Sequence analysis

A knowledge-based approach to predict intragenic deletions or duplications

Krishna R. Kalari¹,², Thomas L. Casavant¹,²,³,⁴ and Todd E. Scheetz¹,²,³,⁴,*

¹Department of Biomedical Engineering, ²Center for Bioinformatics and Computational Biology, ³Department of Electrical and Computer Engineering and ⁴Department of Ophthalmology and Visual Sciences, University of Iowa, Iowa

Received on December 19, 2007; revised on June 17, 2008; accepted on July 14, 2008

Advance Access publication July 21, 2008

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Despite recent improvements in high-throughput or classic molecular biology approaches it is still challenging to identify intermediate resolution genomic variations (50 bp to 50 kb). Although array-based technologies can be used to detect copy number variations in the human genome they are biased to detect only the largest such deletions or duplications. Several studies have identified deletions or duplications occurring within a gene that directly cause or predispose to disease. We have developed a novel computational system, SPeeDD (system to prioritize deletions or duplications) that utilizes machine learning techniques to predict likely candidate regions that delete or duplicate exon(s) within a gene.

Results: Data mining and machine learning methods were applied to identify sequence features that were predictive of homologous recombination events. The logistic model tree (LMT) method yielded the best results. Sensitivity varied from 20% to 71.6% depending on the specific machine learning model used, but specificity exceeded 90% for all methods evaluated. In addition, the SPeeDD system successfully predicted and prioritized a recently published novel BRCA1 mutation.

Conclusions: Results suggest that the SPeeDD system is effective at prioritizing candidate deletions and duplications within a gene. Use of SPeeDD enables more focused screening, which reduces the labor and associated costs of the molecular assays and may also lead to targeted design of new array-based screens to focus on candidate areas to accelerate the process of mutation discovery.

Contact: tscheetz@eng.uiowa.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Although a variety of methods are readily available to identify mutations, there are several reasons that the disease-causing mutations for specific individuals may not be identified. These include genetic heterogeneity, phenocopy and categories of mutations that are not commonly assayed. This last case includes mutations present in alternatively spliced exons, non-coding regions of genes and changes in the copy number of genes or gene components. This study focuses on the latter class of mutations—the subset that results in duplications or deletions of gene components such as exons. The process whereby sequences recombine and insert or delete a portion(s) of DNA within a gene is termed an intragenic deletion or duplication (IDD).

The Human Gene Mutation Database (HGMD) consists of mutations that are responsible for human inherited diseases. IDDs are an under-surveyed class of mutation; only 6% of the mutations available in HGMD are IDDs (Stenson et al., 2003). It is known that IDDs are non-randomly distributed in the human genome (Mitelman, 2000) and are responsible for the predisposition or cause of disorders. IDD genotypes have been known to cause phenotypes such as breast cancer, colon cancer, amelogenesis imperfecta and pseudoxanthoma elasticum (Mitelman, 2000; Puget et al., 1999; Rabbitts, 1994; Stankiewicz and Lupski, 2002).

Deletions and duplications (DDs) can be molecularly confirmed using several different assays, with each technique having its own advantages and disadvantages. Which assay is best depends on a variety of factors including the size of the mutation, the number of samples to be assayed and how specifically the ends of the DDs are known. For example, large-scale mutations such as chromosomal abnormalities are usually detected by performing fluorescence in situ hybridization (FISH), Southern blotting, Comparative Genome Hybridization (CGH) and arrayCGH. If the number of samples to be assayed is large, then the overhead cost of acquiring and adopting an arrayCGH platform may be amortized across a large number of samples. Smaller mutations affecting a few tens or hundreds of bases can be confirmed using techniques such as PCR or Southern blotting when the sample size is low, but they are often time consuming and laborious. On the other hand when the sample size is large, whole genome-based arrays such as SNP or CGH arrays can be used to detect deletions or duplications. Although these high-throughput methods are available, they are costly and do not have sufficient resolution to robustly identify many of the reported IDD mutations.

Previous studies revealed that IDDs occur during unequal crossover events associated with repetitive elements (Abeysinghe et al., 2003; Batzer and Deininger, 2002; Hsieh et al., 2005; Rudiger et al., 1995; Sen et al., 2006). Although repetitive elements have been suggested as the mediator of IDDs, very little information is available on the specific mechanism(s) of these events. The study by Abeysinghe et al. in 2003 concluded that there was no specific motif that was found to be overrepresented at the mutation site but

*To whom correspondence should be addressed.
found the presence of repetitive elements such as Alus at the break point sites of IDDs. They also concluded that homologous unequal recombination played a role during deletion mutagenesis and that deletion break points were significantly over-represented with AT-rich and alternating purine–pyrimidine sequences (Abeysinghe et al., 1976). Little is known about the involvement of particular sequences and sequence features in unequal recombination. Another study revealed that length and similarity between sequences correlate positively with the distance between sequences during recombination; features such as length, similarity and distance between sequences play a role during unequal crossing over (Lupski, 1998). Hence, that study focused mainly on ‘significant regions of homology’ that are required for recombination and looked into the sequence-specific features that play a major role during this process.

Availability of genomic sequences, annotations, HapMap data and software such as BL2SEQ—to obtain similar sequences (Tatusova and Madden, 1999), TMalign—to measure melting temperature characteristics permits new approaches to compare and contrast different IDDs already published using data mining algorithms. For more than a decade data mining has been popular for identifying patterns that could not be found by manual analysis. Data mining methods utilize combinations of statistical analysis, artificial intelligence and machine learning technologies. There are several studies that used data mining algorithms such as neural networks, hidden Markov models and support vector machines (SVMs) to identify core promoters (Ben-Hur and Brutlag, 2003; Ohler et al., 2002; Pavlidis et al., 2001; Reese, 2001; Sharan and Myers, 2005). Our goal is similar to design and develop a method to identify IDDs using data mining algorithms.

Rather than conducting biological experiments on the hundreds of candidates with an exhaustive search to identify medium size mutations, our study utilizes a novel computational and machine learning system termed SPeeDD (System to Prioritize Deletions or Duplications) to narrow the list of IDD candidates. The high-level view of our approach is shown in Figure 1. We manually obtained and curated the features of previously published IDD case and control sequences and trained the machine learning model to predict the most likely disease-causing IDDs for a given gene (case and control sequences are defined in the Section 2). SPeeDD prioritizes the screening of IDD candidates, allowing investigators to focus on only the most likely ones. The SPeeDD process thus reduces the labor and associated costs of biological assays, and accelerates the process of mutation discovery.

2 METHODS

2.1 Data collection

We manually obtained and curated the definitions of all DD mutations from the comprehensive database of published human mutations maintained in HGMD (Stenson et al., 2003). This search was performed for all of the 12,371 gene names with official HUGO symbols (Povey et al., 2001). Approximately 6.5% of the Human Genome Organization (HGMD) contained large deletions or duplications. All of the published deletions or duplications were obtained from HGMD and stored in our database of human gross deletions or duplications. This locally maintained database consists of 441 genes with known deletions or duplications and 1463 links to the primary literature in which they were published. We manually examined these publications and collected the set of genes with IDDs having published sequences of break point regions. From the literature review and follow-up contact with authors through email we were able to collect 102 IDDs across 72 genes with exact break points published. The remaining IDDs had indeterminate break points due to: (1) vague description of the exact mutation break points or deleted regions, (2) assay variations some of which referred to genomic, and some to RNA/cDNA context or (3) incomplete description of the recombination site. The 102 IDDs we obtained had Alu elements at their break points. This may be due to the abundant presence of these elements in human genome and also due to the similarity between these elements (Batzer and Deininger, 2002).

After collecting 102 different cases from the literature for which the exact break points were known, we computed all potential pairs of recombining sequences using the computational approach described below. For each case there are typically numerous combinations of similar sequence that could recombine, resulting in IDDs with identical consequence to the gene structure [e.g. deletion of specific exon(s)], we collected all such potential DNA sequences from the BL2SEQ output and classified them as cases (observed) and controls (unobserved) for training the SPeeDD system. Further details of case and control dataset construction are included in the Supplementary methods section.

2.2 Computational method to identify candidate break point sequences

To identify candidate break points a computational system was developed using the UCSC genome annotation database (Karolchik et al., 2003) and the BL2SEQ program. For each gene under study, the genomic sequence of the gene plus 5 kb of the 3′ and 5′ flanking sequences were obtained and aligned against themselves using BL2SEQ. The filtering criterion of the BL2SEQ output is shown in Figure 2. Only the non-redundant set of alignments was maintained. Only those alignments at least 30 bp in length, 80% identity and <50 kb away were considered. The exact context (intron, exon, promoter or UTR) of these similar sequences was obtained using the gene structure annotation from the RefSeqs (Pruitt et al., 2007). Details of similar sequences that span exons were obtained as shown in the computational pipeline (Fig. 2). For each gene, the fully characterized break point sequence was identified in the pairs of similar sequences obtained from the BL2SEQ output. These break point sequences were used to calculate the set of informative features for data mining. For each candidate IDD, the sequence-specific haplotype block structure and melting temperature features were calculated.

2.3 Features obtained for data mining

(a) Sequence-specific features—software was developed to calculate the sequence-specific features. These include analysis of the sequence length, identity and distance between the flanks directly from the BL2SEQ output file or analysis of the sequence files themselves. (b) Melting temperature
features—the TmAlign program was obtained from IDT (Integrated DNA Technologies; unpublished software) to calculate melting temperature ($T_m$) of the two sequences flanking the candidate IDDs. TMalign uses energy files, hybridization temperature, oligo concentration and salt concentration during alignment. Recombination is more likely when the homologous sequences contain longer perfect matches (Deininger and Batzer, 1999). Hence, longest exact subsequence properties were also utilized in the analysis.

(c) The start and end positions of the haplotype blocks defined in (Hinds et al., 2005) were obtained, and stored in our local database. The haplotype block features used in SPeeDD are classified based upon whether the break points are in the same block, in different blocks or not in any haplotype block (INSAME, SPAN and OUT, respectively). Distance to the nearest haplotype blocks is obtained for both ends of candidate IDD break points.

Although several features were obtained (Supplementary Table 1), not all were used in the final analyses. Highly correlated features were removed as they added no additional information. The following features were utilized in the final model to classify the control and case datasets efficiently—length, percent identity, GC content, melting temperature and BL2SEQ alignment score (bl2score) of the candidate break points, the distance and GC content between the break points, the haplotype block status and the nearest haplotype block boundary to both ends of the IDD were included in the final analysis.

2.4 Validation of classifiers

The fitness of each classifier was evaluated using cross-validation methods. Two cross-validation techniques were performed on our dataset ($N$-fold cross-validation and ‘leave-one-out’). Cross-validation methods are commonly used for estimating the performance and generalization error. The advantage with the cross-validation method is it estimates the generalization error on multiple different subsets rather than a single subset.

2.5 Evaluation of the classifier

The entire feature set data for the cases and controls was obtained and data mining was performed using Weka (Witten and Frank, 2005), an open source machine learning toolkit. Performance of the SPeeDD system was measured in terms of sensitivity and the specificity. Sensitivity of a set of predictions is described as the percent of positives that are correctly classified (TP/TP + FN), and the specificity is defined as the percent of negatives that are correctly predicted (TN/TN + FP). The four different outcomes for a two-class classifier are TP (true positive), TN (true negative), FP (false positive) and FN (false negative). TP and TN denote the two correct classifications.

Conversely, when a case input is incorrectly classified as control candidate it is termed an FN; and when a control input is incorrectly classified as a case candidate it is termed an FP. Because only a fraction of bona fide IDDs have been reported to date, it is certain that some elements identified as FP will later be found to be TP. Thus, we use the terminology case and control rather than positive and negative when referring to those sets of candidate IDDs.

3 RESULTS

Despite recent advances in the field of high-throughput assays and bioinformatics, identification of small insertions or deletions (50 bp to 50 kb) using existing methods continues to be a challenge (Bhangale et al., 2005; Conrad et al., 2006; Sharp et al., Weber et al., 2002). Our study enhanced the knowledge of small and intermediate level mutations by: (a) mining sequence features of deletions or duplications reported to cause disease; and (b) building an effective machine learning classification model to prioritize candidate unknown DDs.

3.1 Feature selection

The results of any automated analysis depend on the nature and quality of the data analyzed. Therefore, the selection of appropriate DNA features is a very important step. There are several features to be examined for each potential pair of recombining elements. For example, Stankiewicz and Lupski demonstrated (Stankiewicz and Lupski, 2002) that during the process of recombination, the more identical and the closer together the elements are the greater the chances of recombination. Hence, parameters such as the length of the homologous sequences, and the percent identity and the distance between them play a vital role. It is also known that relative GC content densities in the genome play a role in the recombination rates (Fullerton et al., 2001). Hence, GC content parameters of the DNA sequences are included. In the same way melting temperature and haplotype characteristic features were also included in the analysis. A list of the features selected, and the computational methods used to obtain them are found in Supplementary Table 1.

3.2 Building the classifier model

The classifier was built using the feature set data obtained from the case and control dataset as described in the Section 2. The classifier model was trained to predict and prioritize IDD candidate regions based on a confidence score. The model used Weka machine learning system (Witten and Frank, 2005). The training data and candidates to be evaluated were converted into the Attribute-Relation File Format (ARFF) format natively used by Weka. Each ARFF record consists of a set of features and is assigned a class or outcome. Based on training data, the classifier model determines the properties to distinguish case and control datasets.

3.3 Selecting the appropriate machine learning system

A variety of classifiers including artificial neural networks (ANNs), decision trees, logistic model tree (LMT), simple logistic (SL), simple naive bayes (NBS), k-nearest neighbor and SVMs were evaluated on the training dataset. A 10-fold cross-validation was used to evaluate the models, and the best model was selected based on its performance (prediction of TP and FP) and its ability to deal with missing or noisy data. Overall, sensitivity of the system varied from 20% to 74% but the specificity exceeded 90% for all the
methods that were assessed (Fig. 3). The high specificity may be due, in part, to the utilization of an unbalanced dataset. The majority of the training dataset consisted of control sequences, with cases making up only 4.2% of the dataset. This is typical of many real-world problems, in which there are often far fewer cases than controls.

As shown in Figure 3 the sensitivity of the SL and ANN models were inferior to the other models. A marked increase in performance was observed with the NBS, decision tree (J48) and SVM methods with sensitivity ranging from 60% to 65%, while specificity of all three methods exceeded 95%. However, the LMT model, which is a combination of the logistic regression and decision tree methods, yielded the best results with sensitivity and specificity of 74.2% and 97.2%, respectively (Fig. 3). In case of the LMT model, an additional benefit is that the rules used in the classification are easily interpretable providing an insight into how the classifier works. In addition to the 10-fold cross-validation, the leave-one-out validation was performed using the LMT model on the dataset. The leave-one-out yielded similar results with a little increase in our sensitivity to 81%.

3.4 Role of each feature in predicting the candidates
The importance of each feature in predicting the TPs was evaluated by removing each feature, one at a time, from the collection of features. The effect was measured by evaluating the performance of 10-fold cross-validation in predicting the IDD candidates. The influence of each feature was estimated by comparing the difference between the performance with and without the feature. As shown in Figure 4 removal of GCbwn (GC between the IDD candidate break points) feature had the largest impact on performance and reduces the power of true prediction drastically. Percent identity and distance between the IDD break point sequences also had significant impacts on the performance of the classifier when removed (Fig. 4).

As noted above, all of the 102 cases identified in the literature were flanked by Alu elements. However, when included as a feature in the machine learning process this did not increase our ability to accurately identify IDDs.

3.5 In silico BRCA1 mutation verification
Subsequent to the development and tuning of the model and classifier used in SPeeDD, a novel duplication within the BRCA1 gene was identified in a population of Chinese ethnicity (Yap et al., 2006). This mutation is a duplication of 8.4kb, resulting in an additional copy of exon 13. Because this mutation had been recently reported and had not been included in our training set we considered it as a candidate to validate the SPeeDD system. When the BRCA1 gene was processed with SPeeDD, this exon 13 spanning IDD was the sixth best candidate out of the 44 candidates predicted as likely to be involved in an exon deletion or duplication event (from a total of 5645 pairs of potential sites, and two previously described IDDs).

4 CONCLUSIONS
SPeeDD successfully enriches lists of candidate IDD regions within genes based upon a compact set of decisions or rules that look at sequence-based features. Given the abundance of repetitive elements within a typical gene, it is impractical to exhaustively test for gene-wide deletions or duplication candidates. SPeeDD is an efficient method to prioritize candidate deletion or duplication regions. This strategy is not meant to replace experimental validation within the lab. However, it can save both time and effort by allowing investigation to focus on the set of most likely IDDs within a gene of interest. SPeeDD utilizes tools that are already available and tools that are internally developed; the suite of software with instructions will be made available upon request (see Supplementary methods section for result interpretation).

Current high-resolution SNP or CGH platforms (Illumina human1M, Affymetrix SNP array 6.0, Nimblegen arrays, etc) allow the identification of copy number variations but are limited by uneven probe distribution and number. SPeeDD is more specific when identifying candidate deletions or duplication events within genes of interest. In other words, it can identify candidate IDDs at a higher resolution than are robustly detectable by current array-based assays, and also identifies the exact break points predicted to be involved for all predicted candidates. The computational system developed in this study elucidated several important features that play a role in intermediate size deletion or duplication mutations. It enables investigators to identify the set of most likely IDDs for which a custom array/assay can be quickly and accurately fabricated. The findings in this article will not only aid in the rapid design of high-throughput assays in focused regions or genes, but will also help in identifying novel disease-causing mutations and accelerate the process of mutation discovery.

ACKNOWLEDGEMENTS
The authors thank Dr Andy Peek (IDT-Integrated DNA Technology) for providing early access to the TmAlign program for accurate
calculation of melting temperatures. T.E.S. was partially supported through a Career Development Award from Research to Prevent Blindness. We thank the reviewers for their many helpful suggestions.

Conflict of Interest: none declared.

REFERENCES


