Carlos E. Ferreira
Satoru Miyano
Peter F. Stadler (Eds.)

LNBI 6268

Advances in Bioinformatics and Computational Biology

5th Brazilian Symposium on Bioinformatics, BSB 2010
Rio de Janeiro, Brazil, August/September 2010
Proceedings

Springer

Lecture Notes in Bioinformatics 6268

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

Subseries of Lecture Notes in Computer Science

Carlos E. Ferreira Satoru Miyano
Peter F. Stadler (Eds.)

Advances in Bioinformatics and Computational Biology

5th Brazilian Symposium on Bioinformatics, BSB 2010
Rio de Janeiro, Brazil, August 31-September 3, 2010
Proceedings

Springer

Volume Editors

Carlos E. Ferreira
Universidade de São Paulo
IME-USP (MAC)
Rua do Matão, 1010, 05508-090 São Paulo, SP, Brazil
E-mail: cef@ime.usp.br

Satoru Miyano
University of Tokyo
Institute of Medical Science, Human Genome Center
4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan
E-mail: miyano@ims.u-tokyo.ac.jp

Peter F. Stadler
University of Leipzig
Department Computer Science and Interdisciplinary Center for Bioinformatics
Bioinformatics Group
Härtlestr. 16-18, 04107 Leipzig, Germany
E-mail: studla@bioinf.uni-leipzig.de

Library of Congress Control Number: 2010931861

CR Subject Classification (1998): J.3, I.2, F.1, H.2.8, I.5, H.3

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743
ISBN-10 3-642-15059-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-15059-3 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

This volume contains the accepted full papers and extended abstracts of the 5th Brazilian Symposium on Bioinformatics held in Búzios, Rio de Janeiro, Brazil from August 31 to September 3, 2010.

The first three meetings of this series, which took place 2002, 2003, and 2004, were called WOB (Workshop on Bioinformatics). In 2005, the conference got its current name BSB and has since published its proceedings as a special issue of the series *Lecture Notes in Bioinformatics* (volumes 3594/2005, 4643/2007, 5167/2008, and 5676/2009).

Its topics of interest vary in many areas of bioinformatics, including sequence analysis, motifs, and pattern matching; biomedical text mining; biological databases, data management, integration; biological data mining; structural, comparative, and functional genomics; protein structure, modeling, and simulation; gene identification and regulation; gene expression analysis; gene and protein interaction and networks; molecular docking; molecular evolution and phylogenetics; computational systems biology; computational proteomics; statistical analysis of molecular sequences; algorithms for problems in computational biology; as well as applications in molecular biology, biochemistry, genetics, and associated subjects.

We would like to thank all referees and Program Committee members for their careful work in preparing this proceedings volume. Also, we want to acknowlegde the local organizers and their staff for making this meeting possible.

September 2010

Carlos E. Ferreira
Satoru Miyano
Peter F. Stadler

Organization

BSB 2010 was promoted by the Brazilian Computing Society (SBC) and was organized by the Informatics Department of the Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio).

Executive Committee

Conference Chair:	Sergio Lifschitz (PUC-Rio, Brazil)
Program Chairs:	Carlos E. Ferreira (University of São Paulo, Brazil)
	Satoru Miyano (University of Tokyo, Japan)
	Peter F. Stadler (University of Leipzig, Germany)
Steering Comittee:	André C. de Carvalho (University of São Paulo, Brazil)
	Carlos E. Ferreira (University of São Paulo, Brazil)
	Katia Guimarães (Federal University of Pernambuco, Brazil)
	Sergio Lifschitz (PUC-Rio, Brazil)
	Francisco M. Salzano (UFRGS, Brazil)
	João Carlos Setubal (Virginia Tech, USA)
	Osmar Norberto de Sousa (PUC-RS, Brazil)

Program Committee

Emil Alexov	Clemson University, USA
Nalvo F. Almeida Jr.	UFMS, Brazil
Masanori Arita	University of Tokyo, Japan
Junior Barrera	USP, Brazil
Ana Lucia C. Bazzan	UFRGS, Brazil
Igor Berezovsky	University of Bergen, Norway
Marcelo M. Brigido	UnB, Brazil
André C.P.L. Carvalho	USP, Brazil
Saikat Chakrabarti	NCBI, NIH, USA
Zanoni Dias	Unicamp, Brazil
Jessica Fong	NCBI, NIH, USA
Oxana Galzitskaya	Institute of Protein Science, Russia
Kátia S. Guimarães	UFPE, Brazil
Ronaldo F. Hashimoto	USP, Brazil
Paul Horton	AIST, Japan

Maricel Kann	UMBC, USA
Carl Kingsford	UMIACS, UMD, USA
Kengo Kinoshita	Tohoku University, Japan
Eugene Krissinel	EBI, UK
Ney Lemke	UNESP, Brazil
Thomas Madej	NIH, USA
Wojtek Makalowski	University of Münster, Germany
Ion Mandoiu	University of Connecticut, USA
Natalia F. Martin	Embrapa, Brazil
Wellington S. Martins	UFG, Brazil
Hideo Matsuda	Osaka University, Japan
José Carlos M. Mombach	UFSM, Brazil
Anna Panchenko	NIH, USA
Ron Pinter	Technion, Israel
Teresa Przytycka	NIH, USA
Marie-France Sagot	University Lyon, France
Cenk Sahinalp	Simon Fraser University, Canada
Francisco M. Salzano	UFRGS, Brazil
Hagit Shatkay	Queen's University, Canada
Alexander Schliep	Rutgers University, USA
Benjamin Shoemaker	NIH, USA
Marcilio M.C.P. Souto	UFRN, Brazil
Jurek Tiuryn	University of Warsaw, Poland
Maria Emilia T. Walter	UnB, Brazil
Ryo Yoshida	Institute of Statistical Math., Japan
Alex Zelikovsky	Georgia State University, Brazil
Jie Zheng	NIH, USA

Table of Contents

Full Papers

Extended Abstracts

Evolution of the Long Non-coding RNAs MALAT1 and MENβ/ϵ

Peter F. Stadler

Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany; Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany; Fraunhofer Institut für Zelltherapie und Immunologie, Perlickstraße 1, D-04103 Leipzig, Germany; Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria; Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

Abstract. MALAT1 is one of the best-conserved long ncRNAs in mammals and shares several characteristics, among them nuclear retention and a non-standard processing of its 3' end, with the longer, but less well conserved, adjacent MENβ RNA. We show that MALAT1 is conserved among gnathostomes (with the possible exception of birds), while MENβ likely originated in the mammalian stem lineage. Evolutionary conserved features of both transcripts are discussed, including RNA secondary structure motifs and short RNA processing products.

Keywords: MALAT1, MENβ, VINC, NEAT1, TncRNA.

1 Introduction

A plethora of diverse non-coding RNAs have been discovered during the last decade, collectively demonstrating that a large fraction of the genomes of higher eukaryotes is transcribed into mRNA-like non-protein-coding transcripts (mlncRNAs) [1, 2]. The evolutionary history of these transcripts is still poorly understood. With very few exceptions, only global statistical information is available to demonstrate that a large number of ncRNAs is under stabilizing selection [3–5]. Nevertheless, most mlncRNAs are poorly conserved at sequence level compared to other functional transcripts [4, 6]. Detailed evolutionary information is available on many families of protein-coding genes and structured "house-keeping" RNAs. For ncRNAs, it is compiled in the `Rfam` database [7] and in specialized data repositories for microRNAs (`miRBase` [8]) and snoRNAs (`snoRNA-LBME-db` [9]). In contrast, evolutionary and phylogenetic information on mlncRNAs is currently neither collected nor organized in a systematic way.

Detailed case-studies are available for only a few prominent transcripts, such as the imprinting-related mammalian H19 ncRNA [10], the *Drosophila roX* RNAs [11], and the eutherian *Xist* transcript [12, 13]. The latter originated by pseudogenization of the protein-coding *Lnx3* gene in the eutherian ancestor [12] under inclusion of repetitive elements [13], which also gave rise to conserved

C.E. Ferreira, S. Miyano, and P.F. Stadler (Eds.): BSB 2010, LNBI 6268, pp. 1–12, 2010.

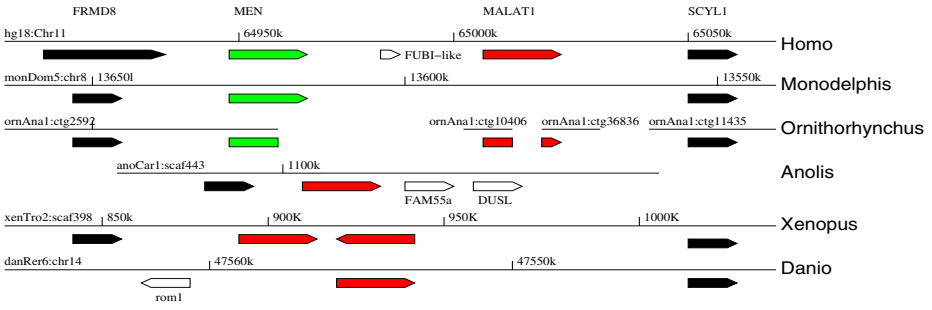


Fig. 1. Overview of the MEN/MALAT locus in different vertebrate species. The non-coding MALAT transcript is linked to at least one of FRM8 and SCYL1 in all species except the stickleback. In *Xenopus*, MALAT1 is duplicated, with one copy arranged in reverse direction.

secondary structure features [14]. *Xist* is one of only three highly expressed poly-adenylated ncRNAs that show strong nuclear retention [15].

The other two transcripts, NEAT1 and NEAT2/MALAT1, are the topic of this contribution. They are located in close genomic proximity at the human 11q13.1 locus. NEAT1 also exists in a longer isoform, known as MENβ [16]. Recent studies showed that MALAT1 and MENβ share a number of peculiar features. Both transcripts are spliced only infrequently [15], a feature that is atypical for transcripts of their size. Most surprisingly, their 3' ends are processed in a non-standard way: RNase P cleaves the primary transcripts before a tRNA-like element [16, 17], which is then processed into an independent cytoplasmic ncRNA. The evolution of these small tRNA-like ncRNAs was studied already in some detail [18].

The $\sim$ 8.7kb MALAT1 transcript (also known as NEAT2 and AlphaTFEB) is overexpressed in a variety of different carcinomas [19–21]. Knockdown of MALAT1 by shRNA implicates the transcript in cell cycle progression [22]. As noted in [15], MALAT1 is exceptionally well-conserved for a long ncRNA. The same study noticed the presence of a homolog in the opossum genome and reported an "apparent absence of the transcript in non-mammalian species". The subnuclear localization of MALAT1 is concentrated in the SC35 splicing domains, indicating a function in pre-mRNA metabolism [15].

The NEAT1 transcript, which has a size of $\sim$ 3.2 kb, is also responsive to diverse disease states. It is induced in mouse brain during infection by Japanese encephalitis virus and rabies virus, and hence was termed "Virus Inducible Non-Coding RNA" (VINC) in [23, 24]. It is located at the *Men1* ("multiple endocrine hypoplasia 1") locus, and hence was named MENϵ in [16, 25]. The transcript contains the shorter "trophoblast non-coding RNA" (TncRNA) that suppresses the expression of major histocompatibility antigens [26–28]. The bovine NEAT1 orthologue shows increasing expression levels during development of cattle muscle [29]. The same locus also produces a much longer isoform ($\sim$ 20kb), called

MENβ. Several groups recently reported the involvement of MENϵ and MENβ in the organization of the paraspeckles [16, 25, 30, 31], reviewed in [32]. Protein interaction regions in the VINC/NEAT1/Menϵ RNA are investigated in [33]. The main biological function of NEAT1/MENϵ is the regulation of gene expression by restricting nuclear export [30, 34].

2 Materials and Methods

Genomic sequence data were retrieved from `ensembl` (v.57). In addition, ESTs and unassembled genomic DNA from NCBI GenBank, and high throughput sequencing data from GEO were analyzed. Initial homology searches were performed with `blast` and extended to global alignments using several alignment tools, including `custalw`, `dialign`, `muscle`, and `mafft`. RNA secondary structures were investigated using the `Vienna RNA Package`. The UCSC genome browser was used for visualization. Due to length restrictions on the manuscript, details and references are given throughout the Results section where necessary.

3 Results

Syntenic Conservation. The genomic location of MENβ/MALAT1 is flanked by FRMD8 ("FERM domain containing 8", a.k.a. FKSG44) on the 5' side and by the highly conserved kinase-like gene SCYL1 throughout Eutheria. A small "FUBI-like" gene (AP000769) is located between MENβ and MALAT1. All these transcripts share reading direction, Fig. 1. The MALAT1 homolog is also linked to FRMD8 and/or SCYL1 in other vertebrates, and the arrangement [5'-FRMD8-MALAT1-SCYL1-3'] appears to be the ancestral state. The assembly of the elephant shark genome, however, does not provide sufficient evidence to test this hypothesis directly; no MALAT1 homolog was detectable in the lamprey genome. In teleosts, synteny is broken between FRMD8 and MALAT1, while

Table 1. Approximate locations of MENβ in several mammalian genomes. The coordinates refer to the (mostly unspliced) ESTs located in the approximate region identified by `blastn` as homologous to human MENβ. The 5'-end of the menRNA is also listed. Dots indicate that there are no ESTs near the position of the menRNA.

Species	Assembly	Chr.	±	5'-MENβ	3'-MENβ	5'-menRNA
Homo sapiens	hg19	11	+	65190269	65213007	65213012
Macaca mulatta	rheMac2	14	−	9009052	...	8979130
Mus musculus	mm9	19	−	5845579	5824708	5824707
Rattus norvegicus	rn4	1	−	208481740	208455951	208456537
Canis familiaris	canFam2	18	−	54794495	54775188	54775783
Equus caballus	equCab2	12	+	25591044	...	25613257
Bos taurus	bosTau4	29	+	45474754	45495959	45495960
Ornithorhynchus anatinus	ornAna1	ctg2592	−	13707	...	—

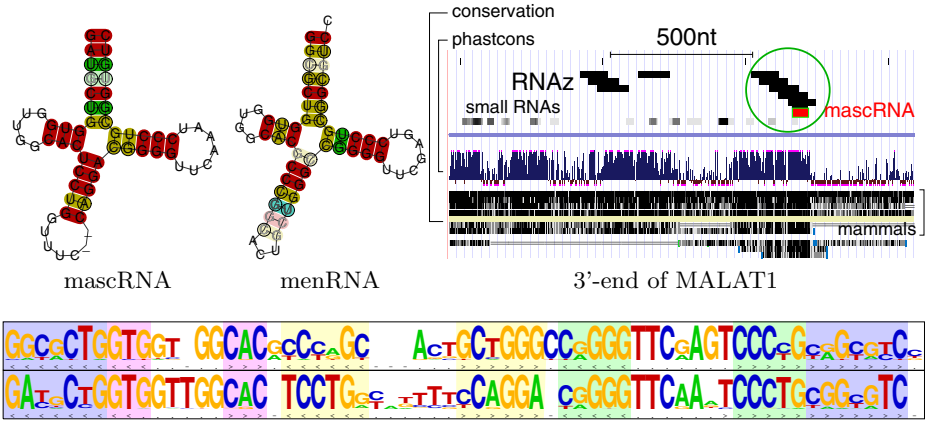


Fig. 2. RNA secondary structure in MALAT and MENβ. By far the best conserved structured signals are mascRNA [17] (upper left) and menRNA [16] (upper middle), respectively. The upper right panel summarizes the `RNAz` predictions of structured RNAs in MALAT1. Besides the mascRNA and the hairpin structure described in [16] at the 3' end (encircled), there is only one additional structured region about 600nt upstream of the mascRNA. Below, the aligned sequence logos of menRNA (above) and mascRNA (below) clearly show that the two ncRNAs are homologous.

SCYL1 is located on a different scaffold in the lizard genome. In *Xenopus* we find two divergent copies of the MALAT1 sequence in tail-to-tail orientation.

Surprisingly, the entire region is missing in all four sequenced bird genomes (chicken, turkey, zebrafinch, and duck). No plausible homolog of SCYL1, FRMD8 (using `tblastn`), or MALAT1 (using `blastn`) are detectable. There are two possible explanations: (1) Birds lost the entire genomic locus. (2) FRMD8-MALAT1-SCYL1 is located on a microchromosome, which are known to be underrepresented in the chicken genome assembly [35]. Given that MALAT1 can be identified in all other sequenced gnathostomes and the high level of sequence conservation of the two flanking genes (whose functions appear to be unrelated to that of MALAT1 and MENβ/ϵ), we suspect that we see a data bias rather than a true loss of the entire locus.

MENβ is clearly present in all mammals. Within eutheria, the homology is easy to establish and the loci can be found by simple `blastn` searches using e.g. the human sequence as query. In several species the presence of the MENβ and/or MENϵ transcripts is supported by (predominantly unspliced) ESTs mapping to the location of the `blastn` hit, see Tab. 1. Due to gaps, break-points between scaffolds, and inaccuracies in the genome assemblies, it is a bit more complicated to trace MENβ in marsupials and in platypus. Unambiguous `blastn` hits to large portions of MENβ are easily obtained, however. No EST support is available in marsupials. The expression of a MENβ transcript in platypus is supported by a handful of ESTs (EY202075, EY201405, EH004653, EG34158) as well as several 454 reads listed in `ensembl` (v.57).

Table 2. Approximate positions of the MALAT1 homologs in non-mammalian vertebrates determined from EST information. Data are from the UCSC genome browser, except for lizard, which was taken from `ensembl` (version 57).

Species	Assembly	Chr.	±	5'-MALAT1	3'-MALAT1	5'-mascRNA
Anolis carolinensis	anoCar1	s.443	+	1097270	1104518	1104209
Xenopus tropicalis +	xenTro2	s.398	+	900641	910280	910287
Xenopus tropicalis -	xenTro2	s.398	−	936780	924118	925349
Danio rerio	Zv8	14	−	47564239	47570301	47564238
Tetraodon nigroviridis	tetNig2	1	−	8135090	...	8130318
Takifugu rubripes	fr2	Un	+	240961971	...	240966770
Gasterosteus aculeatus	gasAcu1	IV	−	5027768	...	5022122
Oryzias latipes	oryLat2	10	+	8151439	8156861	8156576

MascRNA and menRNA. Both MALAT1 and MENβ have a highly structured 3' end, consisting of a hairpin structure, the genomically encoded polyA motive, and the tRNA-like structure that is cleaved off and becomes a stable cytoplasmic ncRNA. This common structure is described in some detail in [17] for mascRNA (MALAT1 associated RNA) and in the supplemental material of [16] for menRNA, see also Fig. 2. The menRNA is by far the best-conserved part of the MENβ transcript. It is easily identified in the two metatheria (*Monodelphis* and *Macropus*) [18]. Although a menRNA homolog is missing from both shotgun traces and the genome assembly of platypus, it is possible to identify other homologous sequences near the 3' end of MENβ. In contrast, no potential ortholog of MENβ/ϵ or menRNA can be found in outside mammalia.

A short region in the lizard genome aligned in the UCSC genome browser to the menRNA region (anoCar1, scaffold 944:1210-1465[-]) cannot be identified unambiguously as the 3'-end of a MENβ ortholog, because a `blastn` search yields 42 similar homologs throughout the lizard genome. Their sequences were retrieved together with about 200nt flanking sequence and aligned (with `clustalw`) to the corresponding regions surrounding mascRNAs and menRNAs. All lizard sequences clearly appear as monophyletic group in this tree (Fig. 3), indicating lineage-specific proliferation of mascRNA. The data are consistent with (but do not provide a conclusive proof for) the origin of MENβ through a duplication of MALAT1, probably in the mammalian stem lineage. The frog genome contains two divergent, and hence ancient, copies of MALAT1 in an unexpected tail-to-tail configurations. The phylogenetic analysis does not provide any evidence that one of these copies might be the ancestor of MENβ.

MascRNA and menRNA are clearly homologous [18], Fig. 2, and the similarities of MENβ and MALAT1 extend upstream of the cleavage site to include a hairpin structure and the genomically encoded poly-A tract [16]. At least parts of the MENβ thus may have arisen from a duplication of MALAT1 in the mammalian ancestor. The lack of recognizable homologies further towards the 5' end could be explained by the poor overall conservation of MENβ.

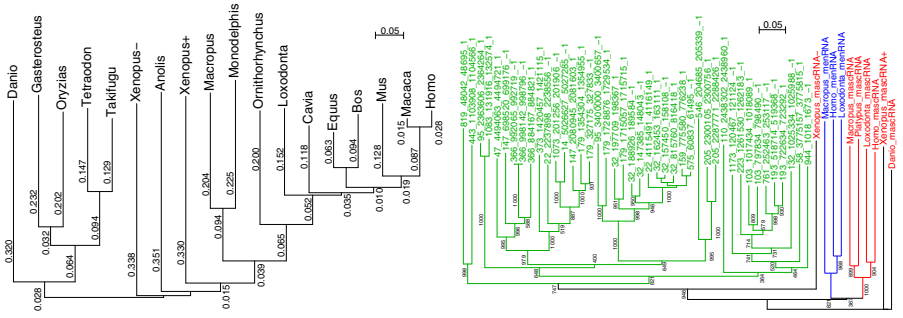


Fig. 3. Left: The vertebrate MALAT1 sequences are alignable and a neighbor-joining tree reproduces the established vertebrate phylogeny quite well, except for the positioning of the two marsupials (*Monodelphis* and *Macropus*) outside of the platypus sequence. **Right:** Neighbor-joining tree of the mascRNA and menRNA loci with about 200nt flanking sequence on both sides. Lizard sequences are shown in green, tetrapod mascRNAs in red, and tetrapod menRNAs blue.

Gene Phylogeny. The MALAT1 transcript is easily recognizable in all mammals [15]. Significant `blastn` hits can also be found in the available genome data of all five sequences teleosts, the elephant shark, the frog, and the lizard. In particular, the mascRNA can be identified unambiguously [18]. In addition to sequence homology, EST data can be used to determine the approximate extent of the MALAT1 transcript in several non-mammalian gnathostomes, see Tab. 2. Approximate full-length sequences were retrieved from the genomic data compiled in the UCSC genome browser and aligned using `clustalw` [36]. Visual inspection of the alignment shows that it indeed consists of homologous sequences. The neighbor-joining tree constructed from this alignment is shown in Fig. 3. It conforms to the established phylogeny of vertebrates with the exception of the relative position of marsupials and platypus, which can be interpret as a long branch attraction artifact.

Promoters. Not much is known about the transcriptional regulation of MALAT1 and MENβ/ϵ. There is evidence for alternative transcription start sites for the human MALAT1 transcript(s): In addition to the longer transcript reported e.g. in [15], a shorter isoform ($\sim$ 7kb) is produced from a CREB-sensitive promoter that can be stimulated by oxytocin [37]. This start sites matches that of mouse *hepcarcin* [20]. Fig. 4 shows that the core promoter region is well conserved within mammals. The alignment of all vertebrate MALAT1 sequences, however, does not provide evidence for a conservation of this feature in other gnathostomes.

Conserved Secondary Structure Elements. Many ncRNAs exhibit evolutionarily conserved secondary structures. Surveys of the human genome [38, 39], for instance, identified tens of thousands of conserved structural motifs. The alignment of the MALAT1 sequences was screened with `RNAz` [38]. As expected,

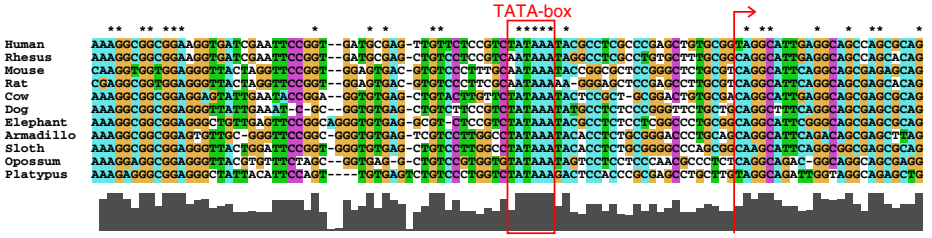


Fig. 4. Conserved promoter of the shorter form of MALAT1 reported in [37], which fits the 5' end of the mouse "hepcarcin" RNA

the mascRNA locus and the adjacent conserved hairpin structure just upstream of the RNase P processing site [16] was identified as structured region. Despite the size of the MALAT1 transcripts, however, `RNAz` detected only one additional structured location about 600nt upstream of the processing site, see Fig. 2. For MENβ, only the menRNA and a small structured region near the 5'-end of the transcript were detected.

Small Processing Products. A plethora of different types of small RNA products have been detected in eukaryotic genomes, ranging from microRNAs, piRNAs, and endogenous siRNAs [40, 41] to multiple families of small RNAs associated with mRNAs [42, 43]. Several studies using modern high throughput sequencing technologies reported that well-known ncRNA loci are also processed to give rise to small RNAs. MicroRNA precursor hairpins, for instance, are frequently processed to produce additional "off-set RNAs" (moRNAs) that appear to function like mature miRs [44, 45], tRNAs are cleaved to yield multiple shorter products [46–49], snoRNAs frequently give rise to specific miRNA-like short RNAs [50], and a functional short RNA product derives from a vault RNA [51, 52]. The production of small RNA products is a ubiquitous phenomenon that is strongly associated with secondary structure [53].

Here, several published short-read sequencing data sets as well as an extensive library of short RNAs from human brains kindly provided by Philipp Khaitovich [45, 51] is re-evaluated. After mapping the entire dataset to the genome with `segemehl` [54], the subset localized in the MALAT1/MENβ region was extracted. Both MALAT1 (Fig. 5) and MENβ (Fig. 6) give rise to relatively high levels of

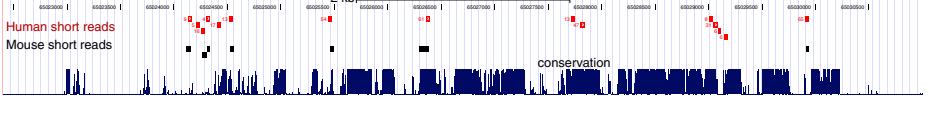


Fig. 5. Conservation of short read expression between human (top) and mouse (below). For comparison, sequence conservation is shown at the bottom of the browser image. Only the most highly expressed blocks of reads are indicated. The genome browser panel covers exactly the annotated human MALAT1 transcript.

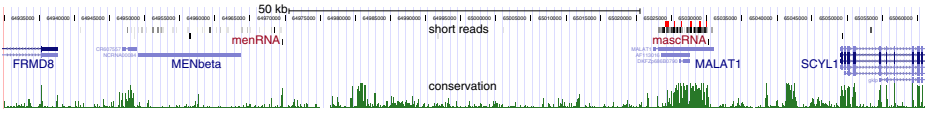


Fig. 6. A diverse set of short-reads is also produced over the complete length of the MENβ transcript. The tRNA-like menRNA is located at small highly conserved locus the very end of the solid EST bar.

short RNA products with a length < 30nt. A comparison of the human and mouse reads shows that several of the most highly expressed locations in the human libraries are also detectable in the much smaller mouse data set GEO: GPL7195 [55]. Surprisingly, this syntenic conservation neither correlates with evolutionary conservation of either sequence or secondary structure. In contrast to MALAT1 and MENβ, most protein-coding transcripts do not give rise to similar patterns of short RNA product. It is unclear whether the processing into small RNAs is a generic feature of nuclear-retained transcripts. The *Xist* transcript, which behaves similarly to MALAT1 and MENβ/MENϵ in several respects [15], does not produce any short reads in any of the investigated libraries. Either *Xist* is simply not expressed under any of the conditions/tissues used here, or its processing is indeed distinct from that MALAT1 and MENβ/MENϵ.

4 Concluding Remarks

The detailed investigation of the MALAT1/MENϵ locus reveals several surprising facts about MALAT1 (conservation at least throughout gnathostomes, the presence of an internal promoter that is conserved across mammals) and MENβ (a probable origin in the mammalian stem lineage) and highlights several commonalities between them: the previously described processing of the 3'-ends by RNase P including the production of small tRNA-like cytoplasmic ncRNA [16, 17], the absence of conserved secondary structures almost everywhere else in the transcript, and the production of many well-defined short RNA products.

On the other hand, this case-study highlighted serious practical difficulties in the comparative analysis of long mlncRNAs. The generally low level of sequence conservation calls for alignment tools that are optimized for this problem. Current alignment editors cannot effectively handle sequences several kb in length and landmarks, such a promoter elements, structured RNA motifs, ESTs, or splice sites cannot be annotated directly in the alignment. Only a few "finished genomes" provide sequences that do not contain gaps or assembly errors over a length of several 10000nt, calling for more efficient ways to explicitly treat missing data in multiple sequence alignents. Thus, detailed case studies are not only of interest in their own right, but are also a necessary prerequisite for the design and development of computational tools that can efficiently assist the analysis of long ncRNAs.

From the biological point of view, the most interesting question concerns the evolutionary origin of mlncRNAs. So far, *Xist* is the only example for which a satisfactory answer — loss of coding capacity of the *Lnx3* transcript and inclusion of adjacent repetitive sequence elements — is known. In the case of MALAT1 and MENβ no candidate for a possible evolutionary precursor could be identified. It seems that mascRNA and menRNA originally derive from tRNAs, similar to, e.g., BC1 and BC200 [56]. MALAT1 and MENβ, like *Xist*, thus are probably composites deriving from several ancestral genomic elements. Interestingly, the large 3' part of MENβ that is not part of the NEAT1/MENϵ transcript consists to a large extent of old SINE (mostly Alu) and a few LINE elements. In contrast, MALAT1 and NEAT1/MENϵ are (nearly) devoid of annotated repeat-derived sequences.

Acknowledgements. Thanks to Phillip Khaitovich for access to short-read sequencing data, to David Langenberger and Steve Hoffmann for access to their short-read maps, and to Manja Marz for comments on the manuscript.

References

1. The ENCODE Project Consortium: Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799–816 (2007)
2. Maeda, N., Kasukawa, T., Oyama, R., Gough, J., Frith, M., Engström, P.G., Lenhard, B., Aturaliya, R.N., Batalov, S., Beisel, K.W., Bult, C.J., Fletcher, C.F., Forrest, A.R., Furuno, M., Hill, D., Itoh, M., Kanamori-Katayama, M., Katayama, S., Katoh, M., Kawashima, T., Quackenbush, J., Ravasi, T., Ring, B.Z., Shibata, K., Sugiura, K., Takenaka, Y., Teasdale, R.D., Wells, C.A., Zhu, Y., Kai, C., Kawai, J., Hume, D.A., Carninci, P., Hayashizaki, Y.: Transcript annotation in FANTOM3: Mouse Gene Catalog based on physical cDNAs. PLoS Genetics 2, e62 (2006)
3. Ponjavic, J., Ponting, C.P., Lunter, G.: Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. 17, 556–565 (2007)
4. Marques, A.C., Ponting, C.P.: Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome Biol. 10, R124 (2009)
5. Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., Cabili, M.N., Jaenisch, R., Mikkelsen, T.S., Jacks, T., Hacohen, N., Bernstein, B.E., Kellis, M., Regev, A., Rinn, J.L., Lander, E.S.: Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458, 223–227 (2009)
6. Pang, K.C., Frith, M.C., Mattick, J.S.: Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genetics 22, 1–5 (2006)
7. Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R., Bateman, A.: Rfam: updates to the RNA families database. Nucleic Acids Res. 37, D136–D140 (2009)
8. Griffiths-Jones, S., Saini, H.K., van Dongen, S., Enright, A.J.: miRBase: tools for microRNA genomics. Nucleic Acids Res. 36, 154–158 (2008)
9. Lestrade, L., Weber, M.J.: snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res. 34, D158–D162 (2006)

10. Smits, G., Mungall, A.J., Griffiths-Jones, S., Smith, P., Beury, D., Matthews, L., Rogers, J., Pask, A.J., Shaw, G., VandeBerg, J.L., McCarrey, J.R., SAVOIR Consortium,, Renfree, M.B., Reik, W., Dunham, I.: Conservation of the H19 noncoding RNA and H19-IGF2 imprinting mechanism in therians. Nat. Genet. 40, 971–976 (2008)
11. Park, S.W., Kang, Y.I., Sypula, J.G., Choi, J., Oh, H., Park, Y.: An evolutionarily conserved domain of roX2 RNA is sufficient for induction of H4-Lys16 acetylation on the Drosophila X chromosome. Genetics 177, 1429–1437 (2007)
12. Duret, L., Chureau, C., Samain, S., Weissenbach, J., Avner, P.: The *Xist* RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science 312, 1653–1655 (2006)
13. Elisaphenko, E.A., Kolesnikov, N.N., Shevchenko, A.I., Rogozin, I.B., Nesterova, T.B., Brockdorff, N., Zakian, S.M.: A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. PLoS One 3, e2521 (2008)
14. Maenner, S., Blaud, M., Fouillen, L., Savoye, A., Marchand, V., Dubois, A., Sanglier-Cianférani, S., Van Dorsselaer, A., Clerc, P., Avner, P., Visvikis, A., Branlant, C.: 2-D structure of the A region of *Xist* RNA and its implication for PRC2 association. PLoS Biol. 8, e1000276 (2010)
15. Hutchinson, J., Ensminger, A.W., Clemson, C.M., Lynch, C.R., Lawrence, J.B., Chess, A.: A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. BMC Genomics 8, 39 (2007)
16. Sunwoo, H., Dinger, M.E., Wilusz, J.E., Amaral, P.P., Mattick, J.S., Spector, D.L.: MEN ϵ/β nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. Genome Res. 19, 347–359 (2009)
17. Wilusz, J.E., Freier, S.M., Spector, D.L.: 3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. Cell 135, 919–932 (2008)
18. Marz, M., Stadler, P.F.: Homology search for small structured ncRNAs. İn: Hartmann, R., Bindereif, A., Schön, A., Westhof, E. (eds.) Handbook of RNA Biochemistry, 2nd edn. Wiley VCH, Weinheim (in press, 2010)
19. Ji, P., Diederichs, S., Wang, W., Böing, S., Metzger, R., Schneider, P.M., Tidow, N., Brandt, B., Buerger, H., Bulk, E., Thomas, M., Berdel, W.E., Serve, H., Müller-Tidow, C.: MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene 22, 8031–8041 (2003)
20. Lin, R., Maeda, S., Liu, C., Karin, M., Edgington, T.S.: A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. Oncogene 26, 851–858 (2007)
21. Guffanti, A., Iacono, M., Pelucchi, P., Kim, N., Soldà, G., Croft, L.J., Taft, R.J., Rizzi, E., Askarian-Amiri, M., Bonnal, R.J., Callari, M., Mignone, F., Pesole, G., Bertalot, G., Rossi Bernardi, L., Albertini, A., Lee, C., Mattick, J.S., Zucchi, I., De Bellis, G.: A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. BMC Genomics 10, 163 (2009)
22. Guo, F., Li, Y., Liu, Y., Wang, J., Li, Y., Li, G.: Inhibition of metastasis-associated lung adenocarcinoma transcript 1 in CaSki human cervical cancer cells suppresses cell proliferation and invasion. Acta. Biochim. Biophys. Sin. (Shanghai) 42, 224–229 (2010)
23. Saha, S., Rangarajan, P.N.: Common host genes are activated in mouse brain by Japanese encephalitis and rabies viruses. J. Gen. Virol. 84, 1729–1735 (2003)

24. Saha, S., Murthy, S., Rangarajan, P.N.: Identification and characterization of a virus-inducible non-coding RNA in mouse brain. J. Gen. Virol. 87, 1991–1995 (2006)
25. Sasaki, Y.T.F., Ideue, T., Sano, M., Mituyama, T., Hirose, T.: MENϵ/β noncoding RNAs are essential for structural integrity of nuclear paraspeckles. Proc. Natl. Acad. Sci. USA 106, 2525–2530 (2009)
26. Peyman, J.A.: Mammalian expression cloning of two human trophoblast suppressors of major histocompatibility complex genes. Am. J. Reprod. Immunol. 45, 382–392 (2001)
27. Geirsson, A., Paliwal, I., Lynch, R.J., Bothwell, A.L., Hammond, G.L.: Class II transactivator promoter activity is suppressed through regulation by a trophoblast noncoding RNA. Transplantation 76, 387–394 (2003)
28. Geirsson, A., Lynch, R.J., Paliwal, I., Bothwell, A.L., Hammond, G.L.: Altered nuclear retention of Human trophoblast noncoding RNA suppresses CIITA promoter III activity in murine B-lymphocytes. Biochem. Biophys. Res. Commun. 301, 718–724 (2003)
29. Lehnert, S.A., Reverter, A., Byrne, K.A., Wang, Y., Nattrass, G.S., Hudson, N.J., Greenwood, P.L.: Gene expression studies of developing bovine longissimus muscle from two different beef cattle breeds. BMC Dev. Biol. 7, 95 (2007)
30. Chen, L., Carmichael, G.: RNAs containing inverted repeats in human embryonic stem cells: Functional role of a nuclear noncoding RNA. Mol. Cell 35, 467–478 (2009)
31. Clemson, C.M., Hutchinson, J.N., Sara, S.A., Ensminger, A.W., Fox, A.H., Chess, A., Lawrence, J.B.: An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol. Cell 33, 717–726 (2009)
32. Bond, C.S., Fox, A.H.: Paraspeckles: nuclear bodies built on long noncoding RNA. J. Cell Biol. 186, 637–644 (2009)
33. Murthy, S.U.M., Rangarajan, P.N.: Identification of protein interaction regions of VINC/NEAT1/Men epsilon RNA. FEBS Lett. 584, 1531–1535 (2010)
34. Scadden, D.: A NEAT way of regulating nuclear export of mRNAs. Mol. Cell 35, 395–396 (2009)
35. Douaud, M., Fève, K., Gerus, M., Fillon, V., Bardes, S., Gourichon, D., Dawson, D.A., Hanotte, O., Burke, T., Vignoles, F., Morisson, M., Tixier-Boichard, M., Vignal, A., Pitel, F.: Addition of the microchromosome GGA25 to the chicken genome sequence assembly through radiation hybrid and genetic mapping. BMC Genomics 9, 129 (2008)
36. Thompson, J.D., Higgs, D.G., Gibson, T.J.: CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. Nucl. Acids Res. 22, 4673–4680 (1994)
37. Koshimizu, T.A., Fujiwara, Y., Sakai, N., Shibata, K., Tsuchiya, H.: Oxytocin stimulates expression of a noncoding RNA tumor marker in a human neuroblastoma cell line. Life Sci. 86, 455–460 (2010)
38. Washietl, S., Hofacker, I.L., Stadler, P.F.: Fast and reliable prediction of noncoding RNAs. Proc. Natl. Acad. Sci. USA 102, 2454–2459 (2005)
39. Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E.S., Kent, J., Miller, W., Haussler, D.: Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput. Biol. 2, e33 (2006)
40. Moazed, D.: Small RNAs in transcriptional gene silencing and genome defence. Nature 457, 413–420 (2009)

41. Tanzer, A., Riester, M., Hertel, J., Bermudez-Santana, C.I., Gorodkin, J., Hofacker, I.L., Stadler, P.F.: Evolutionary genomics of microRNAs and their relatives. In: Caetano-Anolles, G. (ed.) Evolutionary genomics and System Biology, pp. 295–327. Wiley-Blackwell, Hoboken (2010)
42. Kapranov, P., Cheng, J., Dike, S., Nix, D., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermüller, J., Hofacker, I.L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Madhavan, G., Piccolboni, A., Sementchenko, V., Tammana, H., Gingeras, T.R.: RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316, 1484–1488 (2007)
43. Taft, R.J., Kaplan, C.D., Simons, C., Mattick, J.S.: Evolution, biogenesis and function of promoter-associated RNAs. Cell Cycle 8, 2332–2338 (2009)
44. Shi, W., Hendrix, D., Levine, M., Haley, B.: A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate. Nat. Struct. Mol. Biol. 16, 183–189 (2009)
45. Langenberger, D., Bermudez-Santana, C., Hertel, J., Hoffmann, S., Khaitovich, S., Stadler, P.F.: Evidence for human microRNA-offset RNAs in small RNA sequencing data. Bioinformatics 25, 2298–2301 (2009)
46. Jöchl, C., Rederstorff, M., Hertel, J., Stadler, P.F., Hofacker, I.L., Schrettl, M., Haas, H., Hüttenhofer, A.: Small ncRNA transcriptome analysis from *Aspergillus fumigatus* suggests a novel mechanism for regulation of protein-synthesis. Nucleic Acids Res. 36, 2677–2689 (2008)
47. Kawaji, H., Hayashizaki, Y.: Exploration of small RNAs. PLoS Genet. 4, e22 (2008)
48. Cole, C., Sobala, A., Lu, C., Thatcher, S.R., Bowman, A., Brown, J.W., Green, P.J., Barton, G.J., Hutvagner, G.: Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. RNA 15, 2147–2160 (2009)
49. Lee, Y.S., Shibata, Y., Malhotra, A., Dutta, A.: A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). Genes Dev. 23, 2639–2649 (2009)
50. Ender, C., Krek, A., Friedländer, M.R., Beitzinger, M., Weinmann, L., Chen, W., Pfeffer, S., Rajewsky, N., Meister, G.: A human snoRNA with microRNA-like functions. Mol. Cell 32, 519–528 (2008)
51. Stadler, P.F., Chen, J.J.L., Hackermüller, J., Hoffmann, S., Horn, F., Khaitovich, P., Kretzschmar, A.K., Mosig, A., Prohaska, S.J., Qi, X., Schutt, K., Ullmann, K.: Evolution of vault RNAs. Mol. Biol. Evol. 26, 1975–1991 (2009)
52. Persson, H., Kvist, A., Vallon-Christersson, J., Medstrand, P., Borg, A., Rovira, C.: The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs. Nat. Cell Biol. 11, 1268–1271 (2009)
53. Langenberger, D., Bermudez-Santana, C., Stadler, P.F., Hoffmann, S.: Identification and classification of small RNAs in transcriptome sequence data. In: Pac. Symp. Biocomput., vol. 15, pp. 80–87 (2010)
54. Hoffmann, S., Otto, C., Kurtz, S., Sharma, C., Khaitovich, P., Stadler, P.F., Hackermüller, J.: Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comp. Biol. 5, e1000502 (2009)
55. Babiarz, J.E., Ruby, J.G., Wang, Y., Bartel, D.P., Blelloch, R.: Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. Genes Dev. 22, 2773–2785 (2008)
56. Volff, J.-N., Brosius, J.: Modern genomes with retro-look: Retrotransposed elements, retroposition and the origin of new genes. Genome Dyn. 3, 175–190 (2007)

Granger Causality in Systems Biology: Modeling Gene Networks in Time Series Microarray Data Using Vector Autoregressive Models

André Fujita[1,⋆], Patricia Severino[2], João Ricardo Sato[3], and Satoru Miyano[1,4]

[1] Computational Science Research Program, RIKEN, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan
[2] Center for Experimental Research, Albert Einstein Research and Education Institute, Av. Albert Einstein, 627 - São Paulo, 05652-000, Brazil
[3] Center of Mathematics, Computation and Cognition, Universidade Federal do ABC, Rua Santa Adélia, 166 - Santo André, 09210-170, Brazil
[4] Human Genome Center, Institute of Medical Science, the University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan
andrefujita@riken.jp, psever@einstein.br, joao.sato@ufabc.edu.br, miyano@ims.u-tokyo.ac.jp

Abstract. Understanding the molecular biological processes underlying disease onset requires a detailed description of which genes are expressed at which time points and how their products interact in so-called cellular networks. High-throughput technologies, such as gene expression analysis using DNA microarrays, have been extensively used with this purpose. As a consequence, mathematical methods aiming to infer the structure of gene networks have been proposed in the last few years. Granger causality-based models are among them, presenting well established mathematical interpretations to directionality at the edges of the regulatory network. Here, we describe the concept of Granger causality and explore recent advances and applications in gene expression regulatory networks by using extensions of Vector Autoregressive models.

Keywords: Granger causality, vector autoregressive model, regulatory network, time series, gene expression data.

1 Introduction

In order to understand cell functioning as a whole under specific pathological conditions, it is necessary to uncover, at the molecular level, which genes are expressed at distinct time points and infer how their products interact. Interactions between genes are called gene regulatory networks. Due to the high number of genes involved in these networks, which involve activating or suppressing feedback loops, it is difficult to understand their dynamics by means of DNA microarray studies in cell cultures or patient's tissue due to both high costs

⋆ Corresponding author.

C.E. Ferreira, S. Miyano, and P.F. Stadler (Eds.): BSB 2010, LNBI 6268, pp. 13–24, 2010.

of data acquisition and labor intensive experiments. As a consequence, mathematical and statistical approaches for modeling and simulation *in silico* of these networks have become a field of intensive research.

In the particular case of time series gene expression analysis, a key challenge is the identification of the topology (the interconnection between genes) of these networks. The difficulty resides in the dependency in data, a problematic situation for classical methods of statistical analyses.

Vector Autoregressive (VAR) models are promising tools for the interpretation of time series gene expression data through the identification of Granger causalities between genes [11]. First, they are particularly useful for describing processes composed of locally interacting components. Second, statistical foundations for estimating VAR models from observed data, and computational algorithms are well understood and have been applied successfully in several areas such as Economics [22], Neuroscience [28] and more recently in Bioinformatics [6] [7] [23].

The idea is that temporal associations may contain information to suggest causality. More intuitively, it is known that a cause cannot come after the effect. Thus, if a time series x_t affects a time series y_t, the former should help improving the predictions of the latter.

To formalize Granger causality, suppose that $\Im_t$ is the information set containing all the relevant information available up to and including period t. Let $y_t(h|\Im_t)$ be the optimal (minimum MSE (mean squared error)) h-step predictor of the process y_t at origin t, based on the information in $\Im_t$. The corresponding forecast MSE will be denoted by $\Omega_t(h|\Im_t)$. The process x_t is said to cause y_t in Granger's sense if $\Omega_t(h|\Im_t) < \Omega_t(h|\Im_t \backslash \{x_s|s \leq t\})$ for at least one $h = 1, 2, \ldots$, where $\Im_t \backslash \{x_s|s \leq t\}$ is the set containing all the relevant information except for the information in the past and present of the x_t process. In other words, if y_t can be predicted more efficiently when the information in the x_t process is taken into account in addition to all other information, then x_t is Granger-causal for y_t.

In regulatory networks, a gene expression time series x_t is said to Granger cause another gene expression time series y_t, if x_t $(x_{t-1}, x_{t-2}, \ldots)$ provides statistically more significant information about future values of y_t than would be obtained by considering only the past values of y_t $(y_{t-1}, y_{t-2}, \ldots)$. Thus, past values of x_t contain information to predict the future values of y_t. In other words, a gene x_t is said to Granger-cause a gene y_t if it can be shown through a statistical test on lagged values of x_t, that x_t provides statistically significant information about present and future values of y_t. Notice that since this relationship is asymmetric, Granger causality may be interpreted as the direction of information flow [2]. Nevertheless, Granger causality is not "effective causality" in a deep sense of the word (in the Aristothelic sense) because the former is based solely on prediction and numerical calculations. However, it may give insights into biological molecular interactions.

In this work, VAR models will be discussed and applied to both toy models and actual biological data.

2 Methods

2.1 Vector Autoregressive Model (VAR)

In the linear case (when Granger causality is linearly dependent of the predictor's gene expression level), the most traditional way to identify Granger causality is by estimating VAR models. In order to simplify the description, only VAR models of order one will be presented, i.e., only one time lag will be analyzed. Generalizations to higher orders are straightforward.

The first order VAR model is described as shown:

$$\mathbf{y}_t = \mathbf{v} + \mathbf{A}_1 \mathbf{y}_{t-1} + \varepsilon_t \tag{1}$$

where T is the time series length, $\mathbf{y}_t$ is an $(n \times 1)$ vector of gene expressions (where n is the number of genes), $\mathbf{v}$ is an $(n \times 1)$ vector of intercepts, the normally distributed disturbance ε_t is an $(n \times 1)$ vector with zero mean and covariance matrix $\mathbf{\Sigma}$, and $\mathbf{A}_1$ is an $(n \times n)$ matrix of parameters (connectivity). The disturbances ε_t are serially uncorrelated, but may be contemporaneously correlated. Thus, $E(\varepsilon_t \varepsilon_t') = \mathbf{\Sigma}$, where $\mathbf{\Sigma}$ is an $(n \times n)$ matrix which may not be diagonal. It is important to highlight that, in this multivariate model, each gene may depend not only on its own past values, but also, on the past values of the other genes. Thus if $\mathbf{y}_{i,t}$ denotes the i^{th} gene in $\mathbf{y}_t$, the i^{th} row yields

$$\mathbf{y}_{i,t} = \mathbf{v}_i + \mathbf{a}_{i,1}\mathbf{y}_{1,t-1} + \mathbf{a}_{i,2}\mathbf{y}_{2,t-1} + \ldots + \mathbf{a}_{i,n}\mathbf{y}_{n,t-1} + \varepsilon_{i,t} \quad i = 1, \ldots, n \tag{2}$$

Due to its simplicity, the VAR model allows a simple way of identifying Granger causality in weakly stationary processes. A necessary and sufficient condition for the gene $\mathbf{y}_i$ being not Granger causal for the gene $\mathbf{y}_j$ is statistically testing if and only if $\mathbf{a}_{j,i} = 0$. Thus, Granger non-causality may be identified by looking at the autoregressive matrices of VAR models.

This model can be estimated by Ordinary Least Squares (OLS), simply by regressing each variable on the lags of itself and the other variables. It is possible to re-write (1) as $\mathbf{Z} = \mathbf{X}\beta + \mathbf{E}$, where $\mathbf{E} \sim N(\mathbf{0}_{(n\times 1)}, \mathbf{\Sigma})$ and

$$\mathbf{Z} = \begin{pmatrix} y_{1,2} & y_{2,2} & \cdots & y_{n,2} \\ y_{1,3} & y_{2,3} & \cdots & y_{n,3} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1,T} & y_{2,T} & \cdots & y_{n,T} \end{pmatrix}, \beta = \begin{pmatrix} \beta_{1,1} & \beta_{2,1} & \ldots & \beta_{n,1} \\ \beta_{1,2} & \beta_{2,2} & \ldots & \beta_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{1,n} & \beta_{2,n} & \ldots & \beta_{n,n} \end{pmatrix},$$

$$\mathbf{X} = \begin{pmatrix} y_{1,1} & y_{2,1} & \cdots & y_{n,1} \\ y_{1,2} & y_{2,2} & \cdots & y_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1,T-1} & y_{2,T-1} & \cdots & y_{n,T-1} \end{pmatrix}, \mathbf{E} = \begin{pmatrix} \varepsilon_{1,2} & \varepsilon_{2,2} & \ldots & \varepsilon_{n,2} \\ \varepsilon_{1,3} & \varepsilon_{2,3} & \ldots & \varepsilon_{n,3} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{1,T} & \varepsilon_{2,T} & \ldots & \varepsilon_{n,T} \end{pmatrix},$$

where the matrix $\mathbf{Z}$ and $\mathbf{X}$ are given.

The explicit solution of the OLS estimator is $\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$. Therefore, one can carry out separate regression analyses for each gene. In other words,

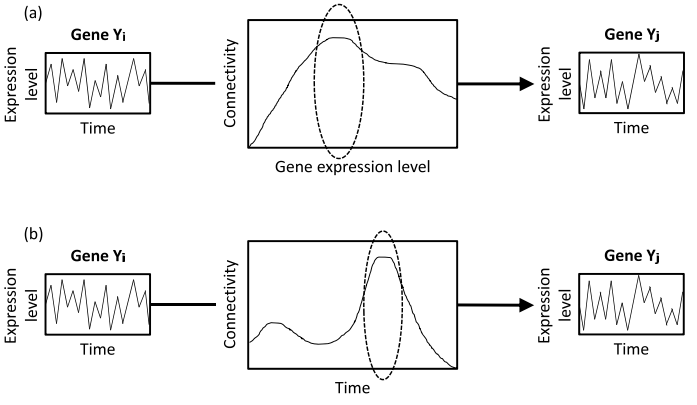


Fig. 1. Illustrative scheme of a regulatory network identified by (a) NVAR and (b) DVAR. (a) Nonlinear Granger causality from gene Y_i to gene Y_j. The connectivity changes in function of gene Y_i's expression level. (b) Time-varying Granger causality from gene Y_i to gene Y_j. The connectivity changes in function of time. Dashed regions indicate when the connectivity is statistically different from zero.

it is possible to separately estimate each column β_i of β: $\hat{\beta}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_i$, $i = 1, \ldots, n$, where $\mathbf{Z}_i$ is the i^{th} column of $\mathbf{Z}$.

In order to test if one gene is Granger causing another gene, the following hypothesis test is set: $\mathrm{H}_0 : \mathbf{C}\beta = 0$ versus $\mathrm{H}_1 : \mathbf{C}\beta \neq 0$.

It may be tested using the Wald statistic conveniently expressed as

$$W = (T-1)(\mathbf{C}\hat{\beta})'(\mathbf{C}\beta\mathbf{C}')^{-1}(\mathbf{C}\hat{\beta}) \tag{3}$$

where $\mathbf{C}$ is a matrix of contrasts ($\mathbf{C} = \mathbf{I}$, for instance). Under the null hypothesis, (3) has a limiting χ^2 distribution with rank($\mathbf{C}$) degrees of freedom.

2.2 Nonlinear Vector Autoregressive Model (NVAR)

Biological experiments show that when a certain gene is over expressed or knocked-down, the regulatory network structure might change. In other words, the gene regulatory network can also be gene expression dependent (presence of nonlinear regulation). This nonlinearity motivated the construction of a VAR model that is able to identify a wider range of regulations, i.e., nonlinear Granger causalities in time series gene expression data, namely NVAR [8] [19] [21].

The results provided by NVAR consist in the identification of Granger causalities that are gene expression dependent, i.e., the structure of the network changes depending on the gene expression levels of the predictors (Figure 1a). One illustrative example is the regulation of gene y_j by gene y_i only when the gene expression value of y_i belongs to a certain interval (represented by a dashed circle in Figure 1a) and they are independent outside this interval.

The NVAR model of order one is defined by $\mathbf{y}_t = \mathbf{v} + \mathbf{A}_1(\mathbf{y}_{t-1}) + \varepsilon_t$, where ε_t is a n-dimensional error vector of random variables with zero mean and covariance matrix $\mathbf{\Sigma}$, $\mathbf{v}$ is the intercept vector and $\mathbf{A}_1(y_{t-1})$ is the coefficient matrix given by

$$\mathbf{A_1(y_{t-1})} = \begin{pmatrix} a_{1,1}(\mathbf{y_{1,t-1}}) & a_{2,1}(\mathbf{y_{2,t-1}}) & \dots & a_{n,1}(\mathbf{y_{n,t-1}}) \\ a_{1,2}(\mathbf{y_{1,t-1}}) & a_{2,2}(\mathbf{y_{2,t-1}}) & \dots & a_{n,2}(\mathbf{y_{n,t-1}}) \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,n}(\mathbf{y_{1,t-1}}) & a_{2,n}(\mathbf{y_{2,t-1}}) & \dots & a_{n,n}(\mathbf{y_{n,t-1}}) \end{pmatrix}. \tag{4}$$

The main idea of NVAR is to set the matrix $\mathbf{A}_1$ as a function of gene expressions y_t. To model this regulatory network in a nonlinear fashion, one may consider the Cubic Splines expansion of the functions in order to estimate the nonlinear functions in $\mathbf{A}_1(\mathbf{y}_t)$.

A function $f(z)$ may be represented by a linear combination of splines functions $\phi_j(z)$. Therefore, considering the truncated spline expansion, the autoregressive coefficient functions $a(y)$ may be written as $a(\mathbf{y}_{i,t}) = \sum_{j=1}^{D} c_j \phi_j(\mathbf{y}_{i,t})$.

A point to be analyzed here is the determination of the parameter D, which controls the number of knots used in the spline expansion. An objective criterion to select the optimum D may be obtained by selecting the value that minimizes the leave-one-out cross-validation residue for each linkage.

Regarding the estimation procedure, Fujita *et al.* [8] proposed the use of Generalized Least Square (GLS) estimation, i.e., by estimating the coefficients of splines expansions in $\mathbf{A}_1(\mathbf{y}_t)$ using GLS. For further details about both, estimation and statistical test, see [8].

2.3 Wavelet Dynamic Vector Autoregressive Model (DVAR)

In this extension, the coefficients in matrices $\mathbf{A}_1$ and $\mathbf{\Sigma}$, can be described as functions of time, allowing the evaluation of time-varying relationships between time series, and thus, testing Granger Causality in a dynamic fashion. Differently from the VAR model, DVAR can infer linear time-varying Granger causalities, thus increasing the power of the test when Granger causality is time-varying [6]. In other words, the DVAR technique allows the identification of one network structure for each time point, being useful to detect when a specific gene is going to be regulated in the set of time series data (Figure 1b). The DVAR model of order one is defined by $\mathbf{y}_t = \mathbf{v}(t) + \mathbf{A}_1(t)\mathbf{y}_{t-1} + \varepsilon_t$, $t = 1, \dots, T$, where $\varepsilon_t \sim N(\mathbf{0}, \mathbf{\Sigma})$ and $\mathbf{\Sigma}(t) = \begin{pmatrix} \sigma^2_{1,1}(t) & \sigma_{2,1}(t) & \dots & \sigma_{k,1}(t) \\ \sigma_{1,2}(t) & \sigma^2_{2,2}(t) & \dots & \sigma_{k,2}(t) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,n}(t) & \sigma_{2,n}(t) & \dots & \sigma^2_{n,n}(t) \end{pmatrix}$. For each time point t, $\mathbf{v}(t)$ and $\mathbf{A}_1(t)$ are the local trend vector and the coefficient matrices, respectively, given by $\mathbf{v}(t) = \begin{pmatrix} v_1(t) \\ v_2(t) \\ \vdots \\ v_n(t) \end{pmatrix}$ and $\mathbf{A}_1(t) = \begin{pmatrix} a_{1,1}(t) & a_{2,1}(t) & \dots & a_{n,1}(t) \\ a_{1,2}(t) & a_{2,2}(t) & \dots & a_{n,2}(t) \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,n}(t) & a_{2,n}(t) & \dots & a_{n,n}(t) \end{pmatrix}$

The intercept, coefficients and the covariance of random error are functions of time. This parameterization allows for the modeling of multivariate time series of regulatory networks in time-varying frameworks, providing a tool to analyze the information flow between genes along the cell cycle, for example.

Sato *et al.* [28] suggested decomposing the time-varying functions $\mathbf{v}(t)$, $\mathbf{A}_1(t)$ and $\mathbf{\Sigma}(t)$ as wavelet expansions. The basic idea is to describe the functions as linear combinations of wavelet functions $\psi_{j,k}$. Thus, considering the DVAR model, the truncated wavelet expansion for the functions $a_{l,m}(t)$ may be written as $a_{l,m}(t) = \sum_{j=1}^{J} \sum_{k=0}^{2^j-1} c_{j,k}\psi_{j,k}(t)$, where j and k are indexes for scale and time-location, respectively and $c_{j,k}$ $(j = -1, 0, 1, \ldots, T-1;\ k = 0, 1, \ldots, 2^j - 1)$ are the wavelet expansion coefficients for the function $a_{l,m}(t)$. Basically, the full estimation of the wavelet DVAR consists in estimating each of the wavelet coefficients $c_{j,k}$ for all the functions in $\mathbf{A}_1(t)$, $\mathbf{v}(t)$ and $\mathbf{\Sigma}(t)$.

Regarding the estimation procedure, Sato *et al.* [28] proposed an iterative GLS estimation composed by two stages which is described in details in [6] and [28].

2.4 Sparse Vector Autoregressive Model (SVAR)

SVAR is suitable to deal with the high dimensional characteristic of gene expression data, i.e., when the number of parameters (genes) is higher than the number of observations (microarray time series length) [7].

Notice in the order one VAR model that, when the number of genes is greater than the number of microarrays, i.e., $(T-1) < n$, where T is the time series length and n is the number of genes, the covariance matrix $\mathbf{X}'\mathbf{X}$ is not invertible and, consequently, the VAR model cannot be estimated. However, sometimes, it is desired to construct larger networks with dozens of genes, i.e., where $(T-1) < n$. In order to overcome this limitation, SVAR was presented by [7]. SVAR is based on the LASSO (Least Absolute Shrinkage and Selection Operator) estimator instead of the standard OLS. The LASSO regression is useful in the identification of regulatory networks when there are more parameters to be estimated than observations since it performs an iterative variable selection at the same time it estimates the coefficients of the regression.

2.5 Microarray Dataset

In order to demonstrate the applicability of VAR, NVAR and DVAR, a public database containing ∼45,000 probes (∼22,000 genes) and 48 time points measured in intervals of one hour [16] was used. This gene expression data was derived from a NIH3T3 cell line using the Mouse Genome 430 2.0 array platform and is available from GEO under the GSE11922 accession number.

3 Results and Discussions

VAR, NVAR and DVAR models were applied to actual microarray data in order to address a specific biological question: the interaction between proteins of the

metalloproteinase family (MMPs), key players in cancer progression, and their regulators. The extracellular matrix holds cells together and maintains the three-dimensional structure of body tissues. MMPs are involved in the breakdown of extracellular matrix in normal physiological processes, such as embryonic development and tissue remodeling, as well as in disease processes, such as cancer [18].

Tumor metastasis is a multistep process involving the dissemination of tumor cells from the primary tumor to secondary sites at a distant organ or tissue. For a tumor cell to metastasize from the primary tumor to other organs, it must locally degrade extracellular matrix components that are the physical barriers for cell migration. Thus, one of the first steps in metastasis is the degradation of the basement membrane, a process in which MMPs have been implicated.

This MMP activity is regulated by Tissue Inhibitors of Metalloproteinases (TIMPs). Inhibitors of MMPs result in inhibition of metastasis, while up-regulation of MMPs led to enhanced cancer cell invasion. In this context, two TIMP genes were chosen from the microarray data set (TIMP2 and TIMP4) and two genes possibly interacting with them were added to the network (RECK and MMP24).

TIMP2 is a member of the TIMP gene family. The proteins encoded by this gene family are natural inhibitors of matrix metalloproteinases. In addition to an inhibitory role against metalloproteinases, the encoded protein has a unique role among TIMP family members in its ability to directly suppress the proliferation of cells, being possibly critical to the maintenance of tissue homeostasis while inhibiting protease activity in tissues undergoing remodeling of the extracellular matrix [5]. TIMP4 also belongs to the TIMP gene family. Literature shows a similar mechanism of action of TIMP2 and TIMP4 when modulating cell surface activation of certain enzymes, such as progelatinase A and C [3]. TIMP4 as well as TIMP2 have been shown to possibly interact with MMP24, the metalloproteinase in our network, as depicted by STRING 8.1 (Figure 2a).

Although not a metalloproteinase, the protein encoded by RECK gene is an extracellular protein with protease inhibitor-like domains. Its expression is suppressed strongly in many tumors and recent work indicates it might have an essential role in cancer metastasis [4] [14]. RECK down-regulation by oncogenic signals may facilitate tumor invasion and metastasis, and it has been shown to be regulated by TIMP2 [26]. The interaction between both these genes and their presence in the context of MMPs regulation can be visualized in Figure 2b.

The complexity of the biological system presented above indicates that relationship among its components cannot be easily inferred. The application of the models VAR, NVAR and DVAR provided novel and complementary information to what is currently known, as shown in Figure 3. Although published data reports a mechanism through which TIMP2 regulates RECK by altering phosphorylation patterns in the cell, our models show that other relationships should be addressed.

Notice in Figure 3b that the auto-loops in RECK, TIMP2 and TIMP4, and the regulation between RECK and TIMP2 are close to linear functions (illustrated

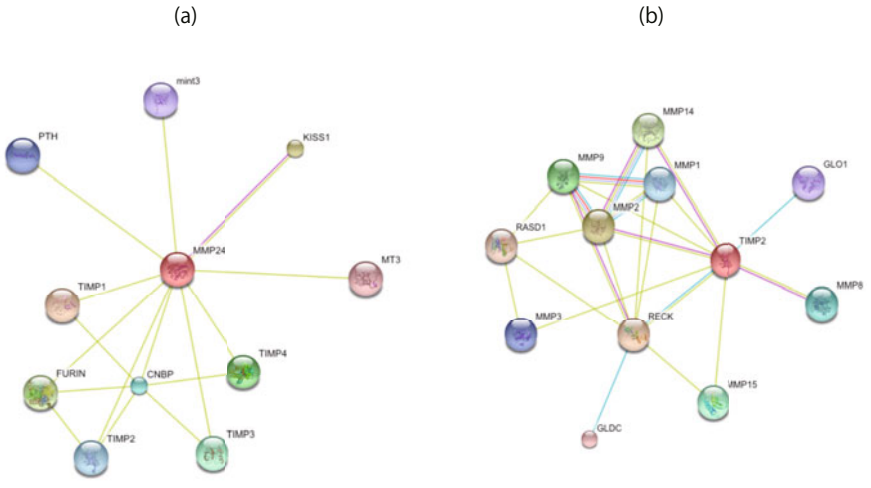


Fig. 2. Evidence of biological interaction as depicted by STRING 8.1. (a) Interaction between MMP24, TIMP4 and TIMP2 is based on proteins being co-mentioned in scientific publications (score MMP24-TIMP4: 0.77; MMP24-TIMP2: 0.65). (b) Interaction between RECK, TIMP2 and MMPs is based on proteins being co-mentioned in scientific publications as well as their association in curated databases (score RECK-TIMP2: 0.9).

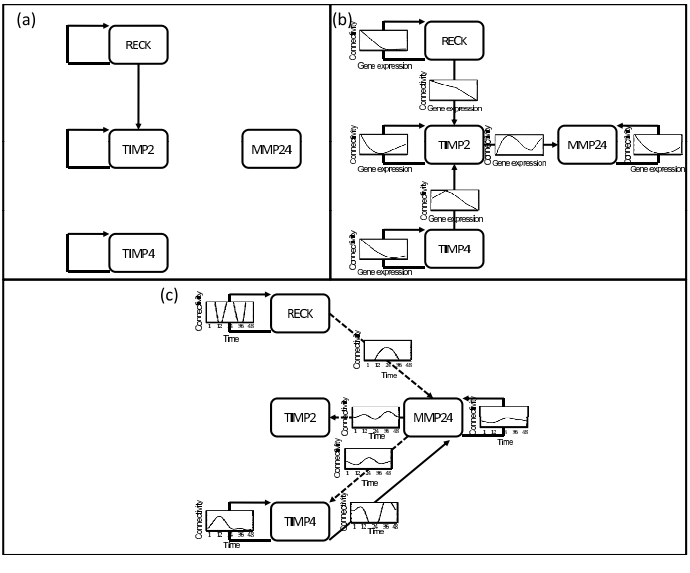


Fig. 3. Application of (a) VAR, (b) NVAR and (c) DVAR models to a network composed by the RECK, TIMP2, TIMP4 and MMP24 genes. The connectivity functions of (b) NVAR and (c) DVAR are shown in each arrow. Dashed arrows are significant connectivities with $p < 0.10$ and solid arrows with $p < 0.05$.

by the boxes shown at the edges of the network), thus, it was also possible to identify them using the VAR method (Figure 3a) which identifies linear Granger causalities. The other interconnections were not identified by VAR since they are nonlinear relationships. In Figure 3c, one can verify the time-varying Granger causalities that were not identified by the other two methods. These interactions occur only during a narrow time frame as can be seen in Figure 3c, in other words, the regulation was not constant in time to be identified by standard VAR (the regulations occur at the peaks illustrated at the edges of the network in Figure 3c). Since each VAR extension was designed for a specific purpose, in order to address a real biological question to which no hypothesis has been postulated, one should apply all of them and verify which regulation could be biologically more reasonable and test the new hypothesis using wet lab experiments.

One limitation of VAR, NVAR and DVAR is the fact that they can be applied only in a limited number of genes, since the estimation of the parameters requires a large time series length. In order to identify linear Granger causalities in a large gene set, one may apply the SVAR model.

By applying SVAR in a large (more parameters to be estimated than observations) and sparse scale-free network [17] containing 100 genes (nodes) and 100 regulations (edges), with time series length equal to $T = 25$ and random noise $\varepsilon_t \sim N(0, 1)$, it was possible to reconstruct in average, 17, 18 and 19% of the entire topology by controlling the false discovery rate (FDR) in 1, 5 and 10%, respectively. It is important to point out here that this partial reconstruction was only possible because the topology is sparse (around n edges out of a total of n^2 possible edges, where n is the number of genes), i.e., scale-free [17] and the LASSO based shrinkage strategy automatically selects the predictors. A biological application of SVAR with dozens of genes can be observed in [7], where they modeled a cancer-related pathway. In a practical application with a low number of genes, the results obtained by SVAR is similar to VAR, since the LASSO under a large number of time points converges to the OLS.

Arnold *et al.* [1] have shown by simulations and also by applying in actual biological data that the LASSO algorithm coupled to VAR model exhibits consistent gain over the canonical pairwise graphical Granger method.

Several other works have been reported based on the shrinkage strategy in order to model high dimensional data. For example, Opgen-Rhein and Strimmer [27] used a James-Stein-type shrinkage while Shimamura *et al.* [30] proposed the use of recursive elastic net instead of using LASSO in the SVAR model. Basically, independent of the shrinkage type used in the estimator, the ideas are similar. At each iteration, the regression coefficients of each gene with all others are weighted according to their current size and several coefficients are successively down-weighted and set to zero.

The shrinkage strategy used in SVAR is a potential solution to model networks when there is a lot of parameters in limited time series length. For example, Kojima *et al.* [19] used group LASSO instead of GLS in the NVAR model, while Marinazzo *et al.* [21] implemented a kernel-based NVAR instead of applying splines expansion. However, the cost to be paid to combine different VAR

models in high dimensional data is the fact that statistical tests become difficult, resulting in a high number of false positive edges.

Thus, the differences concerning VAR, NVAR, DVAR and SVAR can be briefly summarized as: (i) VAR, DVAR and SVAR identify linear Granger causality; (ii) NVAR identifies nonlinear gene expression dependent Granger causality; (iii) DVAR identifies time-varying Granger causality and (iv) SVAR is useful to identify Granger causality in high dimensional data.

4 Final Considerations

The relevance of Granger causality applications in Bioinformatics can be demonstrated by the increasing number of reports published in the last few years. Nagarajan and Upreti [24] and Nagarajan [25] investigated the use of bivariate VAR for acyclic approximations of networks composed of two genes by exploring the parameters defined as transcriptional noise variance, autoregulatory feedback, and transcriptional coupling strength. Krishna *et al.* [20] described a clustering method using Granger causality in order to identify functional modules from temporal gene expression data.

Besides VAR models, other methods that are not model-based (regression models) such as the ones based on partial correlations are adequate for the identification of Granger causality in the presence of unobserved (latent) variables [13] or in order to identify Granger causality between sets of genes [10].

Several studies have addressed other aspects of Granger causality in Bioinformatics. For example, since it is difficult or quite impossible to include all the variables in the model, Guo *et al.* [12] have developed a method to eliminate the influence of latent variables in both time and frequency domains. Zou *et al.* [31] have recently published a report comparing VAR and dynamic Bayesian networks. In a systematic and computationally intensive comparison on both artificial and actual biological data, Zou *et al.* [31] concluded that the critical point is the time series length, i.e., the dynamic Bayesian network inference outperforms VAR when time series length is short, otherwise, VAR is better. Fujita *et al.* [9] developed a VAR model which takes into account microarray measurement error, thus, obtaining more accurate p-values and coefficients. Existing methods to identify Granger causality are based on Wald type test which relies on the homoscedasticity normality assumption of the data distribution. In order to overcome this drawback, Hu *et al.* [15] proposed an estimating equation-based method which is robust to both heteroscedasticity and non-normality of the gene expression data.

Despite clear improvement in the identification of Granger causality in gene regulatory networks in the past few years, current limitations include: (i) the necessity for a model which is able to incorporate biological information (protein-protein interaction, for example) in order to improve the accuracy of the inferred network; (ii) the integration of measurement error in the methods already developed; (iii) a better understanding of the meaning of the information flow from a biological point of view, i.e., what Granger causality means in the regulatory

network. More accurate models should provide additional insights on cellular process and ultimately lead to a better comprehension of disease mechanisms or to the identification of potential drug targets for better treatment.

Acknowledgments. The super-computing resource was provided by Human Genome Center (Univ. of Tokyo). This work was supported by RIKEN - Japan and Albert Einstein Research and Education Institute - Brazil.

References

1. Arnold, A., Liu, Y., Abe, N.: Temporal causal modeling with graphical Granger methods. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, San Jose, pp. 66–75 (2007)
2. Baccala, L., Sameshima, K.: Partial directed coherence: a new concept in neural structure determination. Biological Cybernetics 84, 463–474 (2001)
3. Bigg, H.F., Shi, Y.E., Liu, Y.E., Steffensen, B., Overall, C.M.: Specific, high affinity binding of tissue inhibitor of metalloproteinases-4 (TIMP-4) to the COOH-terminal hemopexin-like domain of human gelatinase A. TIMP-4 binds progelatinase A and the COOH-terminal domain in a similar manner to TIMP-2. J. Biol. Chem. 272, 15496–15500 (1997)
4. Chang, H.C., Cho, C.Y., Hung, W.C.: Silencing of the metastasis suppressor RECK by RAS oncogene is mediated by DNA methyltransferase 3b-induced promoter methylation. Cancer Res. 66, 8413–8420 (2006)
5. Chang, H., Lee, J., Poo, H., Noda, M., Diaz, T., Wei, B., Stetler-Stevenson, W.G., Oh, J.: TIMP-2 promotes cell spreading and adhesion via upregulation of Rap1 signaling. Biochem. Biophys. Res. Commun. 7, 1201–1206 (2006)
6. Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Morettin, P.A., Sogayar, M.C., Ferreira, C.E.: Time-varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method. Bioinformatics 23, 1623–1630 (2007a)
7. Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Yamaguchi, R., Miyano, S., Sogayar, M.C., Ferreira, C.E.: Modeling gene expression regulatory networks with the sparse vector autoregressive model. BMC Systems Biology 1, 39 (2007b)
8. Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Sogayar, M.C., Ferreira, C.E., Miyano, S.: Modeling nonlinear gene regulatory networks from time series gene expression data. Journal of Bioinformatics and Computational Biology 6, 961–979 (2008)
9. Fujita, A., Patriota, A.G., Sato, J.R., Miyano, S.: The impact of measurement errors in the identification of regulatory networks. BMC Bioinformatics 10, 412 (2009)
10. Fujita, A., Sato, J.R., Kojima, K., Gomes, L.R., Nagasaki, M., Sogayar, M.C., Miyano, S.: Identification of Granger causality between gene sets. Journal of Bioinformatics and Computational Biology (in press)
11. Granger, C.W.J.: Investigating causal relationships by econometric models and cross-spectral methods. Econometrica 37, 424–438 (1969)
12. Guo, S., Wu, J., Ding, M., Feng, J.: Uncovering interactions in the frequency domain. PLoS Computational Biology 4, e1000087 (2008a)

13. Guo, S., Seth, A.K., Kendrick, K.M., Zhou, C., Feng, J.: Partial Granger causality - Eliminating exogenous inputs and latent variables. Journal of Neuroscience Methods 172, 79–83 (2008b)
14. Hsu, M.C., Chang, H.C., Hung, W.C.: HER-2/neu represses the metastasis suppressor RECK via ERK and Sp transcription factors to promote cell invasion. J. Biol. Chem. 281, 4718–4725 (2006)
15. Hu, J.: Estimating equation-based causality analysis with application to microarray time series data. Biostatistics 10, 468–480 (2009)
16. Hughes, M.E., DiTacchio, L., Hayes, K.R., Vollmers, C., Pulivarthy, S., Baggs, J.E., Panda, S., Hogenesch, J.B.: Harmonics of circadian gene transcription in mammals. PLoS Genetics 5, e1000442 (2009)
17. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabási, A.L.: The large-scale organization of metabolic networks. Nature 407, 651–654 (2000)
18. Johansson, N., Ahonen, M., Kähäri, V.-M.: Matrix metalloproteinases in tumor invasion. Cell. Mol. Life Sci. 57, 5–15 (2000)
19. Kojima, K., Fujita, A., Shimamura, T., Imoto, S., Miyano, S.: Estimation of nonlinear gene regulatory networks via L1 regularized NVAR from time series gene expression data. Genome Informatics 21, 37–51 (2008)
20. Krishna, R., Li, C.-T., Buchanan-Wollaston, V.: Interaction based functional clustering of genomic data. In: Ninth IEEE International Conference on Bioinformatics and Bioengineering, Washington, pp. 130–137 (2009)
21. Marinazzo, D., Pellicoro, M., Stramaglia, S.: Kernel-Granger causality and the analysis of dynamical networks. Physical Review E 7, 056215 (2008)
22. McCracken, M.W.: Asymptotics for out of sample tests of Granger causality. Journal of Econometrics 140, 719–752 (2007)
23. Mukhopadhyay, N.D., Chatterjee, S.: Causality and pathway search in microarray time series experiment. Bioinformatics 23, 442–449 (2007)
24. Nagarajan, R., Upreti, M.: Comment on causality and pathway search in microarray time series experiment. Bioinformatics 24, 1029–1032 (2008)
25. Nagarajan, R.: A note on inferring acyclic network structures using Granger causality tests. The International Journal of Biostatistics 5, article 10 (2009)
26. Oh, J., Diaz, T., Wei, B., Chang, H., Noda, M., Stetler-Stevenson, W.G.: TIMP-2 upregulates RECK expression via dephosphorylation of paxillin tyrosine residues 31 and 118. Oncogene 25, 4230–4234 (2006)
27. Opgen-Rhein, R., Strimmer, K.: Learning causal networks from systems biology time course data: effective model selection procedure for the vector autoregressive process. BMC Bioinformatics 8, S3 (2007)
28. Sato, J.R., Amaro Jr., E., Takahashi, D.Y., de Maria Felix, M., Brammer, M.J., Morettin, P.A.: A method to produce evolving functional connectivity maps during the course of an fMRI experiment using wavelet-based time-varying Granger causality. Neuroimage 31, 187–196 (2006)
29. Sato, J.R., Morettin, P.A., Arantes, P.R., Amaro Jr., E.: Wavelet based time-varying vector autoregressive modeling. Computational Statistics & Data Analysis 51, 5847–5866 (2007)
30. Shimamura, T., Imoto, S., Yamaguchi, R., Fujita, A., Nagasaki, M., Miyano, S.: Recursive regularization for inferring gene networks from time-course gene expression profiles. BMC Systems Biology 3, 41 (2009)
31. Zou, C., Feng, J.: Granger causality vs. dynamic Bayesian network inference: a comparative study. BMC Bioinformatics 10, 122 (2009)

Semi-supervised Approach for Finding Cancer Sub-classes on Gene Expression Data

Clerton Ribeiro, Francisco de Assis T. de Carvalho, and Ivan G. Costa

Center of Informatics, Federal University of Pernambuco, Recife, Brazil
{craf,fatc,igcf}@cin.ufpe.br

Abstract. The analysis of cancer gene expression is intrinsically a semi-supervised problem, as one is interested in building a classifier for diagnosis, but also on finding new sub-classes of cancer. We propose here a method for Mixture Discriminant Analysis (MDA), which can simultaneously detect sub-classes of cancer and perform classification. We evaluate the method on 10 gene expression data sets. MDA not only improved the classification in some of these data sets, as it detected some known and putative sub-classes of cancer.

Keywords: cancer gene expression, mixture discriminant analysis, semi-supervised learning, constraint based mixture estimation.

1 Introduction

The measurement of the expression of all genes of cancer patients has made possible the development of personalized diagnostics [21]. In this context, a standard approach is the use of machine learning methods to build a classifier for a data set with several healthy and cancer patients or with distinct types of cancer [19]. Moreover, analysis on such data sets have shown the presence of unknown sub-types of cancer by the application of clustering methods [1,10]. Such findings have made the study of gene expression of cancer to be extremely popular, and lead to great advances in cancer diagnosis [21].

These facts indicate that cancer based diagnosis is intrinsically a semi-supervised problem [5]. While the studies generating the gene expression data sets give class labelling of all samples in the data, the frequent discovery of new sub-classes has made the application of both supervised and unsupervised methods routine. Therefore, a method that performs classification of cancer types simultaneously to finding new sub-classes is extremely desirable. By using the detected sub-classes in the classification task, the method can better delineate class boundaries/data distribution, therefore enhancing the overall classification accuracy [11]. Moreover, the detected sub-classes, whenever they are present in the data, are interesting candidates for further analysis by the biomedical experts.

We propose here a semi-supervised method for estimating Mixture Discriminant Analysis (MDA) with Gaussians distributions. MDA, which has been initially proposed in [11], works by fitting a mixture of Gaussian distributions to each class in the data set. One major drawback of this approach is the fact that one needs to estimate the optimal number of components in the mixtures (or sub-classes) for each class independently. This makes the method computationally costly and requires the application

C.E. Ferreira, S. Miyano, and P.F. Stadler (Eds.): BSB 2010, LNBI 6268, pp. 25–34, 2010.

of model selection procedures. We propose here the use of a constraint-based-mixture estimation [13] for estimating the MDA. The method has as input the list all negative pairwise constraints, i.e. all pairs of patients that should not be in the same class. The algorithm, which is based on an extension of the Expectation-Maximization (EM) algorithm, searches for solutions with a pre-determined number of groups K satisfying all negative constraints. That is, we do not have patients of distinct classes in a single group, but we allow patients from the same class to belong to several groups. Therefore, if K is higher than the number C of classes (cancer types), the method will return a classifier with $K - C$ novel sub-classes.

A similar approach has been previously shown to work on the classification of time-series of Multiple Sclerosis patients [6]. In this work, we evaluate the MDA method with several data sets from a cancer gene expression compendium [7]. Furthermore, we apply a Quadratic Discriminant Analysis (QDA), which is equivalent to MDA when $K = C$, to serve as a baseline case. To select the optimal number of sub-classes $K - C$, we use a cross-validation procedure. Finally, apply a consensus method proposed in [16] to evaluate if the sub-classes found are stable over distinct solutions obtained by the cross-validation procedure.

2 Material and Methods

2.1 Data Sets

We use in this study 10 public micro-array data sets with cancer gene expression (`http://algorithmics.molgen.mpg.de/Supplements/CompCancer`).
An overview of these 10 datasets is presented in Table 1.

Table 1. Data set description

Dataset	Classes	n	C	d
Alizadeh-v2	DLBCL(42), FL(9), CLL(11)	62	3	4022
Alizadeh-v3	DLBCL1(21), DLBCL2(21), FL(9), CLL(11)	62	4	4022
Armstrong-v1	ALL(24),MLL(48)	72	2	12582
Armstrong-v2	ALL(24), MLL(20), AML(28)	72	3	12582
Chen	HCC(104), liver(75)	179	2	22699
Golub-v1	ALL(47), AML(25)	72	2	7129
Golub-v2	ALL-B(38), ALL-T(9), AML(25)	72	3	7129
Nutt-v2	CG(14), NG(14)	28	2	12625
Nutt-v3	CO(7), NO(15)	22	2	12625
Yeoh-v1	T-ALL(43), B-ALL(205)	248	2	12625

In Table 1, the second column describes the names of the classes (cancer types), as defined in the original publication, and the number of samples (patients) in each class. For further description of classes see [7]. The third column presents the number of samples (n), the fourth column the number of classes and the last column the number of genes (d). It is quite noticeable from the table that all data sets are sparse with a few samples on a high dimensional space.

The data were pre-processed by the application of an unsupervised filter to discard missing values and genes displaying no differential expression, as described in [7]. The pre-processing performed on data from experiments based on the Affymetrix platform (Alizadeh, Golub, Nutt and Yeoh) has the following steps: (1) all values below 10 and above 16000 were replaced by these bounds,(2) we measured the mean expression of each gene and eliminate 10% of the highest and lowest values to avoid extreme values (3) each expression value was replaced by the base 2 log transformation of the ratio between the expression value and the gene mean expression. For cDNA platform data (Armstrong and Chen), it was not necessary to apply transformations, as they were already in logarithmic scale. The unsupervised filter process was as follows: two l and c thresholds were chosen, where the absolute value of the feature has to be higher than l in at least c patients. Genes that do not fit this restriction were excluded from the data set.

2.2 Classification Algorithms

Let X be a d by n matrix representing a gene expression data set, where x_{ij} denotes the expression value of sample (patient) j and feature (gene) i, x_i is a d-dimensional vector with the expression values of sample (patient) i. We also have associated to each data set a vector Y with dimension n, where $y_i \in \{1, ..., C\}$ denotes the class sample i belongs to.

2.3 Discriminant Analysis

Discriminant analysis (DA) methods perform classification by inference over the posterior distribution $\mathbf{P}[y|x]$ [12]. Let $\mathbf{P}[x_i|y_i = c]$ be the class-conditional density modeling the distribution of samples in class c and π_c be the prior distribution of class c, such that $\sum_{c=1}^{C} \pi_c = 1$ and $\pi_c \geq 0$, we can use Bayes Theorem to derive the posterior probability

$$\mathbf{P}[y_i = c|x_i] = \frac{\pi_c \mathbf{P}[x_i|y_i = c]}{\sum_{c'=1}^{C} \pi_{c'} \mathbf{P}[x_i|y_i = c']}. \tag{1}$$

Therefore, classification of a sample x_i can be performed with the rule

$$\hat{y}_i = \underset{c=\{1,...,C\}}{\arg\max}\ \mathbf{P}[y_i = c|x_i]. \tag{2}$$

as given in Eq. 2, where $\hat{y}_i$ is the predicted class for sample i.

The definition of $\mathbf{P}[x_i|y_i = c]$ is application dependent. In gene expression analysis, a usual choice is a multivariate Gaussian density function [9], which is defined as

$$\mathbf{P}[x_i|y_i = c, \theta_c] = \frac{1}{\sqrt{(2\pi)^d |\Sigma_c|}} \exp^{\frac{1}{2}(x_i - \mu_c)^{\mathbf{T}} \Sigma_c^{-1} (x_i - \mu_c)}, \tag{3}$$

where θ_c are the parameters (μ_c, Σ_c). μ_c and Σ_c can be estimated with the mean and covariance matrices of samples of class c and $\pi_c = n_c/n$, where n_c is the number of samples in class c [12].

Given sparsity of the data (few samples and high dimension), it is usual to assume independence among the attributes given the class. In gene expression analysis this is done by estimating a diagonal parameterization of the covariance matrix Σ_c, i.e. only the diagonal entries are estimated and all other values are set to zero [9]. This variant of DA is known as Diagonal Quadratic Discriminant Analysis (DQDA) and will be used in this study as a baseline method.

2.4 Mixture Discriminant Analysis

With mixture of discriminant analysis (MDA), we assume that class condition densities can be defined as a mixture model, that is

$$\mathbf{P}[x_i|y_i = c] = \sum_{k=1}^{K} \alpha_k \mathbf{P}[x_i|z_i = k], \tag{4}$$

where $\alpha_k, i = 1, ..., K$ are the mixing coefficients. In [11], the estimation of these mixture were performed with the application of the EM algorithm for each class to be classified.

2.5 Mixture Model Estimation with Constraints

A standard mixture model can be defined as

$$\mathbf{P}[x_i|\Theta] = \sum_{k=1}^{K} \alpha_k \mathbf{P}[x_i|y_i = k, \theta_k] \tag{5}$$

as given in Eq. 5,where $\Theta = (\alpha_1, ..., \alpha_k, \theta_1, ..., \theta_K)$ are the model parameters and α_k are the mixing coefficients. By including a set of hidden labels represented by the n-dimensional vector Z, where $z_i \in \{1, .., K\}$ defines the component generating the x_i, we obtain the complete data likelihood

$$\mathbf{P}[X, Y|\Theta] = \mathbf{P}[X|Z, \Theta]\mathbf{P}[Z|\Theta]. \tag{6}$$

We can use then the EM method to estimate the parameters Θ and component assignments Z maximizing the complete likelihood (see [15] for details).

In constrained-based-mixture estimation (and its similar constrained based clustering), the user can define a $n \times n$ matrix W with negative pairwise constraints, where $w_{ij}^{-} = 1$ if samples i and j should not belong to the same mixture component and $w_{ij}^{-} = 0$ otherwise. The constraints are incorporated in the estimation by extending the prior probability of the hidden variable to $\mathbf{P}[Z|\Theta, W] = \mathbf{P}[Z|\Theta]\mathbf{P}[W|Z]$. Assuming $\mathbf{P}[W|Z]$ follows a Gibbs distribution, there is a variation of the EM algorithm for estimating Z and Θ [13,14]. The method requires the redefinition of the posterior assignment distribution as

$$\mathbf{P}[z_i = k|x_i, W] = \frac{\pi_c \mathbf{P}[x_i|z_i = k]}{\mathcal{Z}} \exp^{\sum_{j \neq i} -\lambda^{-} w_{ij}^{-} \mathbf{P}[z_j = k|x_j, W]}, \tag{7}$$

where $\mathcal{Z} = \sum_{k=1}^{K} \mathbf{P}[z_i = k|x_i, W]$ and λ^{-} is the Lagrange parameter defining the penalty weight of constraints violations.

2.6 Constraint-Based Mixture Discriminant Analysis

We propose here the use of the constraint-based mixture estimation method described above for obtaining a MDA classifier. By setting the penalty parameter λ^- with a high value and the constraint matrix W, such that $w_{ij}^- = 1$ if $y_i \neq y_j$ and $w_{ij}^- = 0$ otherwise, we will obtain solutions where samples with distinct classes are not in the same mixture component. Furthermore, by choosing a number of components $K > C$, some of the classes will be related to more than one mixture component. In other words, the mixture will divide some of the classes in sub-classes.

Therefore, we need a procedure to relate the mixture components with the classes. This can be achieved by relating the assignment vector Z of the mixture with the class vector Y. We can estimate the probability of obtaining class c given component k by

$$\mathbf{P}[y = c|z = k] = \frac{\sum_{i=1}^{N} \mathbf{1}(y_i = c)\mathbf{1}(z_i = k)}{\sum_{i=1}^{N} \mathbf{1}(z_i = k)}, \tag{8}$$

where $\mathbf{1}$ is the identity function. From this, we can define the mapping

$$\text{ClassOf}(k) = \underset{c=\{1,\ldots,C\}}{\arg\max}\ \mathbf{P}[y = c|z = k], \tag{9}$$

which defines the class c related to component k.

We can use this mapping and parameters Θ, which has been estimated with the method described in Section 2.5, to define the class conditionals as defined in Eq. 4 and obtain a MDA classifier with the use of Eq. 1.

2.7 Experimental Design and Consensus Analysis

For each data set, we performed a leave-one-out cross-validation. All accuracies described in the following are based on the test set alone. Then we use the Friedman test followed by a multiple comparison correction procedure to access the significance of the ranking of the methods [8]. For the final interpretation of the sub-classes, we need a method for combining the results of the classifiers (training and test sets) for all leave one out runs. For this task, we use a procedure proposed in [4,16]. First, we build a co-occurrence matrix by counting for each pair of samples the number of times they appear in the same component across the different solutions Z. The consensus method works by reshuffling the matrix and clustering samples that share similar groups over solutions [16].

3 Experiments and Results

We investigate here if the use of the Mixture Discriminant Analysis method improves classification accuracy in relation to the baseline method DQDA, which is the equivalent to MDA when $K = C$. Data sets, where the MDA improves or sustains the classification accuracy, are of interest, as these indicate the presence of sub-classes of cancer.

Table 2. Accuracy and standard deviation from classification methods for each data set

Dataset	DQDA	MDA $c+1$	MDA $c+2$
Alizadeh-v1	**95.24** (21.55)	80.95 (39.74)	80.95 (26.07)
Alizadeh-v2	96.77 (17.81)	**100** (0)	**100** (0)
Armstrong-v1	98.61 (11.79)	97.22 (16.55)	98.61 (11.79)
Armstrong-v2	**94.44 (23.07)**	**94.44 (23.07)**	88.89 (11.79)
Chen	91.62 (27.79)	91.06 (28.61)	94.41 (20.72)
Golub-v1	**98.61 (11.79)**	97.22 (16.55)	93.05 (16.55)
Golub-v2	90.28 (29.83)	90.27 (29.83)	90.27 (20.12)
Nutt-v2	78.57 (41.79)	71.42 (46.00)	82.14 (31.50)
Nutt-v3	86.36 (35.13)	90.9 (29.42)	81.81 (38.56)
Yeoh-v1	**96.16 (21.50)**	92.74 (26.00)	91.93 (16.60)

We depict the accuracies and standard deviation in Table 1. Values in bold face represent the method, which obtained a statistically significant improvement as indicated by the Friedman test [8]. For three datasets (Alizadeh-v1, Golub-v1 and Yeoh-v1), DQDA obtained best results. In Alizadeth-v2 MDA with c+1 and c+2 obtained better results and in Armstrong-v2 both DQDA and MDA c+1 were best. In all other cases, there was no statistically relevant difference. Note that we used a leave-one-out cross-validation, due of the small number of samples in the data sets. Such setting, usually lead to low accuracy bias but high deviation, lowering the statistical power of comparisons [3].

As expected, MDA did not obtained a higher accuracy than DQDA in all data sets, not all data sets contain sub-classes. Moreover, the limited number of patients may lead to over-fitting with solutions with many sub-classes (too complex models). In some scenarios MDA was better or equivalent to DQDA. As the existence of sub-classes is interesting from the application problem, we prefer the solution of MDA with more components, whenever accuracy is equivalent to DQDA.

Some of the data sets above, Alizadeh-v1, Armstrong-v2 and Gollub-v2, represent the original classification performed by the specialists, which were latter found to contain sub-classes with the use of unsupervised methods [1,2,10]. In these scenarios, MDA had superior or equivalent accuracies in relation the QDA.

To assess if MDA is successful in detecting the sub-classes, we perform the consensus analysis [4,16] on the Armstrong-v2 data set. In Figure 1, we depict the co-occurrence matrix, where a particular entry indicates the number of times the pair of patients were classified in the same class/sub-class (darker values indicate higher counts). Ideally, the consensus matrix should a block of dark values for each class indicating that the same patients were consistently classified together. As seen in Figure 1 top, DQDA obtained an almost perfect classification and separated all but one patient from the original classes: lymphoblastic leukemias with MLL translocations (MLL) and Acute lymphoblastic leukemias (ALL) [2]. This is indicated in the figure by the two block of dark values.

The original study applied a clustering algorithm and found that 28 patients, which were originally classified as patients with MLL, had distinct expression signatures

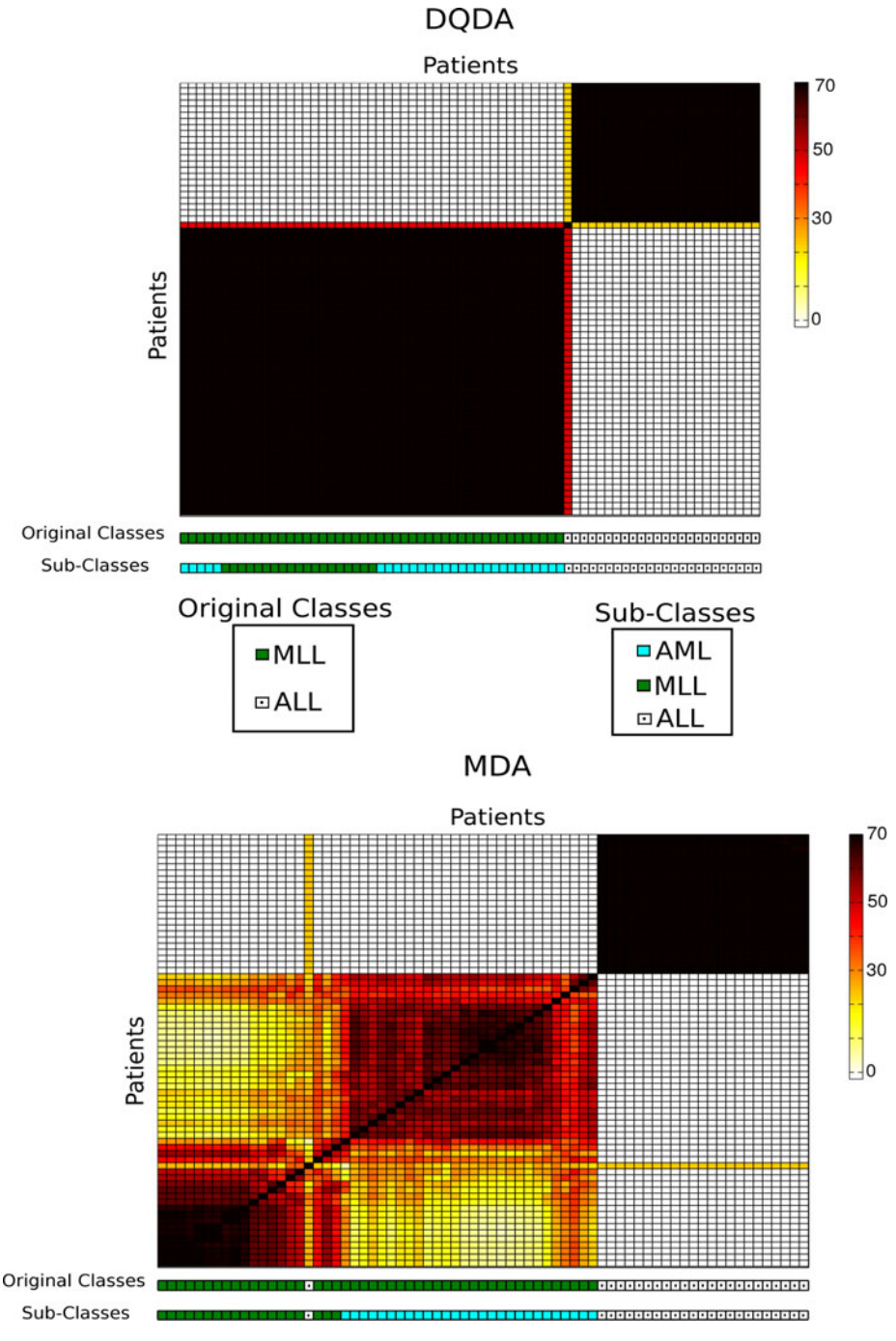


Fig. 1. Consensus Analysis on the Armstrong-v2 data for DQDA (top) and MDA $C + 1$ (bottom)

from other MLL patients [2]. These had their diagnostics changed to akute myelogenous leukemias (AML). As indicated in Figure 1 bottom, MDA with $c + 1$ components, detected the subclasses AML and MLL as indicated by the two blocks of dark values in the left-bottom part of the matrix. Note that in this data set, only the two original classes (MLL and ALL) were given as input for the constraints. This exemplifies a case when MDA successfully finds sub-classes.

Another interesting data set is Nutt-v3, where we see a improvement on the classification accuracy of MDA in relation to DQDA. Moreover, the co-occurrence analysis indicated two sub-classes of patients with non-classic anaplastic oligodendrogliomas, with respectively 11 and 4 patients. These non-classic gliomas are of difficult diagnosis and these sub-classes have not been previously reported in the original study [17]. We detected a significant difference (t-test with p-value < 0.05) in the patient survival time: 672 days for sub-class 1 and 1079 days for sub-class 2.

Next, we explored the genes (features) that are discriminative between these sub-classes by estimating the Fisher discriminant ratio for all genes and ranking them. We selected the 50 most discriminant genes for each class and we performed an enrichment analysis with the g:profiler tool [18]. The analysis revealed that genes up-regulated in sub-class 1 are related to metabolic process and cell cycle, while genes over-expressed in sub-class 2 are related to immune response. These indicate a quite distinct expression signature of these sub-classes, possibly as a result of distinct immune response of the patients to cancer. However, further patient and clinical data are required for the validation of the potential sub-classes.

4 Final Remarks

We propose a new method for estimation of mixture discriminant analysis. This methods improves the original proposal of MDA [11] by requiring only one pass of the EM algorithm to obtain solutions. In the analysis of cancer gene expression, we have shown that MDA can improve classification and successfully indicate the existence of sub-classes of cancer of gene expression data sets. This was exemplified on the classical study from Armstrong et al. [2]. Moreover, interesting sub-classes of non-classical gliomas were found in the Nutt data set.

As future work, we would like to either include new data sets in the study and perform a more detailed biological analysis of the sub-classes found. From a methodological point of view, the MDA can be improved by the use of feature selection methods to cope with the high-dimensionality problem, for example using an approach similar to Shrunken centroids [20].

Acknowledgment

This work has been partially supported by Brazilian research agencies: FACEPE, CNPq and CAPES.

References

1. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., Staudt, L.M.: Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature 403(6769), 503–511 (2000)
2. Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., Korsmeyer, S.J.: Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat. Genet. 30(1), 41–47 (2002)
3. Braga-Neto, U.M., Dougherty, E.R.: Is cross-validation valid for small-sample microarray classification? Bioinformatics 20(3), 374–380 (2004)
4. Brunet, J.-P., Tamayo, P., Golub, T.R., Mesirov, J.P.: Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. USA 101(12), 4164–4169 (2004)
5. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning. MIT Press, Cambridge (2006)
6. Costa, I.G., Schonhuth, A., Hafemeister, C., Schliep, A.: Constrained mixture estimation for analysis and robust classification of clinical time series. Bioinformatics 25(12), 6–14 (2009)
7. de Souto, M.C.P., Costa, I.G., de Araujo, D.S.A., Ludermir, T.B., Schliep, A.: Clustering cancer gene expression data: a comparative study. BMC Bioinformatics 9, 497 (2008)
8. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30 (2006)
9. Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 97(457), 77–87 (2002)
10. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439), 531–537 (1999)
11. Hastie, T., Tibshirani, R.: Discriminant analysis by gaussian mixtures. Journal of the Royal Statistical Society, Series B 58, 155–176 (1996)
12. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: Data mining, inference and prediction. Springer, Heidelberg (2001)
13. Lange, T., Law, M.H., Jain, A.K., Buhmann, J.M.: Learning with constrained and unlabelled data. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 731–738 (2005)
14. Lu, Z., Leen, T.: Semi-supervised learning with penalized probabilistic clustering. In: Saul, L.K., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems, vol. 17, pp. 849–856. MIT Press, Cambridge (2005)
15. MacLachlan, G., Peel, D.: Finite Mixture Models. Wiley Series in Probability and Statistics. Wiley, Chichester (2000)
16. Monti, S., Tamayo, P., Mesirov, J.P., Golub, T.R.: Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning 52(1-2), 91–118 (2003)
17. Nutt, C.L., Mani, D.R., Betensky, R.A., Tamayo, P., Cairncross, J.G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M.E., Batchelor, T.T., Black, P.M., von Deimling, A., Pomeroy, S.L., Golub, T.R., Louis, D.N.: Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res. 63(7), 1602–1607 (2003)

18. Reimand, J., Kull, M., Peterson, H., Hansen, J., Vilo, J.: g:profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. 35(Web Server issue), W193–W200 (2007)
19. Spang, R.: Diagnostic signatures from microarrays: a bioinformatics concept for personalized medicine. BIOSILICO 1(2), 64–68 (2003)
20. Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G.: Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS 99(10), 6567–6572 (2002)
21. van't Veer, L.J., Bernards, R.: Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature 452(7187), 564–570 (2008)

Bounds on the Transposition Distance for Lonely Permutations

Luis Antonio B. Kowada[1],
Rodrigo de A. Hausen[2], and Celina M.H. de Figueiredo[3]

[1] Universidade Federal Fluminense
luis@vm.uff.br
[2] Universidade de São Paulo
hausen@cos.ufrj.br
[3] Universidade Federal do Rio de Janeiro
celina@cos.ufrj.br

Abstract. The problem of determining the transposition distance of permutations is a notoriously challenging one; to this date, neither there exists a polynomial algorithm for solving it, nor a proof that it is NP-hard. Moreover, there are no tight bounds on the transposition distance of permutations in general. Our proposed approach merges two successful strategies: the classical reality and desire diagram proposed by Bafna and Pevzner and the more recent toric equivalence relation proposed by Eriksson et al. We focus on unitary toric equivalence classes and the corresponding lonely permutations. In a previous paper, we considered the case $n+1$ prime, proved that the reality and desire diagram of such lonely permutations has just one odd cycle and succeed in identifying in this subset of lonely permutations, new permutations for which the transposition distance is computed. The present paper extends regularity properties of the cycle structure for general n, yielding tight bounds for the transposition distance of lonely permutations. The subset of lonely permutations that are 3-permutations is characterized and consequently an upper bound is obtained for their transposition distances.

1 Introduction

The transposition distance [1] of permutations is a metric for determining the similarity between chromosomes. This approach compares the gene orders of the chromosomes instead of directly comparing their DNA sequences [2]. A transposition is a rearrangement of the gene order in a chromosome, in which a block of genes is "cut" from a chromosome and "pasted" elsewhere in the same chromosome. A possible biological explanation for this rearrangement is the duplication of a block of genes, followed by the deletion of the original block [3].

The problem of determining the transposition distance of permutations is a notoriously challenging one; to this date, neither there exists a polynomial algorithm for solving it, nor a proof that it is NP-hard. Moreover, there are no tight bounds for the transposition distance of permutations in general. Our proposed

C.E. Ferreira, S. Miyano, and P.F. Stadler (Eds.): BSB 2010, LNBI 6268, pp. 35–46, 2010.

approach merges two successful strategies, which will be presented in Section 1.1: the classical reality and desire diagram proposed by Bafna and Pevzner and the more recent toric equivalence relation proposed by Eriksson et al. We focus on unitary toric equivalence classes and the corresponding lonely permutations. Following the approach of previous works that sought better bounds for the transposition distance when restricted to subsets of permutations [4,5], we focus on the subset of lonely permutations.

This article is organized as follows: The subsequent subsection provides the basic background on transposition distance, the toric classes, the reality and desire diagram, current known bounds and studied subsets of permutations. Section 2 characterizes the cycle structure of the reality and desire diagram of lonely permutations; this characterization is used to identify which lonely permutations have cycles of even length – which have, thus, a known lower bound of $\frac{n+1}{2}$ for their distance. In the same section, we also provide an upper bound that relates lonely permutations of n elements with lonely permutations having a fewer number of elements. Section 3 is devoted to lonely permutations that are also 3-permutations, culminating in an upper bound for their distance. Section 4 contains our concluding remarks.

1.1 Background

For our purposes, the gene order in a chromosome that has n genes is defined as a permutation $\pi = [\pi_1 \pi_2 \dots \pi_n]$, where $\pi_i \in \{1, 2, \dots, n\}$ for $i = 1 \dots n$, and $\pi_i \neq \pi_j$ if, and only if, $i \neq j$. A transposition is defined as follows:

Definition 1. *[1] A* transposition*, denoted by $t(i, j, k)$, where $1 \leq i < j < k \leq n + 1$, is defined as the permutation*

$$t(i, j, k) := [1\ 2 \dots i{-}1\ j\ j{+}1 \dots k{-}1\ i\ i{+}1 \dots j{-}1\ k \dots n].$$

The transposition $t(i, j, k)$ "cuts" the elements between the positions j and $k-1$ (both inclusive) and "pastes" them immediately before the i-th position. Let $\pi = [\pi_1 \pi_2 \dots \pi_{i-1} \pi_i \dots \pi_{j-1} \boxed{\pi_j \dots \pi_{k-1}} \pi_k \dots \pi_n]$, so

$$\pi \cdot t(i, j, k) = [\pi_1 \pi_2 \dots \pi_{i-1} \boxed{\pi_j \dots \pi_{k-1}} \pi_i \dots \pi_{j-1} \pi_k \dots \pi_n],$$

where the product of two permutations is denoted as an action to the right, a composition of two functions, in which π is applied first, and then $t(i, j, k)$.

Definition 2. *[1] The* transposition distance $d_t(\pi)$ *of a permutation π is the length q of the shortest sequence of transpositions $t_1, t_2, \dots, t_q$ such that $\pi t_1 t_2 \dots t_q = [1\ 2\ \dots\ n]$. If we have $\pi = [1\ 2\ \dots\ n]$, then we define $d_t(\pi) = 0$.*

In the study of the problem of determining the transposition distance, it is useful to give special names and symbols for some permutations. These are the *identity permutation of n elements*, denoted by $\iota_{[n]} := [1\ 2\ \dots\ n]$, the *reverse permutation of n elements*, denoted by $\rho_{[n]} := [n\ n{-}1\ \dots\ 2\ 1]$, and the *lonely permutation of n elements, beginning with the element ℓ*, such that $\gcd(n + 1, \ell) = 1$, denoted by $u_{n,\ell} := [\ell\ \overline{2\ell}\ \overline{3\ell}\ \dots\ \overline{n\ell}]$, where $\overline{x}$ is the remainder of the division of x by $n + 1$. One can readily observe that $\iota_{[n]} = u_{n,1}$ and $\rho_{[n]} = u_{n,n}$.

Toric Classes

Eriksson *et al.* [6] proposed an approach to the transposition distance problem: grouping some permutations that have the same distance into what they called *toric classes*. For further details, the reader is referred to [6,7,8].

Definition 3. *The* circularization *of an ordinary permutation π is the circular permutation π° obtained from π by inserting an extra element 0 as both predecessor of π_1 and successor of π_n, and taking the equivalence class under cyclic shifts. Write $\pi^\circ = (0\,\pi_1\,\ldots\,\pi_n)$ where the use of parentheses indicates an equivalence class under cyclic shifts. From a circular permutation π°, we uniquely retrieve the ordinary permutation π by removing the element 0 and letting its successor be the first element of π.*

Definition 4. *[6] Let π be a permutation of n elements, and m an integer. The* m-step cyclic value shift *of the circular permutation π° is the circular permutation $m + \pi^\circ := (\overline{m}\,\overline{m+\pi_1}\,\ldots\overline{m+\pi_n})$, where $\overline{x}$ is the remainder of the division of x by $n+1$.*

Definition 5. *[6] Two permutations π, σ are* torically equivalent *if $\pi^\circ \equiv^\circ m+\sigma^\circ$ for some integer m. Use $\pi^\circ_\circ$ to denote the toric equivalence class of π.*

Theorem 1. *[6] If π, σ are torically equivalent, then $d_t(\pi) = d_t(\sigma)$.*

This result allowed Eriksson *et al.* to reduce the search space to determine the transposition distance by a branch-and-bound algorithm, allowing them to achieve some results on the study of this problem, such as determining the maximum value of $d_t(\pi)$ for permutations having 13, 14 and 15 elements. However, this reduction in the search space is not enough for determining the transposition distance for permutations with a greater number of elements.

A toric class is *unitary* if it contains only one element, such as $\iota^\circ_{\circ[n]}$ and $\rho^\circ_{\circ[n]}$.

Theorem 2. *[7,8] A toric equivalence class is unitary if, and only if, it is of the form $\left[\,\overline{\ell}\;\;\overline{2\ell}\;\;\overline{3\ell}\;\;\ldots\;\;\overline{n\ell}\,\right]^\circ_\circ$, where $\gcd(\ell, n+1) = 1$.*

Therefore, the only unitary toric classes are those that contain a permutation $u_{n,\ell}$, which justifies the name *lonely permutation*. The permutations $u_{n,\ell}$, where $n+1$ is prime, have been studied in detail in [7,9].

The Reality and Desire Diagram

Another approach to the study of the transposition distance has been introduced by Bafna and Pevzner [1] and relies on the structure of the cycles of a graph that captures the structure of a permutation.

Definition 6. *[1,10] Given a permutation π of n elements, the* reality and desire diagram *$RD(\pi)$ is a graph on the following set of vertices:*

$$V(RD(\pi)) = \{0, -1, +1, -2, +2, \ldots, -n, +n, -(n+1)\},$$

and whose set of edges is partitioned into two sets R and D, respectively reality *and* desire *edges, defined as*

$$R = \{(+\pi_i, -\pi_{i+1}) \mid i = 1, \ldots, n-1\} \cup \{(0, -\pi_1), (+\pi_n, -(n+1))\},$$
$$D = \{(+i, -(i+1) \mid i = 1, \ldots, n-1\} \cup \{(0, -1), (+n, -(n+1))\}.$$

By the definition of the reality and desire diagram, every vertex has degree 2. Therefore, $RD(\pi)$ can be partitioned into a collection of disjoint cycles, and these cycles are alternating, i. e., the reality and desire edges in each cycle alternate. We say that a cycle has length k if it contains exactly k reality edges (which is the same as having k desire edges). If the length of a cycle is even, we say that it is an *even cycle*; otherwise it is an *odd cycle*. The *number of odd cycles* in the reality and desire diagram is denoted as $c_{odd}(\pi)$, and it is used to derive some known bounds for the transposition distance, as discussed further in Section 1.1.

Known Bounds and Studied Subsets of Permutations

By analyzing how a transposition applied to a permutation affected its reality and desire diagram, Bafna and Pevzner were able to provide the first non-trivial bounds for the transposition distance.

Theorem 3. *[1] A permutation π of n elements satisfies*

$$\frac{1}{2}\Big(n + 1 - c_{odd}(\pi)\Big) \leq d_t(\pi) \leq \frac{3}{4}\Big(n + 1 - c_{odd}(\pi)\Big).$$

Bafna and Pevzner were also the first to notice that the distance of the reverse permutation $\rho_{[n]}$ was in the range $\frac{n}{2} \leq d_t(\rho_{[n]}) \leq \frac{n+1}{2}$, theorizing that it was equal to the upper bound, and that the distance of any permutation of n elements would be $\frac{n+1}{2}$ at most, since the reverse permutation seemed to be the hardest permutation to transform into the identity by transpositions. Meidanis, Walter and Dias [10] proved the prediction $d_t(\rho_{[n]}) = \frac{n+1}{2}$ to be correct, but it was later found by Eriksson *et al.* [6] that there were permutations of 13 and 15 elements that had distance greater than $\frac{n+1}{2}$.

Improving on the earlier results, Elias and Hartman [4] were able to provide instances of permutations of n elements, for every odd $n \geq 17$, that were farther than $\frac{n+1}{2}$ transpositions from the identity. In the same paper, they also studied permutations whose cycles in the reality and desire diagram had length 3.

Definition 7. *A k-permutation is one in which every cycle in the reality and desire diagram has length k.*

Theorem 4. *[4] If π is a 3-permutation, then*

$$d_t(\pi) \leq 11 \left\lfloor \frac{n+1}{24} \right\rfloor + \left\lfloor \frac{3\left(\frac{n+1}{3} \bmod 8\right)}{2} \right\rfloor + 1.$$

Elias and Hartman were not the first to study k-permutations. Christie [11] determined previously the distance of any 2-permutation.

Theorem 5. *[11] If π is a 2-permutation, then $d_t(\pi) = \frac{n+1}{2}$.*

The fruitful results on the transposition distance for restricted classes of permutations, in a problem that has evaded a general solution so far, encouraged further efforts on the study of specific classes: Labarre's so-called γ-permutations [5] and the lonely permutations of n elements, where $n+1$ is prime [7,9]. In our previous studies of lonely permutations, we have found the following lower bounds.

Theorem 6. *[7] If $n+1$ is prime and $1 < \ell < n+1$, then $d_t(u_{n,\ell}) \geq \frac{n}{2}$.*

Theorem 7. *[9] If $n+1$ is prime, $\ell = \frac{n}{2} + 2$ and $\ell^* \equiv \ell^{-1} \pmod{n+1}$, then*

$$\frac{n}{2} + 1 \leq d_t(u_{n,\ell}) = d_t(u_{n,\ell^*}).$$

2 Reality and Desire Diagram of Lonely Permutations

Following the approach of previous works that sought better bounds for the transposition distance for some classes of permutations, we will focus on establishing bounds for lonely permutations. Lemma 1, which describes the structure of the reality and desire diagram of lonely permutations, lays the groundwork upon which our bounds for the transposition distance of lonely permutations rest. This lemma establishes a regularity of $RD(u_{n,\ell})$ according to the length of its cycles. This lemma and its proof are a generalization of the result in [7], which states that $u_{n,\ell}$ has just one cycle if $n+1$ is prime.

Lemma 1. *Let $u_{n,\ell}$ be a lonely permutation, with $\ell > 1$. Then $RD(u_{n,\ell})$ satisfies:*

1. *each cycle has length $k = (n+1)/\gcd(n+1, \ell-1)$, therefore $u_{n,\ell}$ is a k-permutation;*
2. *the number of cycles in the reality and desire diagram is $\gcd(n+1, \ell-1)$;*
3. *$+i, +\overline{i+\ell-1}, +\overline{i+2(\ell-1)}, \ldots, +\overline{i+(k-1)(\ell-1)}$ is the sequence of non-negative elements in every cycle, for $i = 0, \ldots, \gcd(n+1, \ell-1) - 1$.*

Proof. Let $+\pi_x$ be a vertex with a non-negative label in the reality and desire diagram $RD(\pi)$. The *non-negative successor* of $+\pi_x$, denoted as $s(+\pi_x)$, is the vertex $+\pi_y$ such that $+\pi_x, -\pi_{x+1}, +\pi_y$ appear in this order, in one of the two orderings of the vertices in the cycle containing $+\pi_x$.

Let $\pi = u_{n,\ell}$, that is, $\pi_i = \overline{i\ell}$. The finite sequence $0, s(0), s(s(0)), \ldots$ is thus equal to $0, +\overline{i_1\ell}, +\overline{i_2\ell}, \ldots, +\overline{i_{k-1}\ell}$ for integers $i_1, i_2, \ldots, i_{k-1}$, where k is to be determined later.

For $p = 1 \ldots k-2$, we thus have $s(+\overline{i_p\ell}) = +\overline{(i_p+1)\ell - 1}$ and $s(+\overline{i_p\ell}) = +\overline{i_{p+1}\ell}$. These two equalities, along with the base case $p = 1$, give us the following recurrence relation:

$$\begin{cases} i_1\ell \equiv \ell - 1 \pmod{n+1} \\ i_{p+1}\ell \equiv i_p\ell + \ell - 1 \pmod{n+1}, \text{ for } p > 1, \end{cases}$$

which is solved by the equivalence relation $i_p\ell \equiv p(\ell-1) \pmod{n+1}$.

Hence the sequence $(0, +\overline{i_1\ell}, +\overline{i_2\ell}, \ldots, +\overline{i_{k-1}\ell})$ becomes $(0, +\overline{\ell-1}, +\overline{2(\ell-1)}, \ldots, +\overline{(k-1)(\ell-1)})$. As we must have that $s(+\overline{(k-1)(\ell-1)}) = 0$, we must find the smallest positive value of k such that $k(\ell-1) \equiv 0 \pmod{n+1}$. It is easy to show that $k = (n+1)/\gcd(n+1, \ell-1)$.

We will now prove that for $i = 0, \ldots, gcd(n+1, \ell-1) - 1$, then all of the sequences of the form

$$(+i, +\overline{i+\ell-1}, +\overline{i+2(\ell-1)}, \ldots, +\overline{i+(k-1)(\ell-1)})$$

have no repeated elements. If there existed two repeated elements, then we would have that $+\overline{i+x(\ell-1)} = +\overline{j+y(\ell-1)}$, for some $0 \leq i, j \leq gcd(n+1, \ell-1)-1$ and integers x, y. Without loss of generality, we will suppose that $i < j$. Then, the equivalence $i+x(\ell-1) \equiv j+y(\ell-1) \pmod{n+1}$ can be written as $(x-y)(\ell-1) \equiv j-i \pmod{n+1}$, or as a diophantine equation $(\ell-1)x' + (n+1)y' = j-i$, where $x' = x-y$ and y' is integer. Bézout's Lemma states that the equation only has a solution if, and only if, $j - i \geq \gcd(\ell-1, n+1)$; but $j-i$ cannot be greater than $\gcd(n+1, \ell-1) - 1$. Hence, there are no repeated elements.

Each of these sequences – we have exactly $\gcd(n+1, \ell-1)$ of them – corresponds to a cycle in $RD(u_{n,\ell})$. Since every reality edge has one positive element, the length of every cycle is the number of positive elements in the cycle, so every cycle has the same size k. □

Given a lonely permutation $u_{n,\ell}$, we refer to the length of the cycles of its reality and desire diagram as the *cycle length of* $u_{n,\ell}$. Lemma 1 together with Bafna and Pevzner's lower bound in Theorem 3 give as a corollary an immediate lower bound when the cycle length is even; in the specific case that $u_{n,\ell}$ is a 2-permutation, Theorem 5 asserts us that the lower bound equals the distance.

Corollary 1. *If the cycle length* $k = n+1/\gcd(n+1, \ell-1)$ *is even, then* $(n+1)/2 \leq d_t(u_{n,\ell})$. *If* $k = 2$, *or equivalently* $\ell = \frac{n+3}{2}$, *the lower bound equals the transposition distance.*

The assumption that k is even implies that $n+1$ is even, and so n is odd. Since $\gcd(n+1, \ell) = 1$, the assumption that n is odd implies that $\gcd(n+1, \ell-1)$ is even. Therefore, the assumption that k is even implies that $n \equiv 3 \pmod{4}$. Corollaries 2 and 3 identify, respectively, an infinite set of values of n such that all lonely permutations $u_{n,\ell}$ have even cycle length, and conditions for the existence of a value of ℓ such that $u_{n,\ell}$ has even cycle length. Those infinite families satisfy $(n+1)/2 \leq d_t(u_{n,\ell})$, the distance of the reverse permutation and close to the distance of the farthest permutations known to date [4,6,10].

Corollary 2. *If* $n = 2^q - 1$ *for* q *integer, then* every *lonely permutation* $u_{n,\ell}$, *with* $\ell > 1$, *has even cycle length.*

Corollary 3. *The lonely permutation* $u_{n,3}$ *such that* $n \equiv 3 \pmod{4}$ *and* 3 *does not divide* $n+1$ *has even cycle length.*

The established regularity in the cycle length is further studied in Lemma 2 and its Corollary 4 with the goal of determining the effect of a transposition in the reality and desire diagram.

Lemma 2. *Consider the sequence of non-negative elements in the reality and desire diagram* $0, +\overline{\ell-1}, +\overline{2(\ell-1)}, \ldots$ *in the cycle in* $RD(u_{n,\ell})$ *that contains the vertex* 0*. The position of the corresponding elements* $0, \overline{\ell-1}, \overline{2(\ell-1)}, \ldots$ *in the permutation is, respectively,* $0, m, \overline{2m}...$*, where* $m = \overline{1-\ell^{-1}}$*.*

Proof. Let $\pi = u_{n,\ell}$. We want to find the value x such that $\pi_x = \overline{y(\ell-1)}$, for $y = 1 \ldots n$. Since π is a lonely permutation, we have that $\overline{x\ell} = \pi_x = \overline{y(\ell-1)}$, that is, $x\ell \equiv y(\ell-1) \pmod{n+1}$, which, multiplied by ℓ^{-1} to the right on both sides of the equivalence becomes $x \equiv y(1-\ell^{-1}) \pmod{n+1}$. □

Corollary 4. *The position of the element* $\overline{i + y(\ell-1)}$ *is* $\overline{i\ell^{-1} + ym}$*, where* $m = \overline{1-\ell^{-1}}$*. This means that the cycles containing* $+i, \overline{+i+(\ell-1)}, \overline{+i+2(\ell-1)}, \ldots$ *and* $0, \overline{+\ell-1}, \overline{+2(\ell-1)}, \ldots$ *have the same structure.*

Proof. Notice that the element i is the $i\ell^{-1}$-th element in $u_{n,\ell}$, and proceed with the proof in a fashion similar to the proof of Lemma 2. □

Bafna and Pevzner [1] proved that, after applying a transposition to a permutation, the number of odd cycles in the reality and desire diagram changes in one of the following ways: i) it *increases* by two units; ii) it does not change; or iii) it *decreases* by two units. Every transposition can be classified according to its effect on the number of odd cycles into a: i) 2-move; ii) 0-move; or iii) −2-move, respectively. We will now characterize the existence of 2-moves that create cycles of length 1. This characterization will be used in Section 3 for providing an upper bound on the transposition distance for 3-permutations.

Definition 8. *[1] We say that a transposition* $t(i,j,k)$ affects *a cycle* c *in the reality and desire diagram* $RD(\pi)$ *if* c *contains at least one of the following reality edges:* $+\pi_{i-1}, -\pi_i$ *or* $+\pi_{j-1}, -\pi_j$ *or* $+\pi_{k-1}, -\pi_k$*.*

Lemma 3. *For a lonely permutation* $u_{n,\ell}$*, there exists a* 2*-move that creates* 2 *cycles of length* 1 *if, and only if,* $\overline{1-\ell^{-1}} > n/2$*.*

Proof. Since all the cycles have the same structure, as per Corollary 4, it does not matter which of the cycles we base our analysis of the effect of a transposition. We will, therefore, suppose that a transposition affects the cycle containing $0, +\overline{\ell-1}, +\overline{2(\ell-1)}, \ldots$ The corresponding elements of the permutation occur, respectively, in the 0-th, $\overline{m}$-th, $\overline{2m}$-th, ... positions in the permutation, where $m = \overline{1-l^{-1}}$. For a transposition to break this cycle into three cycles, with two of them having length 1, we must have that there are three consecutive elements in positions $\overline{xm}$, $(x+1)m$, $(x+2)m$, with $0 \leq x \leq \gcd(n+1, \ell-1) - 3$, where those three elements are in a decreasing circular order – i. e. $x(\ell-1) > (x+1)(\ell-1) > (x+2)(\ell-1)$ or $(x+1)(\ell-1) > (x+2)(\ell-1) > x(\ell-1)$ or $(x+2)(\ell-1) > x(\ell-1) > (x+1)(\ell-1)$. A case analysis (which we will omit for brevity) shows us it only happens in the case that $\overline{1-\ell^{-1}} > n/2$. □

A sequence of 2-moves, if it exists, is the shortest way of transforming a permutation into the identity. Although we cannot guarantee the existence of such a

sequence that transforms a lonely permutation into the identity, we will show in Theorem 8 that there is a sequence of 2-moves that generate 1-cycles for $u_{n,\ell}$, in the case that ℓ divides $n+2$, which relates it to another lonely permutation $u_{n-2(\ell-1),\ell}$, a permutation that has less elements.

Definition 9. *[11] The reduced permutation of π, denoted by $gl(\pi)$, is a permutation whose reality and desire diagram $RD(gl(\pi))$ is equal to $RD(\pi)$, without the 1-cycles, and keeping the order of the elements.*

Definition 10. *[11] Two permutations π and σ are* equivalent by reduction, *if $gl(\pi) = gl(\sigma)$.*

Christie [11] proved that, if two permutations are equivalent by reduction, then they have the same transposition distance.

Definition 11. *[11] A permutation π is* irreducible *if $gl(\pi) = \pi$.*

Notice that every lonely permutation that is not the identity is irreducible, for a lonely permutation only has one cycle of length 1 in the reality and desire diagram if every cycle has length 1.

Definition 12. *A permutation σ is an r-reduction of π if there is a sequence of r transpositions that are 2-moves that transforms π into a permutation that is equivalent by reduction to σ.*

Corollary 5. *If σ is an r-reduction of π, then $d_t(\pi) \leq d_t(\sigma) + r$.*

Theorem 8. *Let $u_{n,\ell}$ be a lonely permutation such that ℓ divides $n+2$. Then $d_t(u_{n,\ell}) \leq d_t(u_{n-2(\ell-1),\ell}) + \ell - 1$.*

Proof. We will show that $u_{n-2(\ell-1),\ell}$ is an $(\ell-1)$-reduction of $u_{n,\ell}$. The result will then follow as a consequence of Corollary 5.

Since ℓ divides $n+2$, consider $d := \frac{n+2}{\ell}$. It immediately follows that $\ell d \equiv n+2 \equiv 1 \pmod{n+1}$, which implies that $d = \overline{\ell^{-1}}$. The position of the element i is, thus, $\overline{i\ell^{-1}} = i\overline{\frac{n+2}{\ell}}$.

Calculating the positions of the elements $1, 2, \ldots, \ell-1$, we have the increasing sequence $1d, 2d, \ldots, (\ell-1)d$, since $i\frac{n+2}{\ell} \leq n$ for $i \leq \ell - 1$. For the elements $n+2-\ell, n+2-\ell+1, n+2-\ell+2, \ldots, n+2-\ell+\ell-2 = n$, their corresponding positions are $1d-1, 2d-1, \ldots, (\ell-1)d-1$. Let $n' = n+2-\ell$; the lonely permutation $u_{n,\ell}$ has the following structure:

$$[\ell \ldots n' \ \boxed{1 \ldots} n'+1 \ \boxed{2 \ldots} n'+2 \ \boxed{3 \ldots} \ldots n \ \boxed{\ell-1 \ldots n\ell}]$$

The sequence of $\ell - 1$ transpositions $t(1, d, 2d-1), \ldots, t(i, id, (i+1)d-1), \ldots, t(\ell-1, (\ell-1)d, n+1)$ transforms $u_{n,\ell}$ into

$$[\underbrace{1 \ 2 \ 3 \ \ldots \ \ell-1}_{\ell-1 \text{ elements}} \ \ell-1+\ell \ \ell-1+2\ell \ \ldots \ \underbrace{n' \ n'+1 \ n'+2 \ \ldots n}_{\ell-1 \text{ elements}}],$$

which is easy to show that is equivalent by reduction to $u_{n-2(\ell-1),\ell}$.

It remains to show that all of the $\ell - 1$ transpositions applied to $u_{n,\ell}$ are 2-moves. Since the permutation $[1\ \ 2\ \ \ldots\ \ \ell{-}1{+}\ell\ \ \ell{-}1{+}2\ell\ \ \ldots n'\ \ \ldots n]$ that comes after applying the transpositions to $u_{n,\ell}$ is equivalent by reduction to $u_{n-2(\ell-1),2}$, it must have the number of cycles of $u_{n-2(\ell-1),\ell}$ plus the number of cycles of length 1.

The number of cycles in $u_{n-2(\ell-1),\ell}$ equals the number of cycles of $u_{n,\ell}$, for $\gcd(n-2(\ell-1)+1, \ell-1) = \gcd(n+1, \ell-1)$. One can easily observe that the $2(\ell-1)$ elements $1, 2, \ldots, \ell-1, n', n'+1, \ldots, n$ create, each of them, a cycle of length 1 in $[1\ \ 2\ \ \ldots\ \ \ell{-}1{+}\ell\ \ \ell{-}1{+}2\ell\ \ \ldots n'\ \ \ldots n]$. Therefore, the permutation that is equivalent by reduction to $u_{n-2(\ell-1),\ell}$ has $2(\ell-1)$ more cycles than $u_{n,\ell}$. Since we can only add, at most, 2 cycles after applying a transposition, we must deduce that every transposition is a 2-move. □

3 An Upper Bound for Lonely 3-Permutations

A permutation is said to be *simple* if every cycle in the reality and desire diagram is of length at most 3. A really promising approach to estimate the distance of a permutation is to transform it into a simple permutation by a series of operations, and then calculate the distance of the resulting simple permutation [4]. In Section 2 we have already dealt with lonely 2-permutations. In this section we study the remaining simple permutations that are also lonely permutations – the lonely 3-permutations – with the aim of providing tight bounds for their distance. We begin with a characterization of those permutations in Lemma 4, and conclude this section with an upper bound for their distance in Theorem 9, which seems to be tight, as discussed in Section 4.

Lemma 4. *A lonely permutation $u_{n,\ell}$ is a 3-permutation if, and only if,*

$$\begin{cases} n \equiv 2 \pmod 9 \text{ if } \ell = \frac{n+4}{3} \\ n \equiv 5 \pmod 9 \text{ if } \ell = \frac{2n+5}{3} \\ n \equiv 8 \pmod 9 \text{ if } \ell = \frac{n+4}{3} \text{ or } \frac{2n+5}{3} \end{cases}$$

Proof. Since every cycle of a lonely permutation has the same length, it is also a 3-permutation if, and only if, $n+1$ is a multiple of 3, i. e. $n+1 \equiv 0 \pmod 3$. Therefore, we have three distinct possibilities for n: n is equivalent to 2, 5 or 8 modulo 9.

Since $n+1 \equiv 0 \pmod 3$, we can write it as $n+1 = 3q$, for some q integer. Since $u_{n,\ell}$ is a 3-permutation, we have that q equals the number of cycles, that is, $q = \gcd(n+1, l-1)$. Since $\ell < n+1 = 3q$, and q divides $\ell - 1$, we must consider only two cases for $\ell - 1$: either $\ell - 1 = q$ or $\ell - 1 = 2q$.

If $\ell - 1 = q$, it follows that $\ell = q + 1 = \frac{n+1}{3} + 1 = \frac{n+4}{3}$. Since ℓ cannot be a multiple of 3, we have that $n \equiv 2 \pmod 9$ or $n \equiv 8 \pmod 9$.

If $\ell - 1 = 2q$, it follows that $\ell = \frac{2n+5}{3}$. Since ℓ cannot be a multiple of 3, we have that $n \equiv 5 \pmod 9$ or $n \equiv 8 \pmod 9$. □

Definition 13. *[1] A permutation is* oriented *if it has a 2-move. Otherwise, it is unoriented.*

Lemma 5. *A lonely 3-permutation is oriented if: i)* $n \equiv 2 \pmod 9$ *and* $\ell = \frac{n+4}{3}$*; or ii)* $n \equiv 8 \pmod 9$ *and* $\ell = \frac{2n+5}{3}$*. It is unoriented if: iii)* $n \equiv 5 \pmod 9$ *and* $\ell = \frac{2n+5}{3}$*; or iv)* $n \equiv 8 \pmod 9$ *and* $\ell = \frac{n+4}{3}$*.*

Proof. Calculating $m = \overline{1 - \ell^{-1}}$ for each case, we have that $m > n/2$ in the cases *i* and *ii*, and $m \leq n/2$ in the cases *iii* and *iv*. The validity of the hypothesis comes as a consequence of Lemma 3. □

Lemma 6. *Every 2-move that is applied to an oriented lonely 3-permutation* $u_{n,\ell}$*, with* $n > 2$*, results in an unoriented permutation that is equivalent by reduction to* $u_{n-3,\ell'}$*, where*

$$\ell' = \begin{cases} \ell - 1 & \text{if } \ell = \frac{n+4}{3} \\ \ell - 2 & \text{if } \ell = \frac{2n+5}{3} \end{cases}$$

Proof. The possible 2-moves are $t(i, \overline{2m+i}, \overline{m+i})$, where $m = \overline{1-\ell^{-1}}$, $i = 1 \ldots m-1$; these 2-moves generate two 1-cycles according to Lemma 3. Apply them, considering $\ell = \frac{n+4}{3}$ or $\ell = \frac{2n+5}{3}$, and eliminate the 1-cycles. □

Definition 14. *An irreducible permutation* σ *is a* (t, r)*-reduction of a permutation* π *if there is a sequence of* t *transpositions, of which* r *of them are 2-moves, that transforms* π *into a permutation that is equivalent by reduction to* σ*.*

Corollary 6. *If* σ *is a* (t, r)*-reduction of* π*, then* $d_t(\pi) \leq d_t(\sigma) + t$*.*

Lemma 7. *Let* $u_{n,\ell}$ *be an unoriented lonely 3-permutation, with* $n > 5$*. Then the oriented lonely 3-permutation* $u_{n-6,\ell'}$ *is a* $(3, 2)$*-reduction of* $u_{n,\ell}$*, where*

$$\ell' = \begin{cases} \ell - 2 & \text{if } \ell = \frac{n+4}{3} \\ \ell - 4 & \text{if } \ell = \frac{2n+5}{3} \end{cases}$$

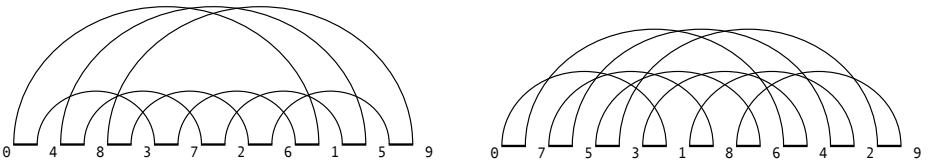
Proof. Apply the sequence $t(1, m+1, \overline{2m+1})$, $t(2, m+2, \overline{2m+2})$ and $t(1, m+1, \overline{2m+1})$ to $u_{n,\ell}$, with $m = \overline{1-\ell^{-1}}$, where the first one is a 0-move and the others are 2-moves, considering separately $\ell = \frac{n+4}{3}$ and $\ell = \frac{2n+5}{3}$. □

Theorem 9. *Let* $u_{n,\ell}$ *be a lonely 3-permutation. We have that*

$$d_t(u_{n,\ell}) \leq \begin{cases} \frac{4n+1}{9} & \text{if } n \equiv 2 \pmod 9 \\ \frac{4n+7}{9} & \text{if } n \equiv 5 \pmod 9 \\ \frac{4n+4}{9} & \text{if } n \equiv 8 \pmod 9 \end{cases}$$

Proof. If $n \leq 8$, the lonely 3-permutations are: $u_{2,2}$, $u_{5,5}$, $u_{8,4}$ and $u_{8,7}$. It is easy to show that their distances are 1, 3, 4 and 4 respectively. For $n > 8$, consider the same 4 cases of Lemma 5.

In the case *i* we have $n \equiv 2 \pmod 9$ and $u_{n,\ell}$ is oriented. We can apply Lemma 6 to $u_{n,\ell}$, obtaining $u_{n-3,\ell'}$, and apply Lemma 7 to $u_{n-3,\ell'}$, obtaining the oriented permutation $u_{n-9,\ell''}$ (the values of ℓ' and ℓ'' are not important in this case). Therefore, we have the recurrence relation $d_t(u_{n,\ell}) \leq d_t(u_{n-9,\ell''}) + 4$,



(a) (b)

Fig. 1. Reality and desire diagrams of: (a) $u_{8,4}$ and (b) $u_{8,7}$

Table 1. Lonely permutations. Shaded cells correspond to lonely 3-permutations

$d_t(u_{n,\ell})$		ℓ																
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
n	2	**1**	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	3	—	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	4	2	2	3	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	5	—	—	—	**3**	—	—	—	—	—	—	—	—	—	—	—	—	—
	6	3	4	3	4	4	—	—	—	—	—	—	—	—	—	—	—	—
	7	—	4	—	4	—	4	—	—	—	—	—	—	—	—	—	—	—
	8	4	—	**4**	4	—	**4**	5	—	—	—	—	—	—	—	—	—	—
	9	—	5	—	—	—	5	—	5	—	—	—	—	—	—	—	—	—
	10	5	6	6	6	5	6	6	6	6	—	—	—	—	—	—	—	—
	11	—	—	—	**5**	—	6	—	—	—	6	—	—	—	—	—	—	—
	12	6	7	7	7	6	6	7	7	7	6	7	—	—	—	—	—	—
	13	—	7	—	7	—	—	—	7	—	7	—	7	—	—	—	—	—
	14	7	—	7	—	—	7	7	—	—	**7**	—	7	8	—	—	—	—
	15	—	8	—	8	—	8	—	8	—	8	—	8	—	8	—	—	—
	16	8	9	8	9	9	9	8	8	9	9	9	8	9	8	9	—	—
	17	—	—	—	9	—	**8**	—	—	—	9	—	**8**	—	—	—	9	—
	18	9	9	10	10	10	10	9	10	9	10	9	9	9	9	10	10	10

with the base case $d_t(u_{2,2}) = 1$. This relation has the following closed formula: $d_t(u_{n,\ell}) \leq \frac{4n+1}{9}$. Case *iii* can be solved similarly; observe that Lemma 7 must be applied first, and then Lemma 6, and that the base case is $d_t(u_{5,5}) = 3$.

Cases *ii* and *iv* can be dealt with as just one case. The permutation $u_{n,\ell}$ can be transformed into $u_{n-9,\ell''}$ with 4 transpositions – just apply Lemmas 6 and 7 in the appropriate order. The recurrence relation $d_t(u_{n,\ell}) \leq d_t(u_{n,\ell''}) + 4$ has either $d_t(u_{8,4}) = 4$ or $d_t(u_{8,7}) = 4$ as a base case, which means the solution is the same for both cases: $d_t(u_{n,\ell}) \leq \frac{4n+4}{9}$. □

4 Conclusion

There are two lonely 3-permutations for every $n \equiv 8 \pmod 9$, but just one for every $n \equiv 2$ or $5 \pmod 9$, and observe an interesting property of the values of ℓ for each n such that $u_{n,\ell}$ is a 3-permutation: for $n \equiv 2$ or $5 \pmod 9$, we have that $\ell \equiv \ell^{-1} \pmod 9$, whereas for $n \equiv 8 \pmod 9$ each of the two possible values is the inverse of the other modulo $n + 1$. This means that for $n \equiv 8 \pmod 9$ not only the upper bound in Theorem 9 for the two lonely 3-permutations is the

same, but they also must have the same distance, according to a result in [9]. Even more puzzling is the fact that one of them is *unoriented* – Figure 1a – while the other is *oriented* – Figure 1b. It could be expected that, since one of the two permutations does not require a first 0-move, it would be easier to sort.

As in our previous work [7], we have computed the transposition distances of every lonely permutation for n up to 18, in order to compare our bounds with the exact transposition distance. The results are in Table 1.

Notice the tightness of the lower bound for lonely permutations with even cycle length. The upper bound for the lonely 3-permutations in the table is also really tight: it always equals the distance. We believe that the upper bound of Theorem 9 is indeed the transposition distance.

References

1. Bafna, V., Pevzner, P.A.: Sorting by transpositions. SIAM J. Disc. Math. 11(2), 224–240 (1998)
2. Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F., Cedergren, R.: Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. Proc. Natl. Acad. Sci. 89(14), 6575–6579 (1992)
3. Boore, J.L.: The duplication/random loss model for gene rearrangement exemplified by mitochondrial genomes of deuterostome animals. In: Sankoff, D., Nadeau, J.H. (eds.) Comparative Genomics, pp. 133–148. Kluwer Academic Publishers, Dordrecht (2000)
4. Elias, I., Hartman, T.: A 1.375-approximation algorithm for sorting by transpositions. IEEE/ACM Trans. Comput. Biol. and Bioinformatics 3(4), 369–379 (2006)
5. Labarre, A.: New bounds and tractable instances for the transposition distance. IEEE/ACM Trans. Comput. Biol. and Bioinformatics 3(4), 380–394 (2006)
6. Eriksson, H., Eriksson, K., Karlander, J., Svensson, L., Wästlund, J.: Sorting a bridge hand. Discrete Mathematics 241(1), 289–300 (2001)
7. Hausen, R., Faria, L., Figueiredo, C.M., Kowada, L.A.: On the toric graph as a tool to handle the problem of sorting by transpositions. In: Bazzan, A.L.C., Craven, M., Martins, N.F. (eds.) BSB 2008. LNCS (LNBI), vol. 5167, pp. 79–91. Springer, Heidelberg (2008)
8. Hultman, A.: Toric Permutations. Master's thesis, Department of Mathematics, KTH, Stockholm, Sweden (1999)
9. Hausen, R., Faria, L., Figueiredo, C.M., Kowada, L.A.: Unitary toric classes, the reality and desire diagram, and sorting by transpositions. SIAM J. Disc. Math. (to appear, 2010)
10. Meidanis, J., Walter, M.E.M.T., Dias, Z.: Transposition distance between a permutation and its reverse. In: Baeza-Yates, R. (ed.) Proc. 4th South American Workshop on String Processing, pp. 70–79. Carleton University Press (1997)
11. Christie, D.A.: Genome Rearrangement Problems. PhD thesis, University of Glasgow (1999)

Insights on Haplotype Inference on Large Genotype Datasets

Rogério S. Rosa and Katia S. Guimarães

Federal University of Pernambuco, Center of Informatics,
Recife, Brazil
{rsr,katiag}@cin.ufpe.br

Abstract. In this paper we present insights on the problem of haplotype inference for large genotype datasets. Our observations are drawn from an extensive comparison of three methods for haplotype inference using several datasets taken from HapMap. The methods chosen, PTG, Haplorec, and fastPHASE, are among the best known; they are based on different approaches, and are able to deal with large amounts of data. Our analysis controls the execution time and also the accuracy of results, based on the Error Rate and the Switch Error, as well as sequence conservation patterns. The results show that (1) fastPHASE and Haplorec are both more accurate than PTG, (2) fastPHASE is computationally the most expensive of the three methods, while Haplorec may fail to resolve long sequences, and (3) all approaches do better with more conserved sequences, and tend to fail in distinct sequence sites.

Keywords: Haplotype, Genotype, Inference, SNP, fastPHASE, Haplorec, PTG.

1 Introduction

An important challenge in biology is correlating differences among phenotypes with variations in the DNA. It is known that the human genome is highly similar for different individuals of the same population, and that some regions of the DNA sequence are preserved along generations. In these preserved regions, changes can occur in some specific alleles called Single Nucleotide Polymorphism (SNP). Besides a strong relation with phenotypes, SNPs can also be related with genetic diseases, and for that reason it is desirable to map them. Identifying these specific changes is not an easy task because it requires haplotype information. Diploid organisms have two copies of each chromosome, each copy is called a haplotype and a collection of these homologue chromosomes is called a genotype. Due to technological constraints, only the latter is available in large scale. Getting haplotype information directly is costly, therefore computational methods to infer haplotype data from genotype data are highly desirable.

Several computational methods have been proposed for inferring haplotypes from genotype data. Currently the challenge is to infer haplotypes from large scale genotype, due to the high computational costs of these approaches and

C.E. Ferreira, S. Miyano, and P.F. Stadler (Eds.): BSB 2010, LNBI 6268, pp. 47–58, 2010.

the need of real applications. Though there are some comparative studies of haplotype inference methods, they are old [1],[2] and do not include more recent approaches or sometimes methods that are considered as good techniques for inferring haplotypes in large scale.

In this paper we developed an extensive analysis involving methods fastPHASE [3], Haplorec [4], and PTG [5], each of which is based on a different computational technique. These approaches were selected because they are deemed to be among the best algorithms in the literature known to resolve very large datasets. For this benchmark, datasets collected from HapMap project were used. We evaluated the capacity of each method to rebuild the correct haplotype data set, and also the time required for finding a good solution. We also analyzed what type of characteristics of the dataset, such as conservation level, for instance, made a method more effective.

The rest of this paper is organized as follows. The next section gives a formal definition of the Haplotype Inference Problem, and presents related work, briefly introducing the PTG, Haplorec, and fastPHASE methods. In Section 3, the design of the experiments and the data used are described. In section 4, the results of our comparative analysis are shown. A discussion and concluding remarks are presented in Section 5.

2 Haplotype Inference Methods

A genotype (haplotype) can be computationally represented by a vector based on alphabet {0, 1, 2} ({0, 1}), where a symbol 2 represents an ambiguous site. Two haplotype vectors h_1 and h_2 are said to explain a genotype vector g, denoted $h_1 \otimes h_2 = g$, if each one of them has n sites and, for each site i, $1 \leq i \leq n$, $h_1(i), h_2(i) \in \{0, 1\}$ and follow the rule given by: (A) $h_1(i) = h_2(i) = g(i)$, if $g(i) \in \{0, 1\}$; and (B) $h_1(i) = 1 - h_2(i)$, if $g(i)$=2.

The Haplotype Inference Problem basically consists of finding for each genotype g two haplotypes h_1 and h_2 such that $h_1 \otimes h_2 = g$, i.e. h_1 and h_2 explain g in a biologically plausible away. For instance, if g=(0,1,2,2,1,2), possible solutions are h_1=(0,1,0,0,1,0) and h_2=(0,1,1,1,1,1), or still h_1=(0,1,0,1,1,1) and h_2=(0,1,1,0,1,0), among other possibilities. It is easy to see that there are 2^{k-1} candidate haplotype pairs to explain g, where k is the number of ambiguous sites in g. Obviously, there are many solutions for each input g, so a biological model criterion is needed to define a good solution, such as the Pure Parsimony model.

Many combinatorial and statistical methods have been proposed for haplotype inference, such as the Clark Method [6], Integer Programming formulations based on the parsimony principle [7],[8],[9],[10],[11], phylogeny based [12],[13], and Bayesian methods [14],[15]. More recently, other methods were projected, such as methods based on Markov Chain Models [16],[17] and haplotyping via Genetic algorithms [18],[19].

There are two main biological models used to infer haplotypes: Pure parsimony and Perfect phylogeny. The parsimony-based methods search the solution space for the minimum distinct set of haplotypes that explain the genotype

data. Unfortunately, the problem of inferring haplotypes by pure parsimony is NP-hard [7]. The perfect phylogeny-based methods build a tree structure and the haplotypes inferred should form a perfect phylogeny. Inferring haplotypes using this principle is a polynomial-time problem, but the assumption that the DNA sequences were not subject to recombination events is not realistic [13].

Haplotype Inference by Pure Parsimony principle (HIPP) has been used by many approaches because of its innate simplicity and biological soundness. Initially, Integer Programming (IP) was applied to solve HIPP. There are three main problem formulations using IP. The first one, called RTIP [8], has an exponential number of constraints in the worst case, but it has been shown [11] that when this formulation can resolve the input dataset, it may be faster than more recent IP formulations. Two independent groups [9],[10] have introduced polynomial-size IP formulations for HIPP, one of them known as PolyIP, however, this formulation is very slow for large datasets, according to Brown and Harrower [11], who proposed a hybrid formulation with a polynomial number of constraints but faster than PolyIP, called HybridIP.

Other methods for HIPP were published. Among them, Parsimonious Tree-Grow (PTG) offers a good compromise of high accuracy at a relatively low computational complexity. The PTG method explains a set of m genotypes of length n in time $O(m^2 n)$. The accuracy of PTG is demonstrated theoretically and also experimentally [5], using comparisons involving HAPAR, HAPLOTYPER, HAPINFERX and PHASE. However, many operations in PTG are random, so it is necessary to run the method many times, and select the best solution using some metric, in order to have reliable results. The metric applied for quality is the number of distinct haplotypes inferred.

Methods based on Markov Chain Model have been proposed successfully. The Markov Chain was applied to the HI problem originally by Eronen and colleagues [20]. Later works improved this approach [16],[17],[21],[4]. These methods basically build a Markov chain where each state is a possible symbol (0 or 1) and the transition probabilities are calculated from the input data. It is possible to estimate the probability that each candidate haplotype fragment will be part of the optimal solution. The number of haplotype candidates and probability combinations (as the original problem) is clearly exponential. For this reason, heuristics based on Dynamic Programming and Expectation Maximization algorithms have been proposed [16],[21]. These heuristics search the solution space for a good explanation of the input data. The goal is to maximize the probability that a candidate haplotype will be part of resolution of several genotypes. Thus a haplotype vector will resolve a bigger number of genotypes and would be more plausible biologically. Based on this approach the software called Haplorec [4], available on the web, was developed; version Haplorec 2.3 is an improved version that solves very large datasets.

Although not as recent as Haplorec, PHASE [14],[22] and fastPHASE [3] are considered good classical approaches for the HI Problem. These methods use maximum likelihood to estimate haplotype frequencies. The objective is determining the maximum value of this likelihood function. Such methods are stochastic

and each execution of the program may result in a different solution, since the derivations are dependent on the initial configuration which is randomly selected. Basically, fastPHASE is a variation of PHASE for resolving large data sets.

Due to their computational performance, PTG, Haplorec and fastPHASE are good candidates for resolving genotypes in large scale. Some works in the literature compare haplotype inference methods [1],[2], but Haplorec and PTG are not included in those analyses, because they were proposed later. The techniques have different behavior and accuracy on distinct datasets. We can assume that the quality of the solution is closely related to the configuration and properties of the input genotypes. Our goal is to look closer into that connection.

3 Experiments Design

The HapMap project [23] is an important source of information about haplotypes shared among individuals from the same population. This international project was started by NIH (National Institute of Health) in 2002, with the goal of mapping these shared sequences. For experiments in this work, haplotypes were collected from HapMap Phase III, of chromosome 20 from a population of Trios Utah residents with Northern and Western European ancestry; this is 1 of 11 populations in HapMap Phase III, called Caucasian European in Utah (CEU). We selected dataset CEU because of the large number of individuals phased in it (88 individuals).

Chromosome 20 spans about 62 million DNA building blocks and represents approximately 2% of the total DNA in cells. This chromosome has between 700 and 800 genes. Many genetic diseases can be related with it. Changes in chromosome 20 have been identified in several types of cancer such as leukemia and lymphoma. Deletions or duplications of genetic material from chromosome 20 can have a variety of effects, including intellectual disability, delayed development, distinctive facial features, skeletal abnormalities, and heart defects. The raw data presents 36258 SNPs and 88 individuals (two sequences for each individual). We mapped all haplotype fragments of length 100, 200, 400, 800, and 1600 to datasets in classes A, B, C, D, and E, respectively. For each class, the fragments with the smallest, the average, and the biggest number of distinct original haplotypes were selected to form three distinct datasets, called X1, X2, and X3, respectively, in which X can be A, B, C, or D. Since all the sequences in class E had virtually the same number of distinct haplotypes, this class has only one dataset, called E1. All together, 15 datasets were generated.

Each measure presented many candidate sets, for instance, there were 18 sets with 200 SNPs, 88 individuals and 13 distinct haplotypes (13 is the smallest number of distinct haplotypes found for a matrix 200x88 in CEU), for this reason, a random choice for selecting a representative set was needed. In the case of the set with 1600 SNPs, the mapping was discarded because the smallest number of distinct haplotypes found was 175 while the largest was 176 (really similar). Due to that, a subset was randomly selected in CEU with 1600 SNPs.

The purpose of mapping the occurrences of distinct haplotype fragments is to identify the abundance of easy and difficult datasets for resolution in a specific

population. Table 1 shows the amount of distinct haplotypes found in CEU with the specific number of SNPs, and their respective number of representatives in CEU. With the integral data of the map, it is possible to identify the regions of chromosome 20 in CEU in which there is a concentration of haplotype fragments with similar properties, such as conserved sequences. According to the map, an increase in the SNPs number implies in an increase in the number of distinct haplotype fragments. This is crucial for the analysis of the methods because different methods have distinct behaviors when resolving genotypes originated from more conserved or from more recombined sequences.

Table 1. Distinct Haplotype Fragments Found: Number of Distinct Haplotypes (DH) in CEU, and their respective number of representatives

N SNPs	Least DH	Larger DH	Average DH
100	13 / 7	175 / 16	101 / 370
200	38 /18	175 / 671	145 / 330
400	102 / 2	176 / 3187	168 / 919
800	151 / 57	176 / 15923	175 / 12002

The metrics used in the benchmark were Error Rate [15], Switch Error [24], computational time, and number of distinct haplotypes inferred. The error metrics are based on known haplotype sets, and tell us about the capacity that a method has to correctly infer a haplotype set from a genotype set. The computational time is an empiric metric used for estimating computational cost. Although in general it is not the best technique for that, in this case, theoretical analysis cannot be applied to all methods. The number of distinct haplotypes is an important metric because the parsimony principle is being considered, and methods that find a minimal solution set are required. Time and number of distinct haplotypes are quantitative metrics.

Error Rate considers the rate of haplotype sites inferred incorrectly in all haplotype sites previously known. Switch Error is the proportion between the number of genotype ambiguous sites and the number of fragments changes needed between two specific haplotype vectors used to explain a specific genotype.These are considered quality metrics. Basically, Error Rate measures site to site individually, while Switch Error considers the neighbor sites. The goal here is to minimize the Error Rate and maximize the Switch Error (maximum value is 1).

For the comparison experiments, PTG was implemented in MATLAB 2008, for method Haplorec (haplotype inference based on Markov chains) the software version 2.3 was used, and for fastPHASE the version 1.2.3 for windows was used. The experiments run individually in a computer with an Intel Quad Core 2.33GHz processor, with 3GB of RAM. Although there is a version of PTG available on the author's webpage, we have chosen to develop our own version of the method, because, since PTG is based on random operations, in order to find reliable results, PTG needs to be executed many times (in our case, we chose 30 runs), taking the best output. In our experiments, the smaller number of distinct haplotypes was used as metric for selecting the best solution.

Table 2. Comparison Results: Error Rate (ER), Distinct Haplotypes found (DH), Time in seconds (s), minutes (m) or hours (h)

Set	PTG			Haplorec			fastPHASE		
	ER	DH	Time	ER	DH	Time	ER	DH	Time
A1	0,35%	12	21,24 s	0,01%	12	3 s	0,00%	13	13 m
A2	61%	118	40,48 s	3,26%	98	5 s	1,96%	96	11 m
A3	12,98%	166	53,88 s	8,97%	176	9 s	8,97%	176	5 m
B1	3,17%	46	1,16 m	0,70%	64	10 s	0,10%	44	10 m
B2	9,54%	155	1,81 m	4,65%	161	19 s	4,21%	158	25 m
B3	10,01%	168	1,62 m	6,05%	176	20 s	5,78%	176	10 m
C1	3,65%	112	2,63 m	0,75%	120	21 s	1,09%	115	54 m
C2	13,87%	169	3,75 m	6,63%	174	49 s	7,65%	176	52 m
C3	12,53%	174	3,73 m	8,78%	176	49 s	9,49%	176	25 m
D1	13,27%	176	7,89 m	10,48%	176	3,5 m	10,61%	176	41 m
D2	11,47%	173	7,34 m	7,26%	176	1,5 m	7,07%	176	1h 40 m
D3	12,34%	175	7,67 m	9,23%	176	2 m	9,25%	176	54 m
E1	14,19%	174	16,98 m	-	-	-	11,34%	176	3 h

4 Experiments Results

The measures used in the comparison were Error Rate, number of distinct haplotypes inferred, and computational time used to solve each instance. Algorithms Haplorec and fastPHASE were taken from their respective sites in the web.

4.1 Comparison Using Metrics

Considering the qualitative measures, Haplorec and fastPHASE had similar performances. In datasets A and B, fastPHASE had lower Error Rate, this demonstrates the capacity of this method to deal well with conserved sequences. Haplorec was the best in datasets C and D (except for D2), but failed to resolve dataset E1. While fastPHASE took 54 minutes to find a solution with Error Rate 9,25%, Haplorec needed only 2 minutes to find a solution with Error Rate 9,23%. These two methods had similar Error Rate in all tests; the difference between them was not superior to 2% in each instance, however fastPHASE was more expensive computationally: while Haplorec took only seconds or a few minutes to resolve an instance, fastPHASE would take several minutes or hours to do the same task. Nonetheless, Haplorec could not handle the more difficult dataset, E1.

The results are shown in Table 2. For each experiment it is given the execution time, the number of distinct haplotype inferred, the error rate and switch error attained. Method PHASE was not included in the experiments because it failed to resolve even the smallest dataset (A1) in reasonable time. The Haplorec method failed to resolve dataset E1, returning several errors for this instance of the problem.

The experiments show the superior accuracy of fastPHASE and Haplorec considering Error Rate, but they both require a considerable computational time.

Table 3. Switch Error Results

Set	PTG	Haplorec	fastPHASE
A1	0,994	0,999	1
A2	0,856	0,980	0,986
A3	0,623	0,918	0,917
B1	0,938	0,991	0,996
B2	0,714	0,980	0,982
B3	0,770	0,956	0,960
C1	0,856	0,983	0,987
C2	0,575	0,973	0,979
C3	0,623	0,952	0,953
D1	0,521	0,937	0,938
D2	0,599	0,977	0,982
D3	0,558	0,958	0,961
E1	0,535	-	0,959

Although the PTG method had worst accuracy in all tests, it was never by a large difference. On the other hand, the computational costs were much lower than more accurate method fastPHASE. It is important to highlight that, for the hardest dataset, E1, while fastPHASE took 3 hours and Haplorec failed to deliver a result, PTG finished in 17 minutes with a result only 3 percentage points worse than that of fastPHASE.

The results considering Switch Error (Table 3) show that Haplorec and fastPHASE have a similar performance in this measure too. The highest difference among these methods was 0,006 (dataset A2), while the performance of PTG was really poor considering this metric. The standard deviation was also very close for Haplorec and fastPHASE, 0,023 and 0,024, respectively. Although, fastPHASE and Haplorec had close performances, these numbers do not tell if the errors in the haplotypes inferred occur in the same sequence positions or not; we investigated further on the aspect of superposition of the sites wrongly resolved.

4.2 Considering Error Sites

The localization of errors in each method is very important for suggesting the type of sequences that are more subject to errors in a given method. So we mapped by individual and by SNP the positions where each method went wrong, as considered in Error Rate. We built an error map for the results of each dataset with each method. We verified that all PTG error maps follow a very scattered pattern over all datasets, which is expected, since it does not consider any type of neighboring or statistical information. On the other hand, fastPHASE and Haplorec both consider statistical factors (although different ones), based on biological principals and they seem to be sensitive to surrounding signals. We plotted the error maps superposed for each dataset for these two methods, and we observed that the different statistics used lead to errors mostly in different positions.

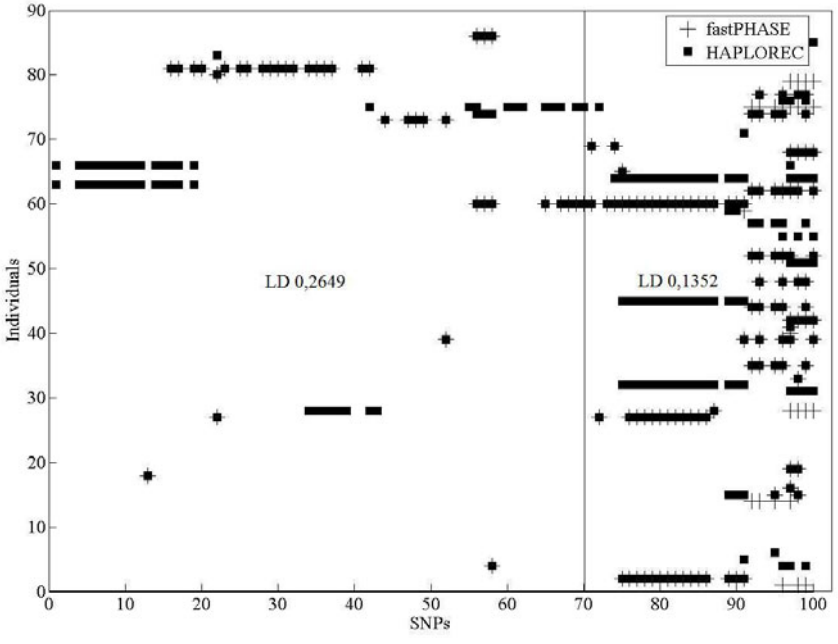


Fig. 1. Error Map for dataset A2: in axis X is presents 100 SNPs and in axis Y the 88 individual of this sample

Due to space limitation, we present only one of the superposed maps. In Figure 1 we present the superposed error maps of fastPHASE and Haplorec for Dataset A2. It is possible to define two main areas of different error concentration. The bigger area, which we will call Area 1, contains the first 70 SNPs, and the smaller area, which we will call Area 2, contains the final 30 SNPs. For both methods, Area 1 contains a very low population of error, whereas Area 2 presents errors much more frequently.

We used Linkage Disequilibrium (LD) to compute how conserved each one of these two regions is. When the sequence is highly conserved, the value of measure LD tends to 1 (maximum value), while it tends to 0 when the individuals' samples do not share sequence among them. Still considering dataset A2, Area 1 presents LD = 0,2649 and Area 2 has LD = 0,1352. In general, for the more conserved datasets, in the more conserved regions fastPHASE and Haplorec both had superior performance. That same type of behavior was observed in the error maps of all the other datasets.

There is an inverse relationship between Error Rate and LD, which is depicted in Figure 2. In SNPs with low LD (least conserved), fastPHASE and Haplorec did more inference mistakes. Regions with high LD are prone to the occurrence of homozygous sites, so a hypothetical explanation for this behavior would be the absence of heterozygous sites, however, Areas 1 and 2 had the same abundance of ambiguous sites: 30% of total sites in each area. A careful study of the curves

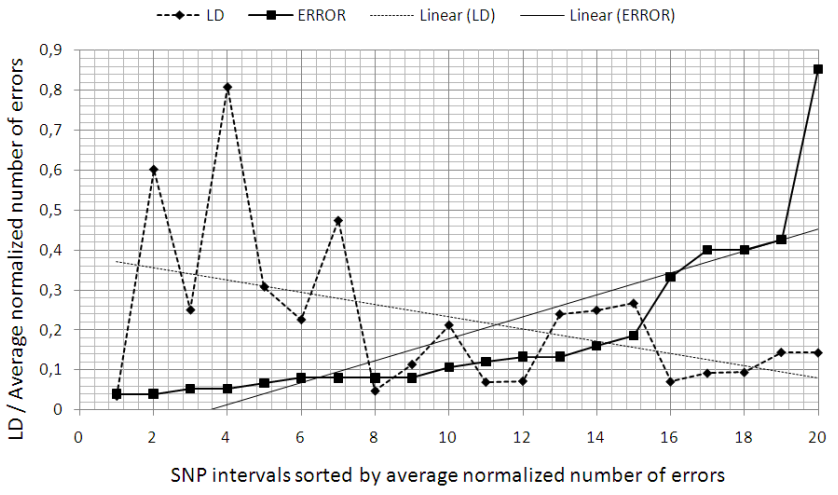


Fig. 2. Relationship between LD and error abundance in dataset A2: In axis X the SNPs are grouped in subsets of 5 neighbors; in axis Y the averages of LD and #errors, where the error abundance was normalized to the interval [0, 1] and then averaged

in Figure 2 leads to the conclusion that LD is not the only factor involved in the Error Rate, although it is certainly a relevant one.

Differently, for the PTG method we could not identify specific regions where the method made more errors; the errors occurred in a more scattered fashion along the sequences. However, we observed that genotype fragments that have original haplotype patterns as a sequence of one repeated symbol (such as $h_1 = 0^k$ and $h_2 = 1^k$, for some value k, for instance), are responsible for most errors observed in PTG. Basically, the errors in method PTG can be explained by its random behavior, and in most cases, there is no relationship with data distribution (genotypes).

The sequence positions where all methods failed to solve are somewhat difficult to explain, because they represent exceptions to the biological principles used to guide the methods. No method in the literature seems to be able to foresee or deal with those cases based only on genotype data. Fortunately, we observed that those sites are relatively few. In our experiments, most of the sites have been correctly explained by at least one of the three methods. In Figure 3, we show four diagrams, each one with the average number of genotype sites inferred incorrectly in datasets A, B, C and D. As can be seen in Figure 3, the average number of errors common to all three methods when dealing with dataset A is given by (fastPHASE(A) $\cap$ Haplorec(A) $\cap$ PTG(A)) = 77, when the total number of sites in A was 962, that is, only 8%. In a similar way, the ratio of errors common to all three methods in datasets B, C, and D, were 5%, 6%, and 8%, respectively.

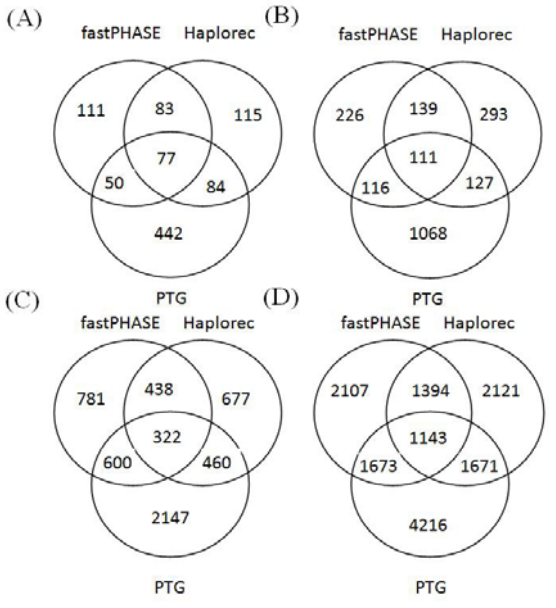


Fig. 3. Error intersection: error average quantity in genotype sites of datasets A, B, C, and D

Another interesting observation is that, from the point of view of the algorithms, as the length of the sequences grows, the number of errors of each algorithm also grows, but not at the same rate.

5 Discussion and Conclusion

We have done an extensive analysis of the performance of methods PTG, Haplorec, and fastPHASE when applied to 13 large genotype datasets, with different lengths and conservation levels. The results of our analysis offer valuable insights on the behavior of those methods.

In general, we found that fastPHASE and Haplorec have somewhat similar accuracy, as indicated be Error Rate, both being more accurate than PTG, although not by a large margin. Interestingly, although fastPHASE is computationally more expensive than Haplorec, the latter may fail to resolve long sequences. While Haplorec did not resolve our longest set of sequences (1600 SNPs), PTG finished within 17 minutes, with Error Rate 14,2%, and fastPHASE finished in three hours, with Error Rate 11,3%. Haplorec failed to solve many genotype datasets with sequences of length larger than 1000 (experiments not shown).

When the Switch Error measure is considered, Haplorec and fastPHASE both achieve almost perfect scores (all above 0,9), while PTG reaches values below 0,7 in about half of the datasets. That can be explained by the fact that the

errors made by the first two methods tend to occur together, while the errors made by PTG tend to be scattered along the sequences.

Naturally, all three methods did better with smaller and more conserved sequences, but conservation seems to particularly favor fastPHASE. We plotted a map of superposed error positions, and we observed that there is a relationship between LD and Error Rate. In general, for the more conserved datasets, in the more conserved regions fastPHASE and Haplorec both had superior performance, while PTG seems not be so sensitive to sequence conservation.

We analyzed the specific sequence positions where all the methods failed, and we observed that they are relatively few, roughly 7,2% altogether, with very little deviation for each specific dataset. The sequence positions where all methods failed to solve are somewhat difficult to explain, because they represent exceptions to the biological principles used to guide the methods. No method in the literature seems to be able to foresee or deal with those cases based only on genotype data.

Finally, we also observed a strong inverse relationship between Error Rate and LD of the SNPs, although LD is not the only factor involved in the Error Rate.

We believe that the insights provided by our analysis can be used for a more effective choice of algorithms used, and can also be explored in the design of better approaches for the Haplotype Inference Problem.

Acknowledgments. RSR and KSG gratefully acknowledge the financial support of Brazilian sponsoring agencies CAPES and CNPq, respectively.

References

1. Adkins, R.M.: Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset. BMC Genetics, 5–22 (2004)
2. Xu, H., Wu, X., Spitz, M.R., Shete, S.: Comparison of haplotype inference methods using a genotypic data from unrelated individuals. International Journal of Human and Medical Genetics 58, 63–68 (2004)
3. Scheet, P., Stephens, M.: A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. American Journal of Human Genetics 78, 629–644 (2006)
4. Eronen, L., Geerts, F., Toivonen, H.: Haplorec: Efficient and accurate largescale reconstruction of haplotypes. BMC Bioinformatics 7, 542 (2006)
5. Li, Z., Zhou, W., Zhang, X.S., Chen, L.: A parsimonious tree-grow method for haplotype inference. Oxford Bioinformatics 17, 3475–3481 (2005)
6. Clark, A.: Inference of haplotypes from PCRamplified samples of diploid populations. Journal of Molecular Biology and Evolution 7, 111–122 (1990)
7. Gusfield, D.: Inference of Haplotypes from samples of diploids populations: Complexity and algorithms. Journal of Computational Biology 8, 305–323 (2001)
8. Gusfield, D.: Haplotype Inference by Pure Parsimony. In: Baeza-Yates, R., Chávez, E., Crochemore, M. (eds.) CPM 2003. LNCS, vol. 2676, pp. 144–155. Springer, Heidelberg (2003)

9. Lancia, G., Pinotti, C.M., Rizzi, R.: Haplotype Haplotyping Populations by Pure Parsimony: Complexity of Exact and Approximation Algorithms. INFORMS J. Computing 16, 348–359 (2004)
10. Halldrsson, B.V., Bafna, V., Edwards, N., Lippert, R., Yooseph, S., Istrail, S.: A Survey of Computational Methods for Determining Haplotypes. In: Istrail, S., Waterman, M.S., Clark, A. (eds.) DIMACS/RECOMB Satellite Workshop 2002. LNCS (LNBI), vol. 2983, pp. 26–47. Springer, Heidelberg (2004)
11. Brown, D.G., Harrower, I.M.: Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. IEEE/ACM Trans. Comput. Biol. Bioinform. 3, 141–154 (2006)
12. Gusfield, D.: Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. In: RECOMB, pp. 166–175 (2002)
13. Gusfield, D.Z., Filkov, V.: A Linear-Time Algorithm for the Perfect Phylogeny Haplotyping (PPH) Problem. Journal of Computational Biology 13, 522–553 (2006)
14. Stephens, M., Smith, N.J., Donnelly, P.: A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics 68, 978–989 (2001)
15. Niu, T., Qin, Z.S., Xu, X., Liu, J.S.: Bayesian haplotype inference for multiple linked singlenucleotide polymorphism. American Journal of Human Genetics 70, 157–169 (2002)
16. Sun, S., Greenwood, C.M.T., Neal, R.M.: Haplotype inference using a Bayesian hidden Markov model. Genetic Epidemiology 31, 937–948 (2007)
17. Wu, L.Y., Zang, J.H., Chan, R.: Improved approach for haplotype inference based on Markov chain. Lecture Notes in Operations Research 9, 204–215 (2008)
18. Wang, R.S., Zhang, X.S., Sheng, L.: Haplotype inference by pure parsimony via genetic algorithm. Lecture Notes in Operations Research 5, 308–318 (2005)
19. Che, D., Tang, H., Song, Y.: Haplotype inference using a genetic algorithm. In: CICB, pp. 31–37 (2009)
20. Eronen, L., Geerts, F., Toivonen, H.: A markov chain approach to reconstruction of long haplotypes. In: Pac. Symp. Biocomput, pp.104–115 (2004)
21. Zhang, J.H., Wu, L.Y., Chen, J., Zhang, X.S.: A fast haplotype inference method for large population genotype data. Computational Statistics & Data Analysis 52, 4891–4902 (2008)
22. Stephens, M., Donnelly, P.: A comparison of Bayesian methods for haplotype reconstruction from population genotype data. American Journal of Human Genetics 73, 1162–1169 (2003)
23. The International HapMap Consortium: The International HapMap Consortium. Nature 426, 789–796 (2003)
24. Lin, S., Cutler, D.J., Zwick, M.E., Chakravarti, A.: Haplotype Inference in Random Population Samples. American Journal of Human Genetics 71, 1129–1137 (2002)

An SVM Model Based on Physicochemical Properties to Predict Antimicrobial Activity from Protein Sequences with Cysteine Knot Motifs

William F. Porto, Fabiano C. Fernandes, and Octávio L. Franco

Centro de Análises Proteômicas e Bioquímicas, Pós-Graduação em Ciências Genômicas e Biotecnologia, Universidade Católica de Brasília, Brasília - DF, Brazil, 70790-160
{williamfp7,fabianofernandesdf,ocfranco}@gmail.com

Abstract. The cysteine knot motifs are widely spread in several classes of peptides including those with antimicrobial functions. These motifs offer a major stability to the protein structure. Nevertheless, the antimicrobial activity is modulated by physicochemical properties. In this paper, we create a model of support vector machine to predict antimicrobial activity from sequences with similar motifs, based on physicochemical properties: net charge, ratio between hydrophobic and charged residues, average hydrophobicity and hydrophobic moment. The support vector machine model was trained with 146 antimicrobial peptides with six cysteines from the antimicrobial peptides database and an equal number of random sequences predicted as transmembrane proteins. The polynomial kernel shows the best accuracy (77.4%) on 10-fold cross validation. Testing in a blind dataset, we observe an accuracy of 83.02%. Through this model, proteins of varied size with a cysteine knot motif can be predicted with good reliability.

Keywords: Support Vector Machine, Antimicrobial Peptides, Physicochemical Properties, Cysteine Knot Motif, Machine Learning.

1 Introduction

In the last decades the conventional antibiotics have decreased their activity against the pathogenic bacteria, due to resistance development [1]. The search for novel antimicrobial peptides (AMP) is increasing, since the antimicrobial peptides appear as an alternative to control those pathogens due to their short length and their fast and efficient action [1-3].

AMPs are a very diverse and abundant group, being divided in several classes, but some of them share some physicochemical characteristics, some studies propose that they are non essential, but the determinant for activity is the tertiary structure [4], however, other studies show that the physicochemical characteristics modulate the antimicrobial activity [5]. Some classes of AMP have a special structural feature, a cysteine knot motif with three disulfide linkages in their structures, which provides a major stability to the structure. Among them are the defensins, which could be found in animals, plants and fungi, showing three, four or five disulfide linkages [1, 6-7].

C.E. Ferreira, S. Miyano, and P.F. Stadler (Eds.): BSB 2010, LNBI 6268, pp. 59–62, 2010.

Another class that contains the cysteine knot motif is the class of plant cyclotides, which comprises macrocyclic peptides with a head-to-tail cyclised backbone [8]. Physicochemical characteristics combined with cysteine knot motif can be used to predict antimicrobial activity of protein sequences. The present study was conducted in order to create a model of Support Vector Machine (SVM) to predict antimicrobial activity.

2 Material and Methods

Starting from the Antimicrobial Peptides Database [3], 207 sequences with six cysteine residues were extracted and at least one repetition of the pattern CXC, where X is any of the 20 natural amino acids. The redundant sequences were removed with Jalview [9], remaining 199 in the positive dataset. The negative dataset was composed of an equal number of random protein sequences predicted as transmembrane by Phobius [10], totaling 398 sequences in main dataset. For each sequence, four physicochemical properties were calculated: net charge at the physiological pH, ratio between hydrophobic and charged residues, average hydrophobicity (H) and the hydrophobic moment (μH). H and μH were measured based on Eisenberg's hydrophobicity scale [11]. Moreover, μH was given by the Eisenberg's equation [11].

Before physicochemical characterizations, the main dataset was divided in two datasets, training and blind dataset. The training dataset was composed of 146 AMPs with redundancy minor than 90% and 146 sequences predicted as transmembrane randomly selected. The blind dataset was composed by the 106 remaining sequences of main dataset. The model of SVM was developed with SVM Perl Module, available in the Comprehensive Perl Archive Network (CPAN) [12]. The kernel function has been chosen according to the results of the 10-fold cross validation, and the performance of the SVM was measured by the following parameters:

$$\text{Sensitivity} = \{TP/(TP+FN)\}*100 \quad (1)$$

$$\text{Specificity} = \{TN/(TN+FP)\}*100 \quad (2)$$

$$\text{Accuracy} = \{(TP+TN)/(TP+TN+FN+FP)\}*100 \quad (3)$$

Where TP is the number of true positives; FN, the false negatives; TN, the true negative; and FP, the false positives.

3 Results and Discussion

Cysteine knot motifs are widely spread in several classes of peptides, such AMPs [8]. These motifs offer a major stability to the protein structure. Nevertheless, the antimicrobial activity is modulated by physicochemical properties, such hydrophobicity,

cationicity and amphipathicity [1, 5]. We propose that the combination of these features can be used to predict antimicrobial activity through supervised machine learning. On the other hand, the use of supervised machine learning to predict antimicrobial activity has two major challenges, (i) the size variation of sequences and (ii) the absence of a non-antimicrobial data base [2]. In this paper, a SVM model to predict antimicrobial activity from peptides with a cysteine knot motif through physicochemical properties was developed, which turn possible the prediction of sequences, independently of size, solving the first challenge, on the other hand, this approach generates a novel problem: shuffled sequences have the same scalar physicochemical properties, generating false positives. However, we included the μH to solve this problem, once μH is not a scalar property and its value depends of the sequence of amino acids in protein. To solve the second challenge, we use the Phobius prediction to select transmembrane proteins, since these proteins have no antimicrobial activity; due to those are no secreted proteins [2].

The adequate kernel function was selected based on a 10-fold cross validation on training dataset. The kernel polynomial has the best accuracy (77.4%) than radial (75.68%), linear (74.65%) and sigmoid (46.91%). Figure 1 shows the accuracy in several K-fold cross validations. Testing the model on blind dataset, we observe a sensitivity of 75.36%, a specificity of 97.3% and an accuracy of 83.02%. Through this model, proteins of varied size with a cysteine knot motif can be predicted with good reliability and can be used to discover and design novel drugs.

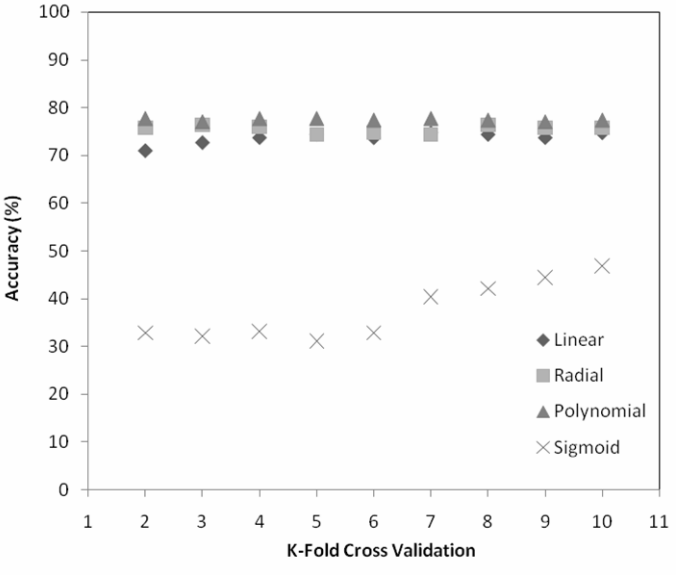


Fig. 1. K-fold cross validation of SVM Model. The kernels linear, radial and polynomial display a similar accuracy, but the polynomial kernel shows the best accuracy in majority of k-fold cross validations.

4 Conclusion

This paper shows that the combination of cysteine knot motif and physicochemical properties can be used to predict antimicrobial activity with good reliability, based on a SVM model, independently of the protein size. This SVM model can be relevant to reveal the antimicrobial activity from proteins which posses other functions, like protease inhibitors, conotoxins or metallothioneins, for example, since these proteins have the cysteine knot motif. In future studies, more physicochemical properties will be added to SVM model in order to improve the accuracy of the model and sequences without described antimicrobial activity will be predicted and tested *in vitro*.

Acknowledgments. This work was supported by FAPDF and UCB.

References

1. Brogden, K.A.: Antimicrobial Peptides: Pore Formers or Metabolic Inhibitors in Bacteria? Nature Microbiology 3, 238–250 (2005)
2. Lata, S., Sharma, B.K., Raghava, G.P.S.: Analysis and Prediction of Antibacterial Peptides. BMC Bioinformatics 8, 263 (2007)
3. Wang, Z., Wang, G.: APD: the Antimicrobial Peptide Database. Nucl. Acids Res. 32, D590–D592 (2004)
4. Loose, C., Jensen, K., Rigoutsos, I., Sphanopoulos, G.: A Linguistic Model for the Rational Design of Antimicrobial Peptides. Nature 443(19), 867–869 (2006)
5. Dathe, M., Wieprecht, T., Nikolenko, H., Handel, L., Maloy, W.L., MacDonald, D.L., Beyermann, M., Bienert, M.: Hydrophobicity, hydrophobic moment and angle subtended by charged residues modulate antibacterial and hemolytic activity of amphipathic helical peptides. FEBS Letters 403, 208–212 (1997)
6. Mygind, P.H., Fischer, R.L., Schnorr, K.M., Hansen, M.T., Sönksen, C.P., Ludvignsen, S., Raventós, D., Buskov, S., Christensen, B., De Maria, L., Taboureau, O., Yaver, D., Elvig-Jørgensen, S.G., Sørensen, M.V., Christensen, B.E., Kjærulff, S., Frimodt-Moller, N., Lehrer, R.I., Zasloff, M., Kristensen, H.H.: Plectasin is a peptide antibiotic with therapeutic potential from a saprophytic fungus. Nature 437(13), 975–980 (2005)
7. Ganz, T.: Defensins: Antimicrobial Peptides of Innate Immunity. Nature Immunology 3, 710–720 (2003)
8. Craik, D.J.: Plant Cyclotides: Circular, Knotted Peptide Toxins. Toxicon 39, 1809–1813 (2001)
9. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J.: Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9), 1189–1191 (2009)
10. Käll, L., Krogh, A., Sonnhammer, E.L.L.: Advantages of Combined Transmembrane Topology and Signal Peptide Prediction–the Phobius web server. Nucl. Acids Res. 35, W429–W432 (2007)
11. Eisenberg, D., Weiss, R.M., Terwilliger, T.C.: The Hydrophobic Moment Detects Periodicity in Protein Hydrophobicity. Proc. Natl. Acad. Sci. 81, 140–144 (1984)
12. Comprehensive Perl Archive Network, `http://search.cpan.org/~lairdm/Algorithm-SVM-0.11/lib/Algorithm/SVM.pm`

Enabling Annotation Provenance in Bioinformatics Workflow Applications

Milene Pereira Guimarães and Maria Cláudia Cavalcanti

Seção de Engenharia de Computação, Instituto Militar de Engenharia (IME),
Praça General Tiburcio n.80, Praia Vermelha, Urca,
22290-270, Rio de Janeiro, RJ, Brazil
milenepgj@gmail.com, yoko@ime.eb.br

Abstract. Nowadays, the resulting data of each step of genomic annotation processes are typically used for annotation provenance. However, recent initiatives in the direction of capturing annotation provenance data do not support the Bioinformatics developer on building their own provenance enabled applications. This work proposes the ArCaP architecture, which aims at supporting the development of provenance enabled applications and at facilitating the access to annotation provenance data. Although the focus is on Bioinformatics applications, this approach is useful to other scientific domains.

Keywords: Data provenance; Annotation Provenance; Scientific Workflows.

1 Introduction

Typically, genomic research centers such as the Oswaldo Cruz Institute [8], use computational tools for processing, analyzing, visualizing and storing genomic data. Some of these tools are embedded in applications also known as genomic annotation systems, which support the whole genomic research process, i.e., since the sequence cleaning, up to the annotation itself. As these systems have been intensely used, there has been an exponential growth of data [2]. More recently, to facilitate data analysis, these systems started to use Database Management Systems (DBMS).

Besides the data generated at the end of a genomic research process, it is also required for genomic research projects to maintain provenance data. A genomic annotation is more valuable as more provenance data are available for the user. Especially in the context of data curation activities, provenance data is a way of addressing data trust issues [6]. Buneman et al. [3] say that provenance information concerning the creation, attribution, or version history of data is crucial for assessing its integrity and scientific value. Moreover, data curation is a long and expensive process [5]. Therefore, to add value to a genomic annotation and to fasten its curation, it is necessary to capture data about annotation provenance and associate them.

This work aims at facilitating the development of provenance enabled Bioinformatics workflow applications, by providing a user guide for the creation of provenance data structures. In this direction we propose a development architecture through which the developer can build his application, enabling it to register

C.E. Ferreira, S. Miyano, and P.F. Stadler (Eds.): BSB 2010, LNBI 6268, pp. 63–66, 2010.

provenance data, which are then made available for user queries. For instance, in a typical genomic process it would be possible to retrieve which sequence fragment is associated to a sequence annotation.

2 Related Work

Application-specific databases do not address data provenance in a uniform and generic way. There are two main initiatives in the direction of providing a more generic provenance registry: metamodels and Scientific Workflow Management Systems (SWfMS). A metamodel is at a higher level of abstraction than a model. It is often called "a model of a model". A provenance metamodel may be useful as it may guide specific application developers on instantiating provenance enabled models, i.e., models that include data structures to store provenance data. Some provenance driven metamodels had been proposed [4][10][11]. However, none of these metamodels are proposed in the context of development support architecture. The developer may build his/her application based or not on such metamodels.

On the other hand, SWfMS are also useful in providing data provenance. A SWfMS is a set of tools that support the definition and execution scientific workflows (sets of tasks), in a coordinated and integrated way, in order to automatically generate and/or derivate new scientific data. Through the use of a SWfMS, specific scientific workflow applications may be developed and used. There are SWfMS that already provide some support for data provenance [9] [1] [7] by using a “behind the scene” model which is able to capture some data provenance. However, none of these SWfMS provide a uniform way to capture provenance, as they use their own specific model for that purpose. Furthermore, the user becomes a hostage of these systems interface, as his/her application depends on these systems to run. Therefore, there is still a need for a generic and uniform development platform that could guide the bioinformatics developer in powering his/her application with provenance capture.

3 ArCaP

In order to address this issue we propose the Architecture for Provenance Capture (ArCaP), which allows the specification of a workflow for the registry and subsequent retrieval of data provenance, based on a provenance metamodel [12]. ArCaP main objective is to support a clear and intuitive development of tools that capture provenance, and also to enable researchers to query about data provenance in a more direct and granular way. Two major types of user are identified in ArCaP: the developer and the researcher. The developer role is related to the user that will specify and develop a workflow application and the researcher role is related to who actually executes a workflow application (e.g. an annotator or a curator).

ArCaP architecture is presented in Fig. 1, and its functionality is summarized as follows. Initially, a developer interacts with the **Workflow Specification** module, in order to register in the **Metamodel Database** the tasks, programs and data (programs input and outputs) that take part on the workflow application. The focus of this

metamodel is on the intermediate data storage, i.e., data that were produced at each step of the process. It includes information related to sources of consumed data, produced data, data derivation, processes and users.

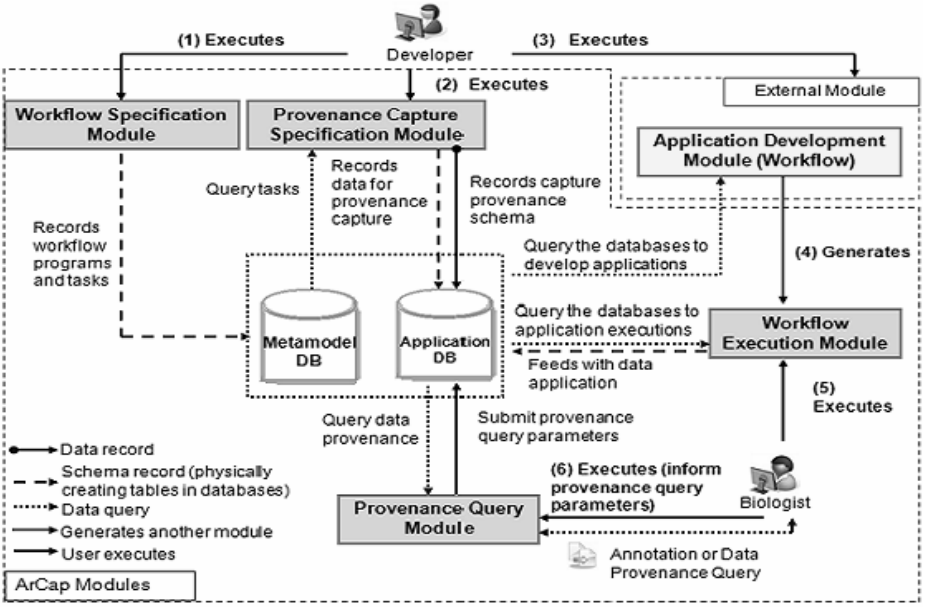


Fig. 1. ArCaP Architecture

Then, the same developer can interact with the **Provenance Capture Specification** module that guides the user on the definition of auxiliary database structures (tables) for registering provenance at the points of the workflow where it is required. The **Metamodel Database** includes the application data schema and the metadata about the tasks and programs of the specified workflow. Based on the data schema, the **Provenance Capture Specification** module is able to create the **Application Database**.

Now, the developer can initiate the development or use of the specified application through the **Application Development** module. This module may be viewed as an external module which complements the architecture, as it can be implemented by any development environment. Based on the data stored in the **Metamodel Database** and on the schema of the application which was used for the creation of the **Application Database**, the developer can build the workflow application, which corresponds to the **Workflow Execution** module. The researcher is now able to interact with this module. As the module is enabled for data provenance capture, the user actions are monitored and provenance data are stored in the **Application Database**. For instance, for each program execution of the workflow, the user and his/her role are associated, and this information is stored in specific tables and provenance structures created by the previous module in the **Application Database.**

Finally, a separate module for **Provenance Query** is also provided, as the interface for obtaining provenance data is not the focus of the application development.

Provenance queries are carried out according to the parameters informed by the researcher, including the provenance granularity required. For example, given an item of some produced data, it is possible to obtain all the data items that were consumed, throughout each step of the executed process (workflow application), to produce it.

4 Conclusion

This work presented the ArCaP architecture, a new approach for data provenance capture based in Bioinformatics scenarios. A prototype of such architecture was implemented based on a real genomic annotation system [13]. Future works include the creation of a complete case study on developing a new Bioinformatics application. Also, we plan to adapt an existing SWfMS to incorporate ArCaP ideas.

Acknowledgments

This work thanks CAPES and CNPq for their financial support and the staff of the BCS Lab at Fiocruz/IOC, for their contribution as Bioinformatics specialists.

References

1. Altintas, I., Barney, O., Jaeger-Frank, E.: Provenance collection support in the kepler scientific workflow system. Journal Provenance and Annotation of Data (2008)
2. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L.: GenBank. Nucleic Acids Research 36, D25–D30 (2008) doi:10.1093/nar/gkm929
3. Buneman, P., Chapman, A., Cheney, J.: Provenance management in curated databases. In: Proc. of the 2006 ACM SIGMOD Int. Conf. on Management of Data, SIGMOD '06, Chicago, IL, USA, June 27-29, pp. 539–550. ACM, New York (2006)
4. Cavalcanti, M.C., Mattoso, M.L., Campos, M.L.M.: Scientific resources management: Towards an in Silico Laboratory. Tech. Rep., UFRJ (2003)
5. Baumgartner Jr., W.A., Cohen, K.B., Fox, L.M., Acquaah-Mensah, G., Hunter, L.: Manual curation is not sufficient for annotation of genomic databases. Bioinformatics 23(13), 41–48 (2007)
6. Lord, P., Macdonald, A., Lyon, L., Giaretta, D.: From Data Deluge to Data Curation. The Digital Archiving Consultancy Limited and the Digital Curation Centre (2004)
7. Menezes, J.G.M.: Gerência distribuida de dados em workflows de bioinformática. Master Thesis, Military Institute of Engineering - IME (2008)
8. Oswaldo Cruz Institute/Fiocruz, `http://www.fiocruz.br`
9. Scheidegger, C., Koop, D., Santos, E., Vo, H., Callahan, S., Freire, J., Silva, C.: Tackling the provenance challenge one layer at a time. Concurrency and Computation: Practice and Experience (2007)
10. Simmhan, Y., Plale, B., Gannon, D.: Karma2: Provenance management for data driven workflows. Int. Journal of Web Services Research 5(1) (2008)
11. Moreau, L., Freire, J., Futrelle, J., Mcgrath, R., Myers, J., Paulson, P.: The open provenance model. Technical report, University of Southampton (2007)
12. Guimarães, M.P.: Uma abordagem para capturar a proveniência de dados na área de Bioinformática. Master Thesis, Military Institute of Engineering - IME (2009)
13. System for Integrate Genomic Resources and Analysis, `http://stingray.biowebdb.org`

BAT: A New Biclustering Analysis Toolbox

Cristian Andrés Gallo[1], Julieta Sol Dussaut[1],
Jessica Andrea Carballido[1], and Ignacio Ponzoni[1,2]

[1] Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC),
Departamento de Ciencias e Ingeniería de la Computación,
Universidad Nacional del Sur, Av. Alem 1253, 8000, Bahía Blanca, Argentina
{cag,jac,ip}@cs.uns.edu.ar
[2] Planta Piloto de Ingeniería Química (PLAPIQUI) - UNS - CONICET Complejo
CRIBABB, Co. La Carrindanga km.7, CC 717, Bahía Blanca, Argentina

Abstract. In this paper, a new biclustering analysis toolbox called BAT, which is based on the BiHEA (Biclustering via a Hybrid Evolutionary Algorithm), is presented. The BiHEA is a memetic approach that integrates a Multi-Objective Evolutionary Algorithm (MOEA) with a local search technique in order to perform microarray biclustering. This method simultaneously considers several goals for optimization, giving as a result a set of biclusters that present a satisfactory trade-off between all of them. The novel software introduced in this article provides the possibility of running the BiHEA along with several pre-processing facilities for the input data and different visualization and statistical tools for the analysis of the biclusters.

Keywords: microarray analysis, biclustering, multi-objective evolutionary computing, software toolbox.

1 Introduction

Microarray technology constitutes one of the most widely used methods that contribute to generate an amazing amount of biological data. Numerous research teams now posses the skills to generate microarray data, whereas in many cases it is still limited their capacity to extract biologically meaningful information from these data. In this sense, the recognition of gene groups with coherent expression values represents a key step in the analysis of gene expression data. Traditional clustering algorithms partition the expression matrix into sub-matrices that extend over the whole set of samples. However, in most cases, the assumption that all genes behave similarly in all conditions may be too restrictive.

To account for this, biclustering approaches carry out the grouping in both dimensions simultaneously: genes and samples [1, 2]. This allows to find subgroups of genes that show the same response under a subset of conditions, e.g. if a cellular process is only active under these conditions. Furthermore, if a gene participates in multiple pathways that are differentially regulated, one would expect this gene to be included in more than one cluster; this cannot be achieved by traditional clustering.

C.E. Ferreira, S. Miyano, and P.F. Stadler (Eds.): BSB 2010, LNBI 6268, pp. 67–70, 2010.

Several biclustering algorithms have been proposed in the literature [3–6]. These techniques vary from simple greedy approaches to complex stochastic evolutionary algorithms. However, in general, the software that implements them is difficult to use or it is not accessible at all. Therefore, our motivation in this work is to provide a biclustering analysis toolbox in order to accomplish the usability of a novel biclustering algorithm recently published [7].

2 Main Toolbox Features

In this work, a user-friendly software toolbox called BAT (Biclustering Analysis Toolbox), which implements the BiHEA algorithm [7] together with several means for visualization and biclustering analysis, is presented. The BiHEA is a memetic evolutionary algorithm for biclustering of Gene Expression Data (GED) that considers the mean squared residue (homogeneity) [3], the row variance and the size of the bicluster as objectives to be optimized during the construction of the solutions. The main facilities of the BAT can be summarized as follows:

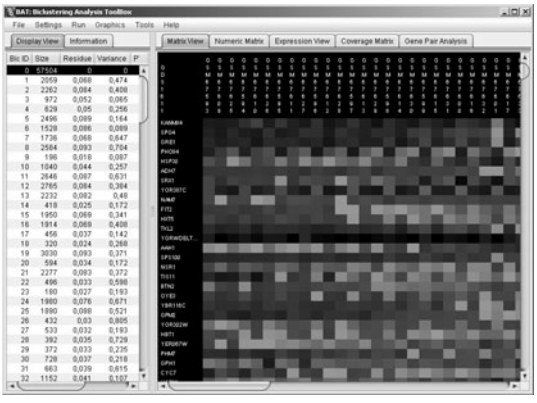
Data handling. The information, consisting on the entire expression matrix and the set of biclusters extracted by the BiHEA, is organized in a list structure that is depicted in the left panel of the graphical user interface (Fig. 1a). This structure allows accessing the dataset and the results for pre-viewing in several forms. Additionally, it also provides information about the size, mean squared residue and row variance of the biclusters.

Considering the stochastic nature of the evolutionary algorithms, the software is capable of performing several runs sequentially, each one with a different seed. In this case, the column in the left panel called trials indicates the amount of runs in which each resulting bicluster has appeared.

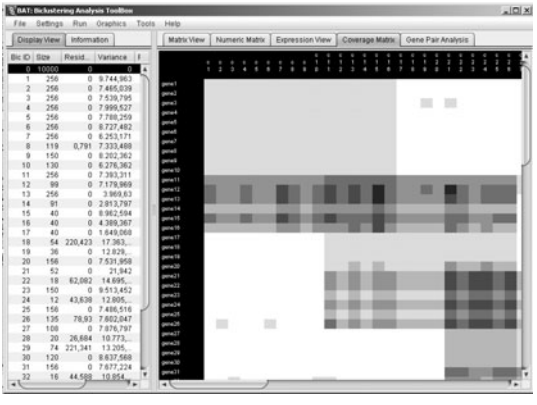
Pre-processing. The input data file can be any CSV file, including annotations of genes and conditions, or a previously saved project. The loaded gene expression data can then be transformed by means of several methods, including normalization (log2) and standardization. Additionally, BAT automatically infers the missing values of the GED employing the BPCA method [8].

Visualization. The GED can be displayed in three different ways: as a heatmap (Fig. 1a), as a numerical matrix or in terms of the coverage of the resulting biclusters (Fig. 1b). The annotations of the conditions run along the top whereas the annotations of the genes are listed on the left hand side. Additionally, the biclusters can be visualized in the form of a heatmap, a numerical matrix, or as a collection of expression profiles (Fig. 1c). The expression profiles display the behavior of those genes that are grouped within a bicluster in which, for each gene, a colored line connects the expression values for the different samples.

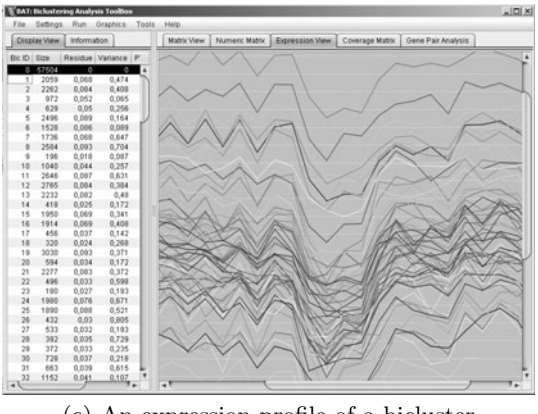
Post-processing. In order to analyze the results, the software offers the possibility of performing a pair-wise gene analysis. More specifically, for each pair of genes, the frequency with which those genes occur together in the same bicluster is calculated. This number of co-occurrence detects those genes that may be functionally related. Additionally, as we mentioned earlier, the degree of coverage of



(a) A heatmap visualization of the GED



(b) The coverage of the biclusters in the GED



(c) An expression profile of a bicluster

Fig. 1. Graphical user interface of the BiHEA software

the resulting biclusters with respect to the gene expression matrix can be visualized as a grey scale matrix (see Fig. 1b). This is useful so as to provide a global view of the areas in the GED in which the biclusters are placed.

The figures, graphs, and results can be exported from the BAT for further usage. For the figures and graphs, it is also possible to adjust the resolution in dpi of the images accordingly to the needs of the user. All the work performed during a session of the BAT can be saved as a project and restored later.

3 Final Considerations and Discussion

In this paper, we have introduced the BAT: a software tool for bliclustering of gene expression data that implements the BiHEA algorithm. As it frequently occurs in bioinformatics, when a new algorithm emerges, the software that implements it is difficult to be found and used, whenever it is available at all. This work is mainly focused on this issue, providing a complete framework in which the biologist can rely in order to perform analysis of gene expression data through a novel biclustering algorithm recently published.

The software, source code, manuals and several examples are freely available at http://lidecc.cs.uns.edu.ar with the aim of offering all the support needed by the user. Additionally, the implementation is OS-independent, and therefore it will work on near all operating systems.

Acknowledgments. This work is kindly supported by CONICET grant PIP 11220090100322 and UNS grant PGI 24/ZN15.

References

1. Madeira, S., Oliveira, A.: Biclustering algorithms for biological data analysis: A survey. IEEE-ACM Trans. Comput. Biol. Bioinform. 1, 24–45 (2004)
2. Madeira, S., Oliveira, A.: A polynomial time biclustering algorithm for finding approximate expression patterns in gene expression time series. Algorithm Mol. Biol. 4(1), 8 (2009)
3. Cheng, Y., Church, G.M.: Biclustering of expression data. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, pp. 93–103. AAAI Press, Menlo Park (2000)
4. DiMaggio, P., McAllister, S., Floudas, C., Feng, X.J., Rabinowitz, J., Rabitz, H.: Biclustering via optimal re-ordering of data matrices in systems biology: rigorous methods and comparative studies. BMC Bioinformatics 9(1), 458 (2008)
5. Mitra, S., Banka, H.: Multi-objective evolutionary biclustering of gene expression data. Pattern Recogn. 39(12), 2464–2477 (2006)
6. Bleuler, S., Prelić, A., Zitzler, E.: An EA framework for biclustering of gene expression data. In: Congress on Evolutionary Computation, pp. 166–173 (2004)
7. Gallo, C.A., Carballido, J.A., Ponzoni, I.: Bihea: A hybrid evolutionary approach for microarray biclustering. In: Guimarães, K.S., Panchenko, A., Przytycka, T.M. (eds.) Advances in Bioinformatics and Computational Biology. LNCS, vol. 5676, pp. 36–47. Springer, Heidelberg (2009)
8. Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K.: A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 19(16), 2088–2096 (2003)

Detection of Protein Domains in Eukaryotic Genome Sequences

Arli A. Parikesit*, Peter F. Stadler, and Sonja J. Prohaska

Bioinformatics Group, Dept. Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany

Abstract. Large-scale studies of the origins and evolution of regulatory mechanisms require quantitative estimates of the abundance and co-occurrence of functional protein domains in the genomes of very diverse organism. Current databases, such as SUPERFAMILY, are not able to provide such quantitative data because of species-specific differences and biases in the existing transcript and protein annotations on which they are based. Here we show that the combination of *de novo* gene predictors and subsequent HMM-based annotation of SCOP domains in the predicted peptides leads to consistent estimates with acceptable accuracy.

Keywords: Protein domains, genome annotation, regulatory mechanisms.

The expression of genomically encoded information is subject to tight regulation. The regulatory rules are implemented in a highly complex network encompassing several biochemically distinct mechanisms involving specific chromatin states, the action of transcription factors, regulated mRNA export, alternative splicing, translational control, post-transcriptional and post-translational modifications, and controlled degradation of both RNA and polypeptides. Surprisingly, different clades appear to emphasize certain types of mechanisms while reducing or even abolishing others. Regulation in eubacteria, for example, appears dominated by transcription factor networks, trypanosomes use the posttranscriptional processing of large polycistronic transcripts, ciliates utilizes extensive amplification of DNA in creating their macro-nuclei, and crown group eukaryotes have evolved an elaborate system of histone modifications. The most direct approach towards an understanding of the origins and evolution of the different regulatory mechanisms is the comprehensive reconstruction of the evolutionary histories of the many protein families that are involved in these processes. In practice, however, this is an exceedingly difficult and tedious task, since homologies even between highly conserved proteins become hard to establish in comparisons across kingdoms. Proteins are composed of recognizable protein domains that implement well-defined functions such as catalytic activities, specific binding, or anchoring in membranes. Over large time-scales, these components have been combined

* Supported by the DAAD.

C.E. Ferreira, S. Miyano, and P.F. Stadler (Eds.): BSB 2010, LNBI 6268, pp. 71–74, 2010.

in a combinatorial fashion to produce new functionalities, so that individual proteins often have multiple ancestors that contributed different domains [1]. A more modest approach thus aims at tracing the distribution of protein domains in a comparative fashion. Considering the co-occurrences of domains in individual proteins, furthermore, reveals clade-specific domain combinations [2], a growing core of combinations in multicellular organisms [3].

Such studies are based on annotations extracted from databases. For instance, [2] used the protein annotation compiled in `KEGG` and `ENSEMBL` together with `Pfam` domains, [4] utilized the `SUPERFAMILY` database, whose HMM models are based on the SCOP domain definitions. The annotation of protein domains, however, is based upon an annotation of "protein models" that are constructed from the genomic DNA sequence, EST and cDNA data, and computational predictions. Large differences in EST and/or cDNA coverage, and in the computational procedures leads to substantially different domain annotations even for phylogenetically very closely related species. For instance, SUPERFAMILY 1.73 annotates 64225 domains in human, but only 45312 in chimpanzee and 21208 in gorilla and only about 14748 in the alpaca. Such biases may explain why a search for putative homologs of the most prominent proteins associated with the microRNA pathway (Drosha, Dicer, DGCR8, TRBP, and TRBP) based on co-occurrences of their known SCOP domains, much to our surprise, did not recover the phylogenetic distributions reported in detailed, homology-based studies [5]. Here, we investigate strategies to construct inventories of protein domains that largely avoid biases arising from the underlying gene annotation. The need for quantitative studies is emphasized by the realization that many protein domains are evolutionarily very old so that functional innovation is reflected by differences and domains abundances and domain combinations.

To this end, we re-annotate protein domains in three different collections of (putative) polypeptides: (1) translations of annotated transcripts, (2) the results of *de novo* gene predictors, and (3) conceptual translations of the entire genomic DNA. We analyze the genomes of three apes, human (GRCh37.57), chimp (CHIMP2.1.57), and gorilla (gorGor3.57) to identify biases among very closely related species. Genomes and transcript annotation were downloaded from `www.ensembl.org` (version 57). Gene predictions were performed using `genscan` [6] and `GeneMark` [7]. In order to save computational resources we randomly selected the Hidden Markov Models of 100 domains from the SUPERFAMILY database (1.73) [8]. The HMMs were mapped using `HMMER 3.0rc1` with $E \leq 10^{-4}$. Among overlapping hits, only the highest-scoring one was retained.

A scatter-plot of the number of domain occurrences measured on the set of annotated transcript and on the *de novo* gene predictions shows a significant correlation, Fig 1(a-c). In contrast, an attempt to estimates the domain numbers by running the HMMs on translated genomic DNA failed miserably: only a small fractions of the known domains can be recovered. This is not surprising since on average a domain contains 3 or 4 introns [9] in human.

In the human data, Fig. 1(a), the majority of domains is observed more frequently in annotated transcripts than in `genscan` predictions. This effect is less

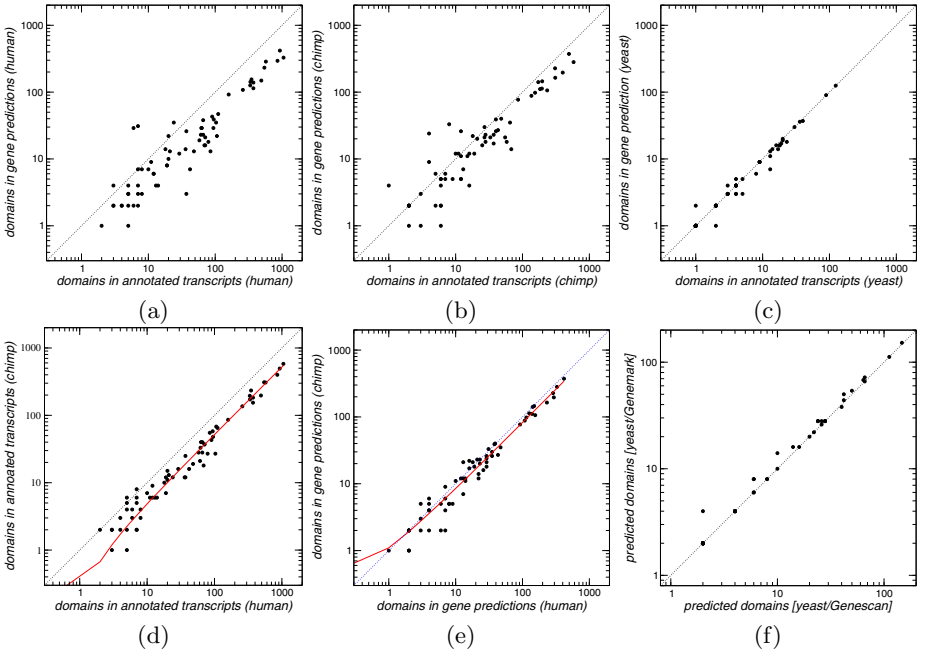


Fig. 1. Correlation of the number of protein domains. Top row: Annotated transcripts compared to *de novo* predicted "genes" for (a) human, (b) chimp, and (c) yeast. Below: While domain prediction based on existing annotation yield systematic differences between human and chimp (d), congruent abundances are obtained from `genscan` predictions (e). Linear regression is shown as red line in panels (e) and (f). Different gene predictors (`genscan` and `GeneMark`) yield comparable results (f), shown here for yeast.

pronounced in chimpanzee, Fig. 1(b). In yeast, on the other hand, the correspondence between transcript-based domain annotation and the `genscan`-based results is excellent. We can understand these differences in terms of the quality and coverage of the transcript annotation. In the human genome, we have a large number of annotated isoforms and alternative transcripts as a result of extensive cataloguing efforts, thus multiple transcripts incorporate the same genomic domain. A comparable density of data is not available for almost all other species, so we under-count the annotated transcripts of the other two ape genomes. Somewhat surprisingly, transcript annotation and `genscan` predictions agree extremely well in yeast.

The sampling bias introduced by working with transcript annotations can be demonstrated by directly comparing domain annotations between the three closely related ape species. We observe strong systematic biases when the domain annotation is based on the currently best transcript annotations for these genomes, shown for human *vs.* chimp in Fig. 1(d): Transcript-based domain annotation yields many more occurrences of domains in human than in chimp and

gorilla. Consistently, we also obtain more domains in chimp than in gorilla. This bias can be explained nearly completely by the differences in the numbers of annotated transcripts: for chimp/human, e.g., the ration of annotated transcripts is 0.446, the ratio of detected proteins domains is 0.522.

In contrast to the transcript-based data, we obtain largely consistent results from the `genscan` predictions. There is some scatter, in particular for relatively rare domains, but no strong systematic bias, Fig. 1(e). The remaining discrepancies between human and the other two apes (about 10%) can probably be explained by the quality and completeness of the current genome assemblies. Finally, we confirmed that the two gene predictors `genscan` and `GeneMark` yield congruent results at the level of predicted protein domains. Fig. 1(f) shows this for yeast. Due to the comparably long run-time of `GeneMark`, corresponding data were computed for a single human chromosome only (not shown).

We conclude that protein domains can be annotated with acceptable accuracy directly from genomic DNA sequences using a *de novo* gene predictor such as `genscan` as an intermediate.

References

1. Koonin, E., Aravind, L., Kondrashov, A.: The impact of comparative genomics on our understanding of evolution. Cell 101, 573–576 (2000)
2. Itoh, M., Nacher, J.C., Kuma, K.i., Goto, S., Kanehisa, M.: Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. Genome Biol. 8, 121 (2007)
3. Wuchty, S., Almaas, E.: Evolutionary cores of domain co-occurence networks. BMC Evol. Biol. 5, 24 (2005)
4. Prohaska, S.J., Stadler, P.F., Krakauer, D.C.: Innovation in gene regulation: The case of chromatin computation. J. Theor. Biol. 265, 27–44 (2010)
5. Murphy, D., Dancis, B., Brown, J.R.: The evolution of core proteins involved in microRNA biogenesis. BMC Evolutionary Biology 8, 92 (2008)
6. Burge, C., Karlin, S.: Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78–94 (1997)
7. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.M.B.: Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33, 6494–6506 (2005)
8. Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., Gough, J.: SUPERFAMILY — comparative genomics, datamining and sophisticated visualisation. Nucleic Acids Res. 37, D380–D386 (2009)
9. Bhasi, A., Philip, P., Manikandan, V., Senapathy, P.: ExDom: an integrated database for comparative analysis of the exon-intron structures of protein domains in eukaryotes. Nucleic Acids Res. 37, D703–D711 (2009)

Discretization of Flexible-Receptor Docking Data

Machado, K.S., Winck, A.T., Ruiz, D.D., and Norberto de Souza, O.

Programa de Pós-Graduação em Ciência da Computação, Faculdade de Informática, PUCRS,
Av. Ipiranga, 6681 – Prédio 32, Sala 602, 90619-900, Porto Alegre, RS, Brazil
{karina.machado,ana.winck,duncan.ruiz,osmar.norberto}@pucrs.br

Abstract. A careful analysis of flexible-receptor molecular docking results, particularly those related to details of receptor-ligand interactions, is essential to improve the process of docking and the understanding of intermolecular recognition. Because flexible-receptor docking simulations generate large amounts of data, their manual analysis is impractical. We intend to apply classification decision trees algorithms to better understand this type of docking results. However, prior to that we need to discretize the target attribute, which in this work is the estimated Free Energy of Binding (FEB) of the flexible receptor-ligand interactions. Here we compare three different discretization methods, by equal frequency (1), by equal width (2) and our proposed method, based on the mode and standard deviation (3) of the FEB values.

Keywords: discretization, molecular docking, flexible receptor, data mining.

1 Introduction

Molecular docking simulations is a computational method used to predict the preferred conformation and orientation of one molecule, usually called a ligand, to a second molecule, named target receptor, to form a stable complex [1]. The output file for each receptor-ligand docking simulation contains much relevant information to help understand the intermolecular recognition details underlying the formation of a stable receptor-ligand complex. The Euclidian space of the intermolecular complex and the value of the objective function which estimates the affinity of the receptor-ligand complex, called Free Energy of Bind (FEB) in the software AutoDock3.0.5 [2], are only two examples of the many types of data that results from a docking simulation. Additionally, when both, ligand and receptor molecules have their flexibility considered in the docking simulations vast amounts of data are generated [3], particularly if the flexible model of the receptor is constructed from a molecular dynamics (MD) [4] simulation trajectory [3][5].

We have previously performed docking simulations of a flexible-receptor to four different ligands [3]. The flexible receptor was modeled as a 3,100 snapshots derived from a 3.1 ns (1 ns = 10^{-9} seconds) MD simulation trajectory [6]. Consequently, this is a very computer-intensive approach. Aiming at reducing this computational demand, we are developing ways to use a smaller number of receptor snapshots without affecting the explicit flexibility of the receptor. However, prior to developing this snapshot-reducing method, a careful analysis of the molecular

C.E. Ferreira, S. Miyano, and P.F. Stadler (Eds.): BSB 2010, LNBI 6268, pp. 75–79, 2010.

docking results is essential to comprehend the details of flexible receptor-ligand interactions and their relationship with the estimated FEB of the complex. These types and amount of data can be efficiently explored by data mining techniques [7]. Hence, we intend to apply classification decision tree algorithm considering as target attribute the FEB values. However, in classification tasks, numerical target attributes must be discretized [7].

In this work we propose and describe a discretization method which uses the mode and standard deviation of the distribution of FEB values. We compare our method to other two classical discretization methods: by equal frequency and by equal width.

2 Materials and Methods

2.1 Receptor, Ligands and Molecular Dynamics and Docking Simulations

In this work we considered the crystal structure of the InhA enzyme from *Mycobacterium tuberculosis* [8] (PDB ID: 1ENY) as the rigid receptor. Starting from this structure, by means of MD simulations [4], we generated the flexible-receptor model of InhA. It is made up of 3,100 snapshots derived from a 3.1 ns MD simulation trajectory [6]. As ligands we used the pentacyano(isoniazid)ferrate(II) (PIF) [9], nicotinamide adenine dinucleotide (NADH) [8], triclosan (TCL) [10], and ethionamide (ETH) [11]. The docking simulations were executed considering the flexible InhA receptor model and each of the four ligands [3] with AutoDock3.0.5 [2] (results summarized in Table 1). The data containing the MD simulation trajectory snapshots and the related docking results were stored in the FReDD database [12].

Table 1. Results of the flexible InhA docking simulation to four different ligands. Columns 1 and 2 contain the ligand names and their total number of atoms after preparation for docking; column 3 displays the total number of valid docking results; column 4, the average and standard deviation of the estimated FEB values (in kcal/mol) and columns 5, 6, and 7 displays the minimum, maximum and the mode of the statistical distribution of FEB values.

Ligands	Atoms	Dockings	FEB	Min. FEB	Max. FEB	Mode
PIF	24	3,042	-9.9 ± 0.6	-11.2	0.0	-9.9
NADH	52	2,823	-12.9 ± 4.2	-20.6	0.0	-16.8
TCL	18	2,837	-8.9 ± 0.3	-10.0	-4.9	-9.0
ETH	13	3,043	-6.8 ± 0.3	-8.2	-5.9	-6.7

2.2 Discretization of the Target Attribute FEB

Since classification decision trees require a categorical target attribute, and being FEB a continuous one, we needed to discretize it [7]. Discretization involves two subtasks: (1) the continuous attribute is sorted and divided into *n-1* split points; and (2) determination of how to map the values of the continuous attribute to the defined categories [7]. Here we considered three unsupervised discretization methods:

- **Method 1 - Discretization by equal frequency interval:** Considering that *k* is the number of intervals defined by the user and *m* the total of instances, this method

divides the continuous variable into k intervals where each interval contains m/k values, approximately.

- **Method 2 - Discretization by equal width interval:** In this approach, for each continuous attribute to be discretized, their values are sorted and then divided into k intervals, defined by the user [7], where each interval has the same width. According to Dougherty et al. [13] this is the simplest discretization method, although it is vulnerable to outliers.
- **Method 3 - Discretization by mode and standard deviation:** The aim of our proposed discretization approach is to divide the sorted attribute into intervals where the border of the instances distribution (best and worst FEB values) fit together in the same class. To achieve this we considered the mode and the standard deviation of the frequency distribution of the attribute that is being discretized. Our discretization method with 5 classes is shown in equation (1) where x is the attribute to be discretized, Mo and σ represents the mode and standard deviation values of the x distribution.

$$Class = \begin{cases} Class1 & if & Mo - 2*\sigma & > & x & & \\ Class2 & if & Mo - \sigma & > & x & \geq & Mo - 2*\sigma \\ Class3 & if & Mo + \sigma & > & x & \geq & Mo - \sigma \\ Class4 & if & Mo + 2*\sigma & > & x & \geq & Mo + \sigma \\ Class5 & if & & & x & \geq & Mo + 2*\sigma \end{cases} \quad (1)$$

3 Results and Discussion

The discretization methods by equal frequency (Method 1), by equal width (Method 2) and by mode and standard deviation (Method 3) were applied to our target attribute FEB. As a result it mapped the FEB values into 5 classes: *Excellent*, *Good*, *Regular*, *Bad* and *Very bad*. The results of the mappings are presented on Table 2. For each class it shows the number of instances for each ligand in each of the three discretization methods tested.

Table 2. Total number of instances for each of the 5 classes considering the three discretization methods: by equal frequency (1), by equal width (2) and by mode and standard deviation (3).

	PIF			NADH			TCL			ETH		
Classes	1	2	3	1	2	3	1	2	3	1	2	3
Excellent	604	2995	7	569	757	205	563	1017	19	619	18	160
Good	607	26	223	559	792	1020	556	1814	158	591	173	512
Regular	620	17	2616	565	839	374	587	4	1866	598	1108	2131
Bad	610	3	173	565	408	903	582	0	645	649	1531	226
Very_bad	601	1	23	565	27	321	549	2	149	586	213	14

From Table 2 we can observe that Method 1 discretized the instances in a balanced form with near the same number of instances in each class. Method 2 discretizes the instances in intervals of same width. It can generate unbalanced classes since instances are not equally distributed. It happens especially for PIF and TCL. For TCL,

FEB varies from -10.0 to -4.9 kcal/mol (see Table 1) and the mode is -9.0 kcal/mol, closer to the minimum than to the maximum FEB. In addition, its FEB standard deviation is 0.3 kcal/mol which means that FEB does not vary much and remains around the mode. Considering the PIF and TCL ligands, since most of their instances were grouped in *Excellent* or *Good* classes, their induced model by classification decision trees algorithms will be distorted. On the other hand, our proposed Method 3 generates unbalanced classes (Table 2). Nevertheless, its main advantage is that it groups the best docking results in one class, as well as the worst docking results in another, different class.

Acknowledgments. This work was supported by grants from MCT/CNPq 14/2008 to DDR and by 410505/2006-4 and 312027/2006-0 to ONS. ONS is a CNPq Research Fellow. KSM is supported by a CAPES PhD scholarship and ATW by CT-INFO/CNPq 17/2007 PhD scholarship.

References

1. Lengauer, T., Rarey, M.: Computational methods for biomolecular docking. Curr. Opin. Struct. Biol. 6, 402–406 (1996)
2. Morris, G., Goodsell, D., Halliday, R., Huey, R., Hart, W., Belew, R., Olson, A.: Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J. Comput. Chem. 19, 1639–1662 (1998)
3. Machado, K., Schroeder, E., Ruiz, D., Norberto de Souza, O.: Automating Molecular Docking with Explicit Receptor Flexibility Using Scientific Workflows. In: Sagot, M.-F., Walter, M.E.M.T. (eds.) BSB 2007. LNCS (LNBI), vol. 4643, pp. 1–11. Springer, Heidelberg (2007)
4. van Gunsteren, W., Berendsen, H.: Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry. Angew. Chem. Int. Ed. Engl. 29, 992–1023 (1990)
5. Lin, J.-H., Perryman, A., Schames, J.R., McCammon, J.A.: Computational drug design accommodating receptor flexibility: the relaxed complex scheme. J. Am. Chem. Soc. 124, 5632–5633 (2002)
6. Schroeder, E., Basso, L., Santos, D., Norberto de Souza, O.: Molecular dynamics simulation studies of the wild-type, I21V, and I16T mutants of isoniazid-resistant Mycobacterium tuberculosis enoyl reductase (InhA) in complex with NADH: toward the understanding of NADH-InhA different affinities. Biophys. J. 89, 876–884 (2005)
7. Tan, P., Steinbach, M., Kumar, V.: Introduction to data mining. Addison Wesley, Boston (2006)
8. Dessen, A., Quémard, A., Blanchard, J., Jacobs Jr., W., Sacchettini, J.: Crystal structure and function of the isoniazid target of Mycobacterium tuberculosis. Science 267, 1638–1641 (1995)
9. Oliveira, J., Souza, E., Basso, L., Palaci, M., Dietze, R., Santos, D., Moreira, I.: An inorganic iron complex that inhibits wild-type and an isoniazid-resistant mutant 2-trans-enoyl-ACP (CoA) reductase from Mycobacterium tuberculosis. Chem. Commun. 3, 312–313 (2004)
10. Kuo, M., et al.: Targeting tuberculosis and malaria through inhibition of Enoyl reductase: compound activity and structural data. J. Biol. Chem. 278, 20851–20859 (2003)

11. Banerjee, A., Dubnau, E., Quemard, A., Balasubramanian, V., Um, K., Wilson, T., Collins, D., de Lisle, G., Jacobs Jr., W.: InhA, a gene encoding a target for isoniazid and ethionamide in Mycobacterium tuberculosis. Science 263, 227–230 (1994)
12. Winck, A., Machado, K., Norberto de Souza, O., Ruiz, D.: FReDD: supporting mining strategies through a flexible receptor docking database. In: Guimarães, K.S., Panchenko, A., Przytycka, T.M. (eds.) Advances in Bioinformatics and Computational Biology. LNCS, vol. 5676, pp. 143–146. Springer, Heidelberg (2009)
13. Dougherty, J., Kohavi, R., Sahami, M.: Supervised and unsupervised discretization of continuous features. In: The Proceedings of the 12th International Conference on Machine Learning, Tahoe City, CA, USA, pp. 194–202 (1995)

Author Index