Visualisation of Associations Between Nucleotides in SNP Neighbourhoods

Kimmo Kulovesi[*,†], **Juho Muhonen**[*], **Ilkka Lappalainen**[‡], **Pentti T. Riikonen**[§], **Mauno Vihinen**[¶,†], **Hannu Toivonen**[*,‖] and **Tomi A. Pasanen**[*]

Abstract

A large number of single nucleotide polymorphisms have been mapped onto the human genome. Mutations are induced through endogenous and exogenous processes, and these procedures have been shown to be sequence-dependent. Association mining is a powerful tool for analyzing sequence neighbourhoods; however, visualisation is essential for pattern recognition because of the abundance of resulting association rules.

A software tool was developed to visualize position interdependencies within the sequence variation data. The software is capable of interactive reorganization of the association rules, enabling fast and easy exploration of the data using a standard web browser. The software and its complete source code is freely available at: http://www.cs.helsinki.fi/group/bioalgss/asvis/

1 Introduction

A single nucleotide polymorphism (SNP) is a site in DNA where at least two different nucleotides occur in a specific population, the less frequent nucleotide(s) occurring at a frequency of 1% or more. The nucleotide variation between individuals forms the genetic background responsible for biological and physical differences such as colour of hair, susceptibility to a disease or response to specific treatment. The International HapMap project aims to characterise the common human sequence variations [International HapMap Consortium, 2005]. The data is publicly available in the dbSNP database [Wheeler *et al.*, 2005].

New mutations arise by errors in endogenous processes involved in maintaining genomic stability, or are induced by various exogenous agents, such as UV radiation [Jiricny, 1998]. The efficiency and specificity of these processes is DNA sequence dependent [Cooper and Krawczak, 1993]. Data mining can be used to analyze the sequence neighbourhoods of neutral and disease-causing SNPs in order to better understand the genetic differences that underlie pathogenic conditions.

Association rules are suitable for this task. Discovery of association rules is popular in data mining and it has a wide variety of applications. In principle, individual association rules describing co-occurrences of sets of attributes in the input data are straightforward to interpret. However, the very large number of resulting association rules seriously hinders their analysis. General-purpose visualisation tools are ill-suited for association rules, and the few available association visu-alisation tools are not suitable for position-dependent SNP neighbourhoods. Furthermore, the existing tools for visualizing SNP neighbouring-nucleotide biases are not applicable to association rules (for example, [Zhang and Zhao, 2005]). Here we introduce a new, publicly available visualisation tool with interactive controls for the selection and arrangement of association rules obtained from SNP data. The tool has a novel position-dependent display of association rules. It can also be used with other similar data, such as protein sequences. The tool provides simple and legible graphical output.

2 System and Methods

For demonstration in this paper, sequence variation data was extracted from the dbSNP (build 124). Only true polymorphisms that could be located within gene coding regions of the human genome (build 35) were used. We consider a sequence neighbourhood of each mutation site that extends up to ten nucleotides on both sides. The mutation position is numbered zero, while positive and negative position numbers denote the distance of following and preceding nucleotides, respectively.

Our software consists of two programs, Firm and AsVis. Firm is used for application-independent discovery of association rules. As an example, the rule $\{0{:}\text{'C}\rightarrow\text{T'}\} \longleftarrow \{+1{:}\text{'G'}\}$ indicates that the substitution of cytosine (C) by thymine (T) is probable when the mutation site is immediately followed by guanine (G) (see Figure 1). To rank the association rules, Firm yields a number of measures for the strength and generality of each rule within the data. Support is the number of records that fully match the rule. Con-

[*]Department of Computer Science, FI-00014 University of Helsinki, Finland

[†]Institute of Medical Technology, FI-33014 University of Tampere, Finland

[‡]Department of Chemistry, Cambridge University, CB2 1EW Cambridge, UK

[§]Department of Information Technology, FI-20520 University of Turku, Finland

[¶]Research Unit, Tampere University Hospital, FI-33520 Tampere, Finland

[‖]Department of Computer Science, University of Freiburg, D-79110 Freiburg, Germany

Consequent	<=	Condition	Support (s)		Frequency (f)		Confidence (c)		Lift (l)		J-Measure (j)	
0:'>T'	<=	0:'C>' 1:'G' 2:'A'	s=	1140	f=	4.1	c=	87.9	l=	2.49	j=	0.012
0:'>T'	<=	0:'C>' 1:'G' -1:'C'	s=	1100	f=	4.0	c=	87.9	l=	2.49	j=	0.012
0:'>T'	<=	0:'C>' 1:'G' -6:'A'	s=	722	f=	2.6	c=	83.9	l=	2.38	j=	0.007
0:'C>'	<=	0:'>T' 1:'G' -1:'C'	s=	1100	f=	4.0	c=	83.9	l=	2.64	j=	0.012

-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	Confidence	Freq.	Support	Lift	J-Measure
										C >T	G	A									87.9	4.1	1140	2.49	0.012
									C	C >T	G										87.9	4.0	1100	2.49	0.012
				A						C >T	G										83.9	2.6	722	2.38	0.007
									C	C >T	G										83.9	4.0	1100	2.64	0.012

Figure 1: Example association rules (above) visualised by AsVis (below). No information is lost in the visualisation, but the positional nature of the rules is immediately apparent.

fidence of a rule is the conditional probability of the consequent (e.g. {0: 'C→T'}) given the condition ({+1:'G'}). Lift is simply the ratio of confidence over the relative frequency of the rule consequent, whereas J-measure [Smyth and Goodman, 1992] is an information-theoretic measure describing the amount of information that the condition gives about the consequent.

Association rules are suitable as such for exploratory data analysis, and a large amount of them can be discovered efficiently, without setting a strong focus on any particular attributes. Usually only support and confidence thresholds are used to limit the number of rules, and algorithms such as Apriori [Agrawal *et al.*, 1996] produce all association rules between any sets of attributes that exceed the thresholds. A rule can have any number of conditions and consequents, but with reasonable threshold values, our data did not give any strong associations for complex dependencies with more than one consequent.

To facilitate visual browsing and exploration of thousands of rules, an interactive web interface, AsVis, was developed. AsVis graphically renders the rules into a form that visually reflects the sequential nature of the data, taking advantage of the relatively small number of dimensions (positions). The association rules are listed in a table, where each row represents a single rule, and each column corresponds either to a position relative to the point of mutation or to a strength or generality measure for the rule (see Figure 1). Consequents and conditions are colour-coded in the positional columns.

The interface enables the user to explore the most interesting rules. Clicking on the column header of a measure sorts the rules by that measure, with the best scores at the top. Clicking on a positional column header limits the display to only those rules that have either a condition or a consequent in that position, while keeping the current sorting. This approach is somewhat similar to the TASA system [Klemettinen *et al.*, 1999].

This simple approach enables quick visual scanning of the rules for an overview of dependencies between positions. The measures for each rule are for closer inspection, providing the full scope of information discovered by the data mining process.

Acknowledgements

This research has been supported by Tekes (the National Technology Agency of Finland), The Medical Reseach Fund of Tampere University Hospital, and the Academy of Finland. We thank Adrian Nickson for valuable discussions.

References

[Agrawal *et al.*, 1996] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. pages 307–328, 1996.

[Cooper and Krawczak, 1993] David N. Cooper and Michael Krawczak. Human Gene Mutation. 1993.

[International HapMap Consortium, 2005] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

[Jiricny, 1998] J. Jiricny. Replication errors: cha(lle)nging the genome. 17(22):6427–6436, 1998.

[Klemettinen *et al.*, 1999] M. Klemettinen, Heikki Mannila, and Hannu Toivonen. Interactive exploration of interesting patterns in the telecommunication network alarm sequence analyzer tasa. *Information and Software Technology*, 41:557–567, 1999.

[Smyth and Goodman, 1992] P. Smyth and R. M. Goodman. An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering*, 4(4):301–316, 1992.

[Wheeler *et al.*, 2005] David L. Wheeler, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Wolfgang Helmberg, David L. Kenton, Oleg Khovayko, David J. Lipman, Thomas L. Madden, Donna R. Maglott, James Ostell, Joan U. Pontius, Kim D. Pruitt, Gregory D. Schuler, Lynn M. Schriml, Edwin Sequeira, Steven T. Sherry, Karl Sirotkin, Grigory Starchenko, Tugba O. Suzek, Roman Tatusov, Tatiana A. Tatusova, Lukas Wagner, and Eugene Yaschenko. Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res.*, 33:D39–45, 2005.

[Zhang and Zhao, 2005] Fengkai Zhang and Zhongming Zhao. SNPNB: analyzing neighboring-nucleotide biases on single nucleotide polymorphisms (SNPs). *Bioinformatics*, 21(10):2517–2519, 2005.