João Carlos Setubal
Sergio Verjovski-Almeida (Eds.)

LNBI 3594

Advances in Bioinformatics and Computational Biology

**Brazilian Symposium on Bioinformatics, BSB 2005
Sao Leopoldo, Brazil, July 2005
Proceedings**

Springer

João Carlos Setubal
Sergio Verjovski-Almeida (Eds.)

Advances in Bioinformatics and Computational Biology

Brazilian Symposium on Bioinformatics, BSB 2005
Sao Leopoldo, Brazil, July 27-29, 2005
Proceedings

Springer

Volume Editors

João Carlos Setubal
Virginia Bioinformatics Institute and Department of Computer Science
Virginia Polytechnic Institute and State University, Bioinformatics 1, Box 0477
Blacksburg, VA 24060-0477, USA
E-mail: setubal@vbi.vt.edu

Sergio Verjovski-Almeida
Universidade de Sao Paulo
Instituto de Quimica, Departamento de Bioquimica
Av. Prof. Lineu Prestes 748, 05508-000 Sao Paulo, SP, Brazil
E-mail: verjo@iq.usp.br

Library of Congress Control Number: 2005929321

CR Subject Classification (1998): H.2.8, F.2.1, I.2, G.2.2, J.2, E.1

ISSN 0302-9743
ISBN-10 3-540-28008-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-28008-8 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11532323 06/3142 5 4 3 2 1 0

Lecture Notes in Bioinformatics 3594

Edited by S. Istrail, P. Pevzner, and M. Waterman

Subseries of Lecture Notes in Computer Science

Preface

The Brazilian Symposium on Bioinformatics (BSB 2005) was held in São Leopoldo, Brazil, July 27–29, 2005, on the campus of the Universidade Vale do Rio dos Sinos (Unisinos). BSB 2005 was the first BSB symposium, though BSB is in fact a new name for a predecessor event called the Brazilian Workshop on Bioinformatics (WOB). WOB was held in three consecutive years: 2002, 2003, and 2004. The change from workshop to symposium reflects the increased reach and quality of the meeting. BSB 2005 was held in conjunction with the Brazilian Computer Society's (SBC) annual conference.

For BSB 2005 we had 55 submissions: 45 full papers and 10 extended abstracts. These proceedings contain the 15 full papers that were accepted, plus 16 extended abstracts (a combination of the accepted abstracts and some full papers that were accepted as extended abstracts). These papers and abstracts were carefully refereed and selected by an international program committee of 40 members, with the help of some additional reviewers, all of whom are listed on the following pages. These proceedings also include papers from three of our invited speakers. We believe this volume represents a fine contribution to current research in bioinformatics and computational biology.

The editors would like to thank: the authors, for submitting their work to the symposium, and the invited speakers; the program committee members and other reviewers for their help in the review process; the Unisinos local organizers, José Mombach and Ney Lemke; Marcelo Walter from Unisinos, coordinator of the SBC conference; Ivan Sendin, from the University of Goiás, who helped with fund raising; Margaret Gabler, from VBI, who helped with the preparation of the proceedings; the symposium sponsors (see list in this volume); Guilherme Telles, Ana Bazzan, Marcelo Brígido, Sergio Lifschitz, and Georgios Pappas, members of the SBC special committee for computational biology; and Springer for agreeing to print this volume.

July 2005

João Carlos Setubal
Sergio Verjovski-Almeida

Organization

BSB 2005 Scientific Program Committee

João Carlos Setubal *Informatics Chair*	(Virginia Bioinformatics Institute, Virginia Tech, USA)
Sergio Verjovski-Almeida *Biology Chair*	(University of São Paulo, Brazil)
Nalvo Almeida Jr.	(Federal University of Mato Grosso do Sul, Brazil)
Ricardo Baeza-Yates	(ICREA-Univ. Pompeu Fabra, Spain & Univ. of Chile)
Valmir Barbosa	(Federal University of Rio de Janeiro, Brazil)
Ana Bazzan	(Federal University of Rio Grande do Sul, Brazil)
Marcelo Brígido	(University of Brasília, Brazil)
Marcelo Briones	(Federal University of São Paulo)
André Carvalho	(University of São Paulo, Brazil)
Julio Collado-Vides	(Autonomous University of Mexico)
Allan Dickerman	(Virginia Tech, USA)
Alan Durham	(University of São Paulo, Brazil)
Carlos Ferreira	(University of São Paulo, Brazil)
James Glazier	(University of Indiana, USA)
Katia Guimaraes	(Federal University of Pernambuco, Brazil)
Lenny Heath	(Virginia Tech, USA)
Victor Jongeneel	(Ludwig Institute, Lausanne, Switzerland)
João Kitajima	(Alellyx, Brazil)
Natalia Martins	(EMBRAPA, Brazil)
Wellington Martins	(Catholic University of Goiás, Brazil)
Marta Matoso	(Federal University of Rio de Janeiro, Brazil)
João Meidanis	(Scylla and University of Campinas, Brazil)
Pedro Mendes	(Virginia Tech, USA)
José Mombach	(Unisinos, Brazil)
Bernard Moret	(University of New Mexico, USA)
Eduardo Jordão Neves	(University of São Paulo)
Ney Lemke	(Unisinos, Brazil)
Sergio Lifschitz	(Catholic University, Rio de Janeiro, Brazil)
Georgios Pappas	(Catholic University of Brasilia, Brazil)
Christian Probst	(Mol.Biol.Institute, Curitiba, Brazil)
Eduardo Reis	(University of São Paulo, Brazil)
Leila Ribeiro	(Federal University of Rio Grande do Sul, Brazil)
Larry Ruzzo	(University of Washington, USA)
Marie-France Sagot	(INRIA, France)
Bruno Sobral	(Virginia Tech, USA)

BSB 2005 Scientific Program Committee (continued)

Siang Song	(University of São Paulo, Brazil)
Osmar Norberto de Souza	(Catholic University of Rio Grande do Sul, Brazil)
Guilherme Telles	(University of São Paulo, Brazil)
Fernando von Zuben	(University of Campinas, Brazil)
Maria Emilia Walter	(University of Brasília, Brazil)

Additional Reviewers

Edson Cáceres
Marcelo Henriques de Carvalho
Jian Chen
Vicky Choi
Lokesh Das
Luciano Digiampietri
Vladimir Espinosa
Katti Faceli
Paulo Roberto Ferreira Jr.
Julio Freyre
Abel González
Marco Gubitoso
Giampaolo Luiz Libralão
Ana Lorena
Sandro Marana
Cleber Mira
Alexey Onufriev
José Augusto Amgarten Quitzau
Cassia Trojahn dos Santos
Marcilio de Souto
Bruno de Souza
Eric Tannier

Local Organizers

José Mombach (Unisinos, Brazil)
Ney Lemke (Unisinos, Brazil)

Sponsoring Institutions

Brazilian Computer Society (SBC)
Universidade Vale do Rio dos Sinos
The Brazilian National Council for Research (CNPq)
The Rio Grande do Sul State Research Agency (FAPERGS)
Hewlett-Packard
GE Healthcare
Invitrogen
Microsoft

Table of Contents

Invited Papers

Full Papers

Extended Abstracts

Differential Gene Expression in the Auditory System

Irene S. Gabashvili[1], Richard J. Carter[1], Peter Markstein[1], and Anne B.S. Giersch[2]

[1] Hewlett-Packard Labs, Computational Biosciences Research, 1501 Page Mill Road, Palo Alto, CA, 94304, USA
{Irene.Gabashvili, Dick.Carter, Peter.Markstein}@HP.com
http://hpl.hp.com/research/cbsr
[2] Department of Pathology, BWH, Harvard Medical School, 75 Francis Street, 02115 Boston, USA
agiersch@rics.bwh.harvard.edu
http://hearing.bwh.harvard.edu/

Abstract. Hearing disorders affect over 10% of the population and this ratio is dramatically increasing with age. Development of appropriate therapeutic approaches requires understanding of the auditory system, which remains largely incomplete. We have identified hearing-specific genes and pathways by mapping over 15000 cochlear expressed sequence tags (ESTs) to the human genome (NCBI Build 35) and comparing it to other EST clusters (Unigene Build 183). A number of novel potentially cochlear-specific genes discovered in this work are currently being verified by experimental studies. The software tool developed for this task is based on a fast bidirectional multiple pattern search algorithm. Patterns used for scoring and selection of loci include EST subsequences, cloning-process identifiers, and genomic and external contamination determinants. Comparison of our results with other programs and available annotations shows that the software developed provides potentially the fastest, yet reliable mapping of ESTs.

1 Introduction

Personalized medicine in the future will be based on the comparison of individual genetic information to reference gene expression, molecular interactions and pathways in tissues and organs, in health and disease. It will be based on advanced genome sequencing, gene expression, proteomic and metabolomic technologies, as well as efficient computational tools for mapping of genes and pathways.

The reliability of computational approaches and models is improving, as "omic" technologies mature and the accuracy of predictions grows with increasing data input. There is a growing need for fast software tools capable of handling massive amounts of data and reanalyzing the data to discover integrated knowledge and identify broken links and wrong connections between intricate processes in individual datasets.

The first step in comparing genomic information is to align DNA sequences, that is, to map nucleotides of expressed sequence tags (ESTs) or full cDNAs to the genome and sequences of known and predicted genes. Sequence alignment is one of

J.C. Setubal and S. Verjovski-Almeida (Eds.): BSB 2005, LNBI 3594, pp. 1–8, 2005.

the oldest and most successful applications of Computer Science to Biology [1-2]. Many local pairwise alignment methods exist [1-6] and most software tools are freely available. These tools, however, are customized for specific tasks and do not allow enough flexibility for new specialized tasks to external users. The most popular generic programs relevant to EST mapping, BLAST from the National Center for Biotechnology Information [6] and BLAT from U.C. Santa Cruz [4], each have their strengths and weaknesses. The BLAST service offered by NCBI is too slow to use for sets of tens of thousands ESTs. Moreover, it does not handle intron gaps well when used for the whole-genome mappings and works best on expressed sequence databases. The BLAT service offered by UCSC is fast, but its interactive nature and 25-sequence submission limit would prevent its use on a large number of sequences.

To direct and control the process of EST mapping, we needed software with problem-specific intelligence that was not available with existing tools. One of the most important tasks in processing experimental data is estimating the errors and potential sources of errors in measurements [7]. Cloning and sequencing artifacts, for example, could be eliminated using pre-screening procedures. Accordingly, we needed not only to align ESTs, but also check for a number of favorable and detrimental signals, to identify the most likely mapping amongst many possibilities.

In this work, we have analyzed over fifteen thousand ESTs expressed in the human cochlea. The cochlea is one of the smallest organs in the body located in the inner ear and responsible for auditory transduction (conversion of sound into the language of the brain). Hearing impairment is always the result of damage to either the middle ear, the cochlea or its associated auditory nerve. Over one hundred genes responsible for deafness have been discovered, but many more candidates apparently exist. A much smaller fraction of molecular-level auditory pathways have been identified [8-10], mostly due to the lack of knowledge of human biology in general.

We have mapped and analyzed genes predominantly expressed in the inner ear and their pathways. We have also studied cochlear genes expressed in low numbers. We show that the vast majority of cochlea-unique genes identified by existing tools and servers are either genomic contaminations or can be also found in other tissues. We have selected a small subset of cochlea-specific genes and they are currently being verified by independent experimental methods.

2 Computational Approach

To speed up alignment of ESTs to the genome and improve the scoring of such mappings, we reduced the problem to that of simultaneous exact matching of multiple motifs within ESTs to localized genome regions. Our approach is illustrated on the example of a particular Morton cochlear EST (Fig. 1).

Mapping and selection of ESTs is realized by dynamic interaction of two in-house programs, *Enhancer2* and *BatchSearch*. *Enhancer2* is a 5000-line C++ program that finds exact matches of a number of input search patterns within a database of sequences (whole genomes, mRNAs, etc). The fast exact string prefix matching algorithm (Dick Carter and Peter Markstein, to be published) was applied to other genome search problems in early stages of its development [11]. Some of the features

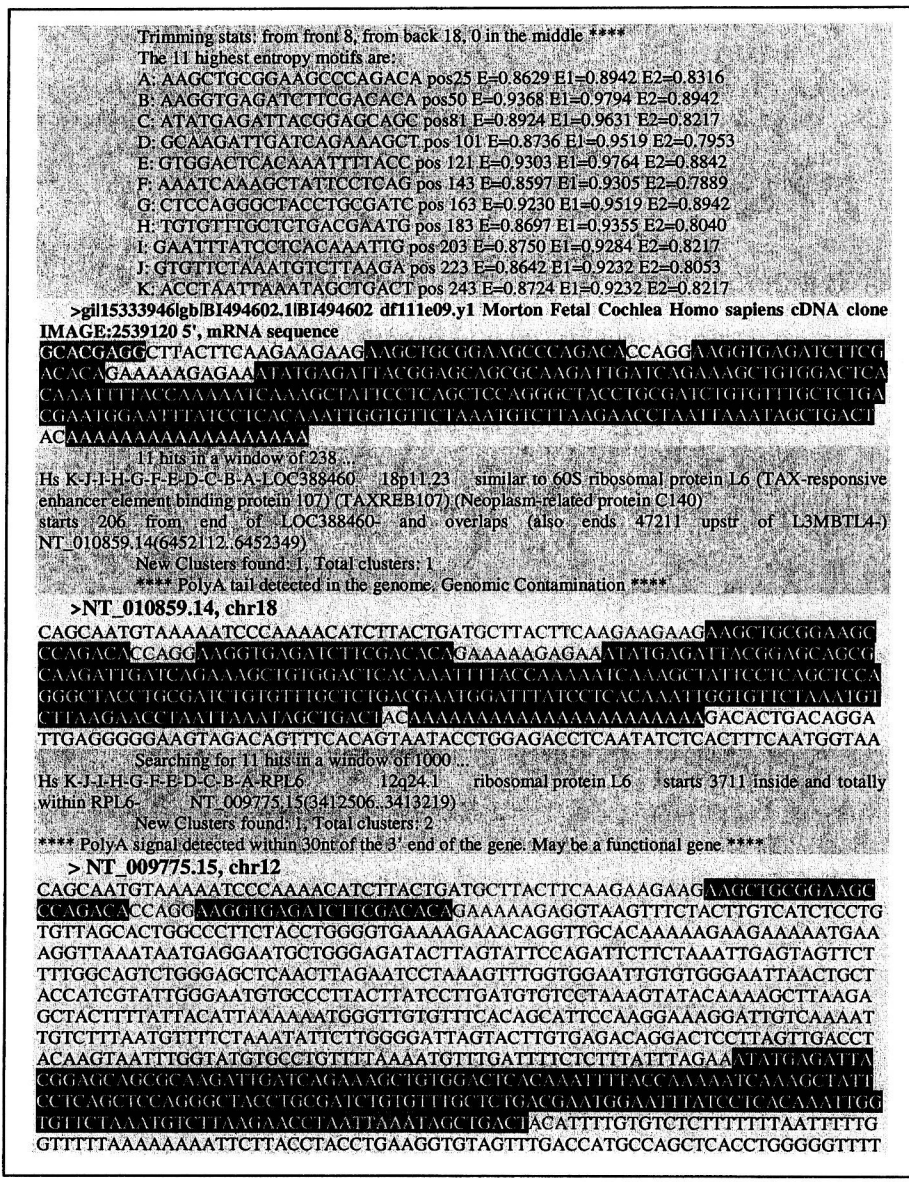

Fig. 1. Our approach to mapping and scoring of results illustrated on the example of a sequence with accession number BI49460. As a first step, we determined detrimental motifs in this sequence (shaded in grey) and trimmed them off. Blue area represents dynamically selected subsequences used for matching to the human genome. The program found two equally well matching regions in chromosomes 12 and 18. A detrimental signal (polyA tail (black shading), in chromosome 18 and a favorable motif in chromosome 12 determined the best mapping. See text for details

of this algorithm are its ability to handle all IUPAC nucleotide codes with little additional overhead and its high parallelization efficiency.

The other component of our EST-mapping solution is *BatchSearch*, a 2500-line C++ program that interacts with *Enhancer2* by giving it search tasks and dynamically responding to its output. Using the fast exact-matching *Enhancer2* speeds the alignment process since EST-mapping would normally require slower inexact matching to cope with introns and frequent EST sequencing errors or single nucleotide polymorphisms (SNPs). Our idea was to divide an EST into smaller fragments and, using *Enhancer2*, find where some of them occur. Normally the bulk of the fragments would be found clustered within the same locale, thus forming the basis for the reported EST mapping. In the majority of cases, we also observed a very high level of identity, as an entire EST sequence after trimming often exactly matched to a localized region within the genome.

The logic of *BatchSearch* involves a number of steps. First, the input EST is trimmed of bases that are artifacts of the sequencing process (Fig.1). Second, a globally optimal set of high-entropy fragments is chosen from the EST using a dynamic programming algorithm. Then, the formulated exact-match search problem is passed to the waiting *Enhancer2* program. Depending on these results, *BatchSearch* can ask *Enhancer2* to refilter its search results, allowing for more widely dispersed clusters to be reported. In addition, clusters of other detrimental and favorable motifs in the genome are taken into account. Fig.1 demonstrates two such motifs – a polyA tail (*black shading*) that is supposed to be located within 30 nucleotides of the 3' end (larger distance may be allowed in the 5' EST) and a polyA signal (see [12], *orange shading*, not be followed by polyA tail in the genome) Alternatively, *BatchSearch* can redo the genome search with smaller EST subsequences, in an effort to identify the most likely mapping. One search for six 20-nucleotide fragments using *Enhancer2* takes about 2.5 seconds on a 2.8 gHz Xeon CPU with one Giga Byte of RAM. A dual-processor HP XW8000 PC workstation requires 5.5 hours to map the entire library of 15000 cochlear ESTs to the human genome. Datasets with less mapping ambiguity are processed faster.

3 Genes and Pathways of the Human Cochlea

Only from 60% to 95% of all deposited ESTs in tissue- and organ-specific libraries are classified by Unigene. Fig.2 demonstrates the ratio of classified vs. unclassified sequences for fetal cochlear, eyes and brain libraries and adult bone and stomach datasets. Only 11,913 human cochlear sequences out of fifteen thousand deposited (dbEST Library ID.371 [13,14]) are annotated in Unigene. We mapped over 98% (all but 276 – area 3 in inset of Fig.2 showing sequences not available in Unigene) of the ESTs in the Morton fetal cochlear library to specific regions in the human genome and genomes of laboratory organisms. Of the unmapped sequences, most correspond to highly conserved regions that can be exactly matched to dozens of proteins in a variety of organisms. The remaining unmapped ESTs seem to be formed by nonspecific recombination events and cannot be confidently attributed to a specific

gene or genome. Non-human contaminations in the dataset (259, area 4 in Fig.2) come from laboratory organisms – mainly yeast, E.coli, phages and cloning vectors, but there are also single occurences of such unexpected species as worm and mouse. Among about five thousand genes identified, almost 2000 genes are represented by single ESTs. Less than 200 genes are supported by ten or more sequences. The most abundant mRNAs were for extracellular matrix genes. This can be explained by the importance of structural support in cochlea. We note that this class of proteins acounts for almost half of nonsyndromic deafness genes.

Less than 10% of all our cochlea sequences were deposited with gene-relevant information in their headers, while 41% of the sequences were annotated based on results of BLAST searches against GenBank databases in early 2000s. Almost 80% from this set are annotated in the latest build of Unigene, although about 8% of these annotations remain hypothetical. We selected many different isoforms among ESTs clustered in the same Unigene clusters. In addition to the 4058 Unigene clusters, we determined almost 1000 additional loci, many of which might represent novel genes or isoforms of known genes (areas 1 and 4 in Fig.2). We found about 20% potential genomic contaminations in the dataset and 1% of sequence flips in EST sequences. Many transcripts corresponding to ESTs present in the dataset might not be expressed as proteins, but instead are degraded by nonsense-mediated mRNA decay or other cell surveillance mechanisms. We revealed a number of incomplete, truncated mRNAs in the library, confirming this possibility.

The inset of Figure 2 shows how sequences extracted from the fetal inner ear and not classified by Unigene are mapped to the human genome and genomes of other species (human pathogens and laboratory organisms). Comparison of our mappings to alignments produced by popular tools, such as BLAST [6] and BLAT [4], shows that our solutions are essentially the same. These other tools, however, offer the best solutions among several other top scoring results, thus requiring post-processing of results, often manually. We note that most of our novel genes are also suggested in the AceView database [15] and are being incorporated into the next build of the human genome. On the one hand, we consider it as another confirmation of the reliability of our findings. On the other hand, we note that the subject of this work is analysis of hearing-specific genes and this was not done by the authors of AceView, GeneScan and other global gene-finding programs.

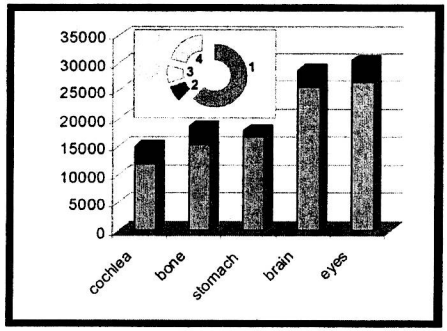

Fig. 2. A bar-chart of sequences of organ-specific libraries classified (white base) and not classified (black top) into Unigene entries. Inset shows our mappings of non-classified cochlear ESTs. Sequences in areas: (1) may be novel isoforms of known genes; (2) are non-human genes; (3) are ambiguous; 4) map to unannotated regions in the human genome