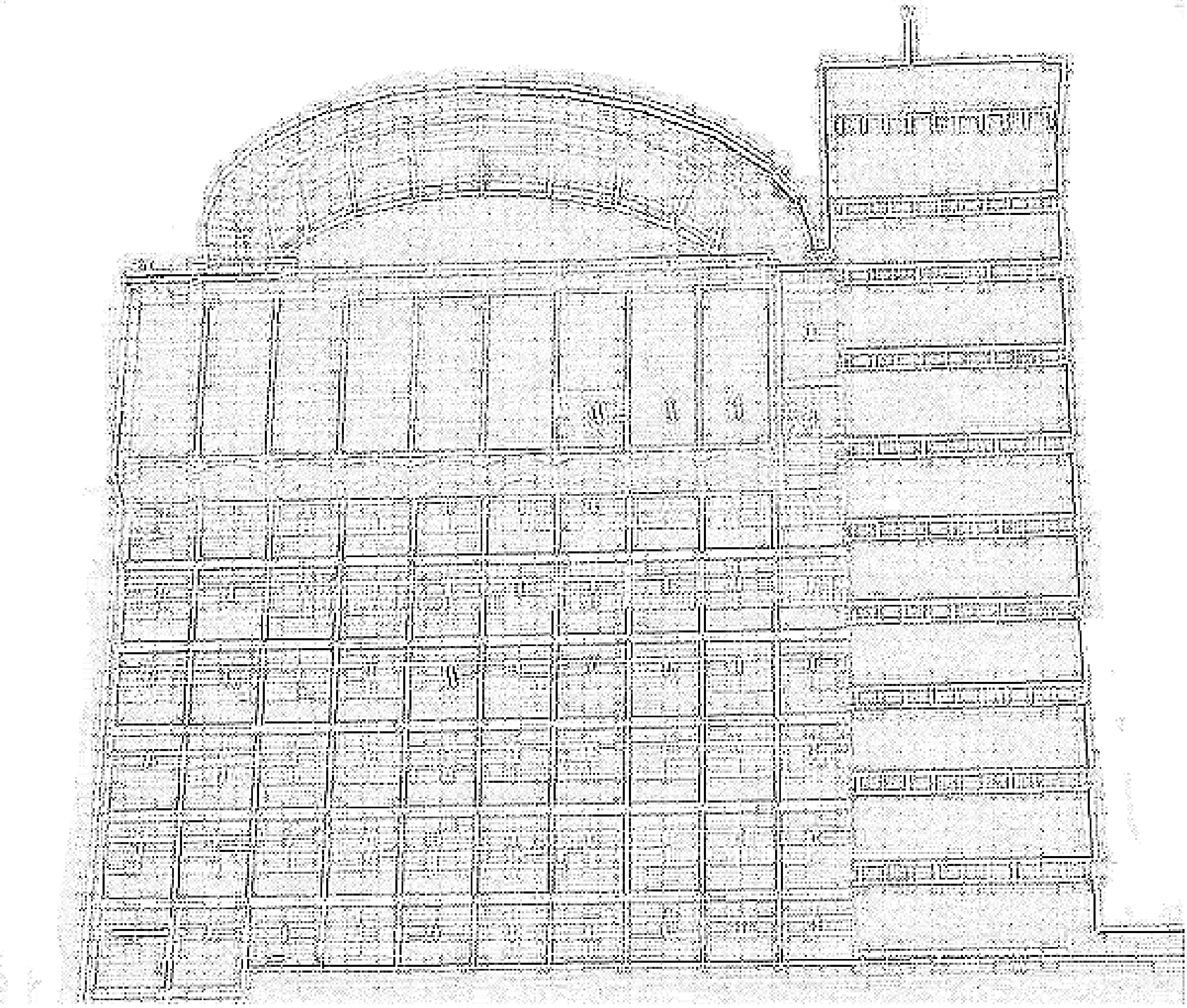
Proceedings of the First Southern African Bioinformatics Workshop

University of the Witwatersrand, Johannesburg

28–30 January 2007



Sponsored by the University of the Witwatersrand, Johannesburg
Department of Science & Technology/National Bioinformatics Network
and the
Centre National de la Recherche Scientifique, France

S. Hazelhurst and M. Ramsay, Eds. *Proceedings of the First Southern African Bioinformatics Workshop (SABioinf 2007)*, Johannesburg, South Africa, 28-30 January 2007. ISBN 978-0-620-38113-0.

ISBN 978-0-620-38113-0

9 780620 381130

Welcome

It is with great pleasure that we warmly welcome you to the First Southern African Bioinformatics Workshop in the summer of 2007 at the University of the Witwatersrand in Johannesburg. Wits Bioinformatics is pleased to host this inaugural workshop.

As a discipline, bioinformatics is still relatively young in South Africa and we were encouraged by the excellent response to the call for papers. The fact we have been able to attract almost 100 delegates and five international invited speakers is a mark of how bioinformatics has progressed in South Africa. Two of our invited speakers, Alan Christoffels and Janet Kelso, are PhD graduates from the South African National Bioinformatics Institute at the University of the Western Cape and we are delighted with their involvement in this workshop. Eric Rivals joins us from the University of Montpellier and CNRS and Kateryna Makova and Anton Nekrutenko are both from Penn State University.

The scientific programme has been arranged according to themes based on the submitted papers and include: New Computational Techniques; Toolkits, Integration and Services; Gene Identification and Expression; Evolution and Phylogenetics; and Transcriptomics. We have also planned panel discussions to rigorously debate the issues of mathematics teaching and education in preparation for a career in bioinformatics and the role of bioinformatics in high performance computing.

The workshop has been generously funded by the University of the Witwatersrand. The workshop is hosted by Wits Bioinformatics, situated on the 12th and 13th floors of the University Corner Building with magnificent views stretching south over the Nelson Mandela Bridge and north towards the tree-rich suburbs. We hope delegates will have an opportunity to visit us.

We have also had generous support to fund delegates from the national Department of Science and Technology, and the Centre National de la Recherche Scientifique (within the framework of the French support of NEPAD's scientific programme). It is supported by the National Bioinformatics Network (NBN) under whose guidance this and future workshops will be held.

32 papers were submitted for reviewing and 26 were finally accepted. Of these, 9 papers were submitted in the *full research paper* category. The full research papers were reviewed by 3-5 reviewers, and finally 3 of these papers were accepted in the full research paper category. The other papers were all sent to external reviewers and considered by a sub-committee of the programme committee. With the exception of one paper, all papers were reviewed by between 2 and 4 referees. In addition over 40 posters have been submitted and are included in the programme.

This is the first of a series of annual bioinformatics workshops that will take place across southern Africa in the years to come. May you embrace this as an opportunity to get to know like-minded scientists, to make new friends and to enjoy the ambiance of our lovely University. You will be richly rewarded if you take a little time to visit the Origins Centre and reflect on the growth and evolution of our species. We wish you a challenging, informative and enjoyable workshop and look forward to your active participation to make this a truly memorable event.

Scott Hazelhurst and Michéle Ramsay
Programme Committee Co-Chairs
January 2007

Thanks

The Organising Committee thanks the following people for their help. Professor Loyiso Nongxa, the Wits Vice-Chancellor has always been supportive of the development of bioinformatics and in particular gave this workshop unstinting support. Professor Belinda Bozzoli, Deputy Vice-Chancellor (Research) and present and past Deans of the Faculties of Science and Health Sciences, Ramesh Bharuthram, Helen Laburn, Max Price and Colin Wright, have also been supportive. We thank the reviewers, many of whom where generous with their time and jumped in at the last moment. Olivier Gascuel and Anne Corval from CNRS were very helpful in making contacts and in promoting the workshop. The administrative support of Varsha Sokay and Brenda Lacey-Smith was invaluable, and Sarah du Plessis and Betty Selebogo of Munlin Travel responded promptly to our needs and helped us recover from our mistakes.

Programme and Organising Committees

—Programme Committee

—Co-chairs: Scott Hazelhurst & Michèle Ramsay, Wits.

—Alia BenKahla, Institut Pasteur, Tunis; Judith Bishop, Pretoria; Greg Blatch, Rhodes; Junaid Gamieldien, NBN Central; Winston Hide, SANBI; Jannie Hofmeyr, Stellenbosch; Dan Jacobson, NBN Central; Fourie Joubert, Pretoria; Heikki Lehväslaiho, SANBI; Daniel Masiga, ICIPE, Nairobi; Nicky Mulder, Cape Town; Hugh Murrell, KwaZulu-Natal; Hugh Patterton, Free State.

—Organising committee

—Scott Hazelhurst (Chair), Jonathan Burke, Pierre Durand, Brenda Lacey-Smith, Khayeni Ndlovu, Andries Oelofse, Michèle Ramsay, Christine Rey

Reviewers

We thank the following people who were responsible for reviewing the papers.

—Alain Denise, Paris-Sud

—Alia BenKahla, Institut Pasteur, Tunis

—Allen Rodrigo, Auckland

—Anna Kramvis, Wits

—Cathal Seoighe, Cape Town

—Christine Rey, Wits

—Darren Martin, Cape Town

—Dean Goldring, KwaZulu-Natal

—Fourie Joubert, Pretoria

—François Coste, IRISA, Rennes

—Frank Dehne, Carleton

—Hugh Murrell, KwaZulu-Natal

—Ian Sanders, Wits

—Judith Bishop, Pretoria

—Karen Megy, EBI

—Michèle Ramsay, Wits

—Nicky Mulder, Cape Town

—Nir Oren, Aberdeen

—Philip Machanick, Queensland

—Pierre Durand, Wits

—Scott Hazelhurst, Wits

—Steven Orzack, Freshpond Institute

—Tracy McLellan, Wits

—Winston Hide, SANBI/UWC

Contents

Welcome iii

Programme and Organising Committees iv

I Invited Lectures 1

Tandem repeats analysis
Eric Rivals 2

Cyprinid Genomics: A GMOD-based framework for cross-species integration of transcripts into closely-related genomes
Alan Christoffels 3

Male mutation bias and X chromosome inactivation in the age of genomics
Kateryna Makova 4

The Neanderthal Genome Project: Bioinformatics challenges
Janet Kelso 5

Dual Coding in Eukaryotic Genomes Defies Statistical Odds
Anton Nekrutenko 6

II Full research papers 7

A two-phase approach for detecting recombination in nucleotide sequences
C.X. Chan, R. Beiko, and M.A. Ragan 9

Identification of orphan environmental sequences by relative abundance of oligonucleotides
O. Reva 17

On the Detection of Recombination in Minisatellite Data
E. Adebiyi and E. Rivals 25

III Short papers and Works-in Progress 33

Using a Tuple Space Web Service for Parallel Processing in Bioinformatics
B. Mueller, L. Schulé, G. Wells 34

Low-resolution structural genomics by structure prediction and filtering of decoys using sparse experimental data
J. M. Bujnicki 38

Progress towards Determining the Structure of Plasmodium falciparum *Pyridoxal Kinase*
Ö. Taştan Bishop, G. Wells, F. Joubert, C. Wrenger, R. D. Walter and A.I. Louw 41

Learning a Subclass of Tree Adjoining Grammars
S. Al Seesi, S. Rajasekaran, R. Ammar 45

Knowledge Driven User Interfaces for Complex Biological Queries
K. O'Neill, A. Garcia-Castro and D. Jacobson 49

MScanner: A tool to expand a corpus of Medline citations
G.L. Poulter, C. Seoighe, D.L. Rubin, R.B. Altman 53

*Mitochondrial DNA shows three cryptic populations of Campbell's and Lowe's monkeys (*Cercopithecus campbelli*, Waterhouse, 1838) in Côte d'Ivoire*
S. Gonedelé Bi, A. Sangare, E. N'Goran, D. Zinner, H. Siedel, L. Vigilant, C. Boesch 59

In-silico Modelling of Helicoverpa armigera *Aminopeptidase-N receptor*
C.V.S. Siva Prasad and A. Singh 63

ISR-mediated gene expression provides insight into the molecular regulation of systemic plant protection by non-pathogenic bacteria
S. Muyanga, S. Meier, C. Gehring and V. Bajic 67

A Computational Approach to Candidate Gene Prioritisation Using Data-Mining and Clinically-Informed Binary Filtering
Z. Lombard, M. Ramsay, and W. Hide 71

IV Reviewed Presentations 75

Development of a web toolkit for the biological interpretation of microarray gene clusters
P.J. Law, C. Claudel-Renard, F. Joubert, D. K. Berger and A.I. Louw 76

A model of directional selection applied to the evolution of drug resistance in HIV-1
C. Seoighe, K. Scheffler, V. Pillay, L. Morris and W. Hide 76

CTL response to HIV-1 subtype C is poorly predicted by epitope motifs
N. Ngandu, C. Seoighe, H. Bredell, C. Williamson, C.M. Gray and the HIVNET028 Study Team 77

An Efficient Algorithm for Oligonucleotides Selection in a Large EST Database
E. Adebiyi 78

What do haplotype inferral programs tell us? An investigation of the human haptoglobin locus
Tracy McLellan, Lindiwe Skhosana and Steven Hecht Orzack 79

Comparison of Developmental Mammalian Gene Expression Using Simplified Ontologies
A. Kruger, O. Hofmann and W. Hide 80

Standards in alternative splicing
O. Hofmann, C. Iseli, C. Mungall, M. Sammeth, E. Eyras, S. Lewis, and W. Hide 80

Molecular Library Building through a Markush-like Molecular Enumeration Technique
W. Cochrane and F. Joubert 81

Integration of HIV clinical and molecular data using Open Source genome managements tools
H. Lehvaslaiho, A. Dawe, A. Kamau, R. Van Rooyen, A. Boardman, Alan Powell And Winston Hide . . . 81

Extracting value from single-read sequence data
G. Koning, S. Hazelhurst, M. Jonas, A. Murray, J. Grzymski, C. Maher, D. Cowan And W. Hide 82

Elucidation of Sulfadoxine Resistance with Structural Models of the Bifunctional Plasmodium falciparum *Dihydropterin Pyrophosphokinase Dihydropteroate Synthase*
T. De Beer, A.I. Louw, and F. Joubert 83

Application of a Robust Method to Reevaluate Evidence for Positive Selection in HIV Coding Sequences
N. Wood, K. Scheffler and C. Seoighe 84

Identification of allele-specific mRNA transcripts through an integrative analysis of genomic and EST data
V. Nembaware, B. Lupindo, K. Scheffler AND Cathal Seoighe 84

V Posters 85

Identification of a new arenavirus in Hylomyscus Simus *in Côte D'Ivoire*
D. N'Golo Coulibaly, C. Akoua Koffi, S.K. Kouassi, B. Becker-Ziaja, B. Allali, B. Liedigk, E. Ekaza, M. Dosso, E. Fichet-Calvet, J. ter Meulen and S. Günther 86

Protein Analysis Made Easy
T. de Beer and F. Joubert 87

Exploring unusual regulatory behaviour in a simple metabolic pathway using PySCeS
B.G. Olivier, J.M. Rohwer and J.-H.S. Hofmeyr 87

Comparative genome analysis of pathogenic and non-pathogenic African Swine Fever Virus genomes
E.P. de Villiers, C. Gallardo, E. Blanco, A. Llorente, F. Kamanu and R. Bishop 88

Prevalence of SNPs at exon-intron junctions of alternatively spliced genes
B. Lupindo AND C. Seoighe 89

Microarray Analysis For Human Malaria Parasite Plasmodium Falciparum Using K-Means Clustering Computational Approaches
V. Osamor and E. Adebiyi 89

Prediction of Nuclear Hormone Response Elements
A. Rebaï 90

Comparative sequence analysis of microbial genomes to identify potential virulence genes in Mycobacterium tuberculosis
G. Jamieson, H. Rabiu, V. Vuppu, and N. Mulder 90

The elucidation of host responses to South African cassava mosaic virus using cDNA microarrays
F. van Schalk and M.E.C. Rey 91

Providing python bindings for the SUNDIALS suite of dynamic system solvers using ctypes
J. Dominy, J. Rohwer and J.-H.S. Hofmeyr 92

Use of oligonucleotide frequencies to derive signatures for prokaryotic genomes
C.F. Davenport, O. Reva and B. Tuemmler 92

An Easy On The Eye Whole Genome Browser
H. Ganesan, A. Rakitianskaia and O. Reva 92

In Silico Identification of Possible Leishmanial Secretory Proteins with Sequence-Specific Nucleic Acid Binding Competency
K.F. Kinyua 93

Detecting Recombination within large groups of sequences using an FPGA
L. Thomson 93

The Crystal Structure of a Mutant Nitrile Hydratase
S. Kwofie, P. Kowlessur, A.O. Bishop, T. Tsekoa, S. Muhammed, T. Sewell and D. Cowan 93

A generic rate equation for catalysed, template-directed polymerisation reactions for use in computational systems biology
O.P. Gqwaka, J.M. Rohwer and J.-H.S. Hofmeyr 94

Homology modelling of human protein cathepsin A
R. Reja and G.K. Maheshwari 95

CAGED: Chromosomal Arrangement of Gene Expression Data
S. Somers, C. Gehring, C. Seoighe 95

Mechanistic insight into the working of spermidine synthase of Plasmodium falciparum *from a homology model*
P.B. Burger, L.-M. Birkholtz, F. Joubert, N. Haider, R.D. Walter and A.I. Louw 96

Plant-scale structural model of carbon fixation and sucrose accumulation in sugarcane
K.L. Meyer, L. Uys and J.M. Rohwer 97

Modelling sucrose accumulation in maturing sugarcane
L. Uys and J. Rohwer 98

A structural bioinformatics approach to Theileria parva *genome annotation*
T. Shah, E. de Villiers, V. Nene, M. Gardner, R. Bishop 99

d^1 EST clustering
D. Kennedy 99

Understanding the control and regulation of the mammalian cell cycle
R. Conradie and J.L. Snoep 100

Oscillatory behaviour and synchronisation in a detailed model of yeast glycolysis
F.B. du Preez and J.L. Snoep 101

A sequence-centric biological workbench as part of a bioinformatics information management system
C.A. Hefer and F. Joubert 102

Review of the implementation of codon usage bias in software utilised in the design of functional synthetic protein coding regions
J.S. Baxter, G.L. Blatch, G. Jäger 103

Identification and characterisation of potential duplicate genes in Mycobacterium tuberculosis *through comparative genomic studies*
V. Vuppu and N. Mulder 104

Annotation of a cDNA Microarray using SSHscreen and MADIBA software
D. Berger and N. Coetzer 105

The JWS Online Project
C. van Gend and J. Snoep 105

Symbolic Control Analysis of Cellular Systems
T.J. Akhurst, J.-H.S. Hofmeyr and J.M. Rohwer 106

Clustering Expressed Sequenced Tags using a Tree-Structured Index Algorithm
B. Kumwenda 107

A Graph-relational Integrated Database of Human and Mouse Omics data
J. Gamieldien, D. Otgaar, N. Southgate, D.A. Jacobson 108

Digr: A web interface for mining Digrabase graph databases
K. O'Neill, H. Murrell, D.A. Jacobson 109

BISI: an integrative, extensible and minable database of South African wildlife and indigenous livestock biomaterial information
A. Maclear, D. Otgaar, J. Gamieldien, P. Bartels, A. Kotze, D. Jacobson 109

BLAST on Virtual Clusters, Affordable High Performance Computing
I. Scheepers, D.A. Jacobson 111

Vector Graphics to improve Blast graphic representations
R. Jimenez, H. Murrell, D. Jacobson 112

DigraBase — A Multigraph-based Data Store for Integrating Biological Data
D. Otgaar, J. Gamieldien, F. Martinez, B.Nortier, D. Jacobson 112
BioFrame — A Java-based Software Development framework for Bioinformatics
F. Martinez, J. Gamieldien, D. Otgaar, D. Jacobson 113
Large-scale design of primers for genes from atypical plants
I. Riani, Y. Trifa, K.W. Berendzen and A. BenKahla 114
Global analysis of Leishmania genes expression using SAGE libraries
S. Smandi, F. Guerfali, D. Laouini, K. Dellagi and A. Benkahla 114

VI Indices 115

Author Index 115

Topic Index 118

An Efficient Algorithm for Oligonucleotides Selection in a Large EST Databases

Ezekiel Adebiyi

Department of Computer and Information Sciences, Covenant University, PMB 1023, Ota, Nigeria.

ABSTRACT

Identifying unique oligonucleotide (oligo) probe sequences is an important step in PCR and microarray experiments. While there are a growing number of complete and annotated genomes, the largest collection of publicly available genetic sequences are expressed sequence tag (EST) sequences. Furthermore, for many organisms that are important to the society, such as barley, the EST is the major data on the expressed genes in a number of these organisms. For the EST sequences, the unique oligo problem is the selection of oligos each of which appears (exactly) in one EST sequence but does not appear (exactly or approximately, for a given hamming difference d) in any other EST sequence.

OligoSpawn, in two phase, has been implemented to efficiently select oligos from ESTs. The notion of a "seed" was used in the construction of OligoSpawn, and its run time is exponential dependent on q (the length of the "seed"). For $q = 11$, it ran on a previous barley dataset of 28MB for 2 hours and 26 minutes using a 1.2GHz AMD machine, but it is very inefficient for large datasets, like the new 43MB barley dataset. We observed this as OligoSpawn, for $q = 11$, runs for about 6 days using a 3.0GHz Pentium IV machine. Furthermore, selection of some important unique oligos (*e.g.*, for which $q = 13$) is unwieldy for OligoSpawn.

In this work, using the suffix tree, we give a careful theoretical characterization of the set of seeds required, and prove a subqradratic time algorithm for extracting these seeds. Using this result, we present an efficient algorithm that takes advantage of the new results, that simplify the solution of the least common ancestor (LCA) problem via the range minimum query (RMQ) problem. The run time of our resulting algorithm is $O(n^3qd/4^{2q})$. For $q = 11$ and $q = 13$, our algorithm runs on the new 43MB barley dataset for 4 days using also a 3.0 GHz Pentium IV. As far as we know, our algorithm is the fastest oligonucleotides selector algorithm for large databases of tens of thousands of EST sequences, such as the barley ESTs.