
 Carole Goble, University of Manchester  
 Concettina Guerra, Università di Padova  
 Dan Gusfield, University of California, Davis  
 Michael Hallett, McGill University  
 Sridhar Hannenhalli, University of Pennsylvania  
 Daniel Huson, Tbingen University  
 Gavin Huttley, Australian National University  
 Jenn-Kang Hwang, National Chiao Tung University  
 Tao Jiang, University of California - Riverside  
 Uri Keich, Cornell University  
 Anand Kumar, University of Leipzig  
 Tak Wah Lam, The University of Hong Kong  
 Doheon Lee, KAIST, Korea  
 Jinyan Li, Institute for Infocomm Research  
 Wentian Li, Feinstein Institute for Medical Research  
 Guohui Lin, University of Alberta  
 Michal Linial, The Hebrew University of Jerusalem  
 Zhijie Liu, University of Georgia  
 Bin Ma, University of Western Ontario  
 Satoru Miyano, The University of Tokyo  
 Laxmi Parida, IBM Thomas J. Watson Research Center  
 Mark Ragan, University of Queensland  
 Marie-France Sagot, University Claude Bernard, Lyon I  
 Akinori Sarai, Kyushu Institute of Technology  
 Vincent Schachter, Genoscope  
 Steven Skiena, State University of New York at Stony Brook  
 Edward Susko, Dalhousie University  
 Yun Song, University of California, Davis  
 Robert Stevens, University of Manchester  
 Alfonso Valencia, Centro Nacional de Biotecnología  
 Michael Waterman, University of Southern California

Ken Wolfe, University of Dublin

Stacia Wyman, Williams College

Hong Yan, City University of Hong Kong

Qiang Yang, Hong Kong University of Science and Technology

Kaizhong Zhang, University of Western Ontario

Liqing Zhang, Virginia Tech

Louxin Zhang, National University of Singapore

## Additional Reviewers

Shandar Ahmad	Marcos Jesus Arauzo Bravo	Alexander Auch
Iris Bahir	Richard Bean	Pierre-Yves Bourguignon
Shihyen Chen	Matteo Comin	Mike Cornell
Tobias Dezulian	Zhihong Ding	Maxime Durot
James Eales	Logan Everett	Paul Fisher
Morihiro Hayashida	Cornelia Hedeler	Duncan Hull
Helen Hulme	Shane Jensen	Noam Kaplan
Gunnar W. Klau	Kiyoung Lee	Yaoyong Li
Jingping Liu	Yaniv Loewenstein	Suryani Lukman
Julia Mixtacki	Jose Carlos Nacher	Niranjan Nagarajan
Hiroshi Nakashima	Kay Nieselt	Elon Portugaly
Magnus Rattray	Jonathan Schug	Charles Semple
Baozhen Shan	Yun S. Song	Kristian Stevens
Ashish V. Tendulkar	Victor Tong	Roy Varshavsky
Balaji Venkatachalam	Li-San Wang	Michael Wilson
Katy Wolstencroft	Yufeng Wu	Lei Xin
Zheng Yuan	Guanglan Zhang	

# CONTENTS

Preface	v
APBC 2007 Organization	vii
<b>Keynote Papers</b>	
Exploring Genomes of Distantly Related Mammals <i>J.A. Marshall Graves</i>	1
Bugs, Guts and Fat - A Systems Approach to the Metabolic 'Axis of Evil' <i>J. Nadeau</i>	3
Protein Identification via Spectral Networks Analysis <i>P. Pevzner</i>	5
<b>Contributed Papers</b>	
Metagenome Analysis using MEGAN <i>D.H. Huson, A.F. Auch, J. Qi, and S.C. Schuster</i>	7
Algorithmic Approaches to Selecting Control Clones in DNA Array Hybridization Experiments <i>Q. Fu, E. Bent, J. Borneman, M. Chrobak, and N. Young</i>	17
Subtle Motif Discovery for Detection of DNA Regulatory Sites <i>M. Comin, and L. Parida</i>	27
An Effective Promoter Detection Method using the Adaboost Algorithm <i>X. Xie, S. Wu, K.-M. Lam, and H. Yan</i>	37
A New Strategy of Geometrical Biclustering for Microarray Data Analysis <i>H. Zhao, A.W.C. Liew, and H. Yan</i>	47

Using Formal Concept Analysis for Microarray Data Comparison <i>V. Choi, Y. Huang, V. Lam, D. Potter, R. Laubenbacher, and K. Duca</i>	57
An Efficient Biclustering Algorithm for Finding Genes with Similar Patterns in Time-series Expression Data <i>S.C. Madeira, and A.L. Oliveira</i>	67
Selecting Genes with Dissimilar Discrimination Strength for Sample Class Prediction <i>Z. Cai, R. Goebel, M.R. Salavatipour, Y. Shi, L. Xu, and G. Lin</i>	81
Computing the All-Pairs Quartet Distance on a Set of Evolutionary Trees <i>M. Stissing, T. Mailund, C.N.S. Pedersen, G.S. Brodal, and R. Fagerberg</i>	91
Computing the Quartet Distance Between Evolutionary Trees of Bounded Degree <i>M. Stissing, C.N.S. Pedersen, T. Mailund, G.S. Brodal, and R. Fagerberg</i>	101
A Global Maximum Likelihood Super-Quartet Phylogeny Method <i>P. Wang, B.B. Zhou, M. Taraeneh, D. Chu, C. Wang, A.Y. Zomaya, and R.P. Brent</i>	111
A Randomized Algorithm for Comparing Sets of Phylogenetic Trees <i>S.-J. Sul, and T.L. Williams</i>	121
Protein Structure-Structure Alignment with Discrete Fréchet Distance <i>M. Jiang, Y. Xu, and B. Zhu</i>	131
Deriving Protein Structure Topology from the Helix Skeleton in Low Resolution Density Map using Rosetta <i>Y. Lu, J. He, and C.E.M. Strauss</i>	143
Fitting Protein Chains to Cubic Lattice is NP-Complete <i>J. Mañuch, and D.R. Gaur</i>	153
Inferring a Chemical Structure from a Feature Vector Based on Frequency of Labeled Paths and Small Fragments <i>T. Akutsu, and D. Fukagawa</i>	165

Exact and Heuristic Approaches for Identifying Disease-Associated SNP Motifs	175
<i>G. Huang, P. Jeavons, and D. Kwiatkowski</i>	
Genotype-Based Case-Control Analysis, Violation of Hardy-Weinberg Equilibrium, and Phase Diagrams	185
<i>Y.J. Suh, and W. Li</i>	
A Probabilistic Method to Identify Compensatory Substitutions for Pathogenic Mutations	195
<i>B.C. Easton, A.V. Isaev, G.A. Huttley, and P. Maxwell</i>	
Exploring Genome Rearrangements using Virtual Hybridization	205
<i>M. Belcaid, A. Bergeron, A. Chateau, C. Chauve, Y. Gingras, G. Poisson, and M. Vendette</i>	
Two Plus Two Does not Equal Three: Statistical Tests for Multiple Genome Comparison	215
<i>N. Raghupathy, R. Hoberman, and D. Durand</i>	
The Distance Between Randomly Constructed Genomes	227
<i>W. Xu</i>	
Computing the Breakpoint Distance between Partially Ordered Genomes	237
<i>Z. Fu, and T. Jiang</i>	
Inferring Gene Regulatory Networks by Machine Learning Methods	247
<i>J. Supper, H. Fröhlich, C. Spieth, A. Dräger, and A. Zell</i>	
A Novel Clustering Method for Analysis of Biological Networks using Maximal Components of Graphs	257
<i>M. Hayashida, T. Akutsu, and H. Nagamochi</i>	
Gene Regulatory Network Inference via Regression Based Topological Refinement	267
<i>J. Supper, H. Fröhlich, and A. Zell</i>	
Algorithm Engineering for Color-Coding to Facilitate Signaling Pathway Detection	277
<i>F. Hüffner, S. Wernicke, and T. Zichner</i>	

De Novo Peptide Sequencing for Mass Spectra Based on Multi-Charge Strong Tags	287
<i>K. Ning, K.F. Chong, and H.W. Leong</i>	
Complexities and Algorithms for Glycan Structure Sequencing using Tandem Mass Spectrometry	297
<i>B. Shan, B. Ma, K. Zhang, and G. Lajoie</i>	
Semi-supervised Pattern Learning for Extracting Relations from Bioscience Texts	307
<i>S. Ding, M. Huang, and X. Zhu</i>	
Flow Model of the Protein-protein Interaction Network for Finding Credible Interactions	317
<i>K. Okada, K. Asai, and M. Arita</i>	
All Hits All The Time: Parameter Free Calculation of Seed Sensitivity	327
<i>D.Y.F. Mak, and G. Benson</i>	
Fast Structural Similarity Search Based on Topology String Matching	341
<i>S.-H. Park, D. Gilbert, and K.H. Ryu</i>	
Simple and Fast Alignment of Metabolic Pathways by Exploiting Local Diversity	353
<i>S. Wernicke, and F. Rasche</i>	
Combining N-grams and Alignment in G-protein Coupling Specificity Prediction	363
<i>B.Y.M. Chen, and J.G. Carbonell</i>	
Author Index	373

## EXPLORING GENOMES OF DISTANTLY RELATED MAMMALS

JENNIFER A. MARSHALL GRAVES

*ARC Centre for Kangaroo Genomics, Research School of Biological Sciences  
Australian National University, Canberra, ACT 2601, Australia*

There are three groups of extant mammals, two of which abound in Australia. Marsupials (kangaroos and their relatives) and monotremes (echidna and the fabulous platypus) have been evolving independently for most of mammalian history. The genomes of marsupial and monotreme mammals are particularly valuable because these alternative mammals fill a phylogenetic gap in vertebrate species lined up for exhaustive genomic study. Human and mice ( $\sim 70$ MY) are too close to distinguish signal, whereas mammal/bird comparisons ( $\sim 310$ MY) are too distant to allow alignment. Kangaroos (180 MY) and platypus (210 MY) are just right. Sequence has diverged sufficiently for stringent detection of homologies that can reveal coding regions and regulatory signals. Importantly, marsupials and monotremes share with humans many mammal-specific developmental pathways and regulatory systems such as sex determination, lactation and X chromosome inactivation.

The ARC Centre for Kangaroo Genomics is characterizing the genome of the model Australian kangaroo *Macropus eugenii* (the tammar wallaby), which is being sequenced by AGRF in Australia, and Baylor (funded by NIH) in the US. We are developing detailed physical and linkage maps of the genome to complement sequencing, and will prepare and array cDNAs for functional studies, especially of reproduction and development. Complete sequencing of the distantly related Brazilian short-tailed opossum *Monodelphis domestica* by the NIH allows us to compare distantly related marsupials. Sequencing of the genome of the platypus, *Ornithorhynchus anatinus* by Washington University (funded by the NIH) is complete, and our lab is anchoring contigs to the physical map. We have isolated and completely characterized many BACs and cDNAs containing kangaroo and platypus genes of interest, and demonstrate the value of comparisons to reveal conserved genome organization and function, and new insights in the evolution of the mammalian genome, particularly sex chromosomes.



