

Identification of High-Polymorphic Dinucleotide Tandem Repeats Using a Machine Learning Approach

HAIFENG LIU * LOO-NIN TEOW

DSO National Laboratories, 20 Science Park Drive, Singapore 118230

ERIC YAP SOCK-HOON NG HUI MIN WU SEO-HWEE GAN

Defence Medical Research Institute, 10 Medical Drive, Singapore 117597

Abstract

It is not a trivial task to decide whether a given short tandem repeat is polymorphic enough to function as a genetic marker. Our main objective is to obtain a large number of tandem repeats that are as highly polymorphic as possible. By applying machine learning to the Genethon human genetic map, we determine the factors or features that may affect the polymorphism (heterozygosity) of dinucleotide tandem repeats. We then use a hybrid approach which combines a K-Nearest Neighbor classifier and two artificial neural networks trained on the feature data of Genethon markers to predict the polymorphism of a given dinucleotide tandem repeat. This approach was validated on the human chromosome 21 and 22 sequences and one DNA segment from human chromosome 12 which is suspected to contain some myopia-related genes. The result shows that our approach achieves a high accuracy of 93% in predicting high-polymorphic DTRs (whose heterozygosity is higher than 50%). Our main contribution lies in helping to speed up the construction of a denser genetic map for haplotype mapping and genomewide linkage disequilibrium studies of complex diseases. In addition, our work brings about a better understanding of the mutation pattern of short tandem repeats by identifying their polymorphism factors. To the best of our knowledge, our work is the first application of machine learning to the problem of computationally selecting candidate DTRs for validation as genetic markers.

Keywords: STR(Microsatellite), Polymorphism, ANN, K-NN, Machine Learning

1 Introduction

Genetic markers play an important role in the localization of human disease loci in positional cloning. Microsatellites or short tandem repeats (STRs) have been the markers of choice since 1989, and several thousands of such polymorphic markers have been identified ([7, 9, 13, 24, 43, 44]). Characterized by high levels of polymorphism and by a large number of alleles, such markers provide ideal tools for pedigree-based linkage analysis. They have also been exploited for population and evolutionary genetics studies. The current widely used method for identification

*Author for correspondence. E-mail: lhaifeng@dso.org.sg, Tel:65-67728220, Fax:65-67759011

of new polymorphic STRs is wet lab (PCR) based, and suffers from the disadvantages of time-consuming procedure and dependency on large-scale mapping facilities and genomic DNA from multiple individuals.

Our main objective is to obtain a large number of tandem repeats that are as highly polymorphic as possible, since high polymorphism greatly facilitates gene-disease mapping. With the availability of the draft human genome sequence ([27]) and the knowledge accumulated on the polymorphism of tandem repeats, it is now possible to develop computational methods to find tandem repeats and predict the polymorphism of a tandem repeat within a single genomic sequence. The degree of polymorphism of a genetic marker is measured by the metric *heterozygosity* which is defined as $H = 1 - \sum p_i^2$, where the sum is taken over all alleles of the marker with p_i denoting the frequency of the i -th allele. Although several software tools ([14, 4, 32, 22]) have been developed for finding tandem repeats within DNA sequences, there have been few research efforts on developing tools for predicting the polymorphism of STRs. We found only one such computational system ([11]), known as POMPOUS, which has been developed for the prediction of polymorphic loci directly from human genomic sequence. However, POMPOUS was designed to distinguish between polymorphic and non-polymorphic STRs, whereas our work aims to distinguish between high-polymorphic STRs and low-polymorphic ones (which may include non-polymorphic STRs). Moreover, POMPOUS's criteria for polymorphism is directly based on the threshold number of repeating units and minimum homogeneity without any analysis of other polymorphism factors. The prediction accuracy for the polymorphism of their sample data (containing 33 loci) is 67%, and the average heterozygosity of the polymorphic loci is only 42%. On the other hand, we could not find any previous work on predicting high-polymorphic STRs.

In this paper, we propose a machine learning approach to specifically predict the polymorphism of dinucleotide tandem repeats (DTRs) which are abundantly distributed in the human genome ([8]). As the well-known Genethon human genetic map ([7]) contains all known polymorphic DTRs (so-called Genethon markers), our work would be able to improve the density of the map. To fulfill the task, we first identify a list of factors or features that may affect the polymorphism of a DTR. These may include repeat length, repeat complexity, repeat composition, and some features of its flanking sequence (A review of polymorphism factors is provided by Schlotterer ([35])). We use K-Nearest Neighbor (K-NN) method to select the features that best predict the heterozygosity of a DTR by learning from the Genethon markers with known heterozygosity. Then, we use the combination of K-NN method and artificial neural networks to predict the heterozygosity of a novel DTR based on its selected features. A high heterozygosity ($> 50\%$) implies high polymorphism; otherwise, low polymorphism is implied by a low heterozygosity ($\leq 50\%$). This approach was applied to all DTRs that are detected from chromosome 21, 22 and a segment of chromosome 12 using a publicly available program Tandem Repeat Finder (TRF) ([4]). The wet lab validation result shows that our approach achieves a high accuracy of 93% in predicting high-polymorphic DTRs. Our main contribution lies in helping to speed up the construction of a denser genetic map for linkage analysis. In addition, our work brings about a better understanding of the mutation pattern of STRs by identifying their polymorphism factors. To the best of our knowledge, our work is the first application of machine learning to the problem of computationally selecting candidate DTRs for validation as genetic markers.

2 Data

From the Genethon genetic map ([1, 7]), we have downloaded the DNA sequences (which are composed of tandem repeats and flanking sequence) of a total of 5264 Genethon markers. The markers have a mean heterozygosity of 70%. We used TRF (with the input parameters as 2, 7, 7, 80, 10, 3, 2) to detect the DTR within each marker. As the TRF program is based on

Classlabel	1	2	3
Heterozygosity(%)	(0, 50]	(50, 80]	(80, 100]
Number of instances	338	3591	909

Table 1: 3 Classes of Genethon markers

an heuristic approach, 4838 DTRs, denoted as set G , were found to be located between the primer pairs of the corresponding markers and taken as our polymorphic DTR examples. For our purpose, based on the different ranges of heterozygosity of the Genethon markers, we classify them into 3 classes. The range of heterozygosity and the number of instances of each class of G are shown in Table 1 where class 1 represents the DTRs with the low polymorphism while classes 2 and 3 represent the high-polymorphic and very high-polymorphic DTRs respectively.

3 Determination of Features Affecting Polymorphism

It has been a long run for researchers to identify the factors that affect the polymorphism of microsatellites. The underlying mutational process may be responsible for the high variability of alleles of microsatellites. Microsatellites have been estimated to mutate at a rate of between 10^2 and 10^5 mutations per gamete ([9, 26]). However, the mutational mechanisms are complex and still poorly understood. Two main mechanisms have been suggested to explain the high mutation rate of microsatellites. The first is recombination between DNA molecules by unequal crossing-over or by gene conversion ([37, 18]). The second mechanism involves slipped-strand mispairing during DNA replication ([23, 36]). There are other analysis contributing to the variation and mutation of microsatellites ([25, 6, 12, 41, 10, 16]). In summary, it has been observed by researchers that the length and composition of a microsatellite repeat may have an effect on its polymorphism. Microsatellites with a larger number of repeats are more polymorphic than those with a smaller number, and microsatellites with a high AT content are more polymorphic than those with a GC content. Complexity of repeat sequences may also affect the polymorphism of microsatellites. That is, the microsatellites that contain interruptions within the repeat or contain more than one type of repeat are less polymorphic than those simple microsatellites. In addition, there are also evidences that the polymorphism of a microsatellite may be correlated with its flanking sequence ([34, 20, 33]). Despite the above preliminary conclusions, it is still difficult for us to list all determinant factors for the polymorphism of microsatellites due to the complex mutation pattern of microsatellites.

Our work begins with the investigation of the polymorphism factors of DTRs by applying machine learning to the known polymorphic DTRs — Genethon markers. For each marker in G , the features listed in Table 2, denoted as F , can be computed from the DNA sequence of the marker and the output of TRF program, and are considered to possibly determine its heterozygosity. Among these features, f_1 , f_2 and f_3 (representing repeat complexity together), and f_4 have been suggested in the literature to be real polymorphism factors, while f_5 , f_6 , f_7 and f_8 are just guesses based on our experience. The task is then to choose some subset of features from F that can best distinguish high-polymorphic DTRs from low-polymorphic ones. This is a typical feature selection task in the field of machine learning. In our work, each input feature is normalized to zero mean and unit standard deviation.

To select the relevant polymorphism features, we applied K-Nearest Neighbor (K-NN)([28]) method which has been proven to be quite effective when it is provided with a sufficiently large set of training data. Our feature selection algorithm aims to select those features that can best separate classes in G from each other, and takes into account possible dependencies

f_1	repeat length
f_2	percentage of matches between adjacent copies overall
f_3	the alignment score (obtained from TRF)
f_4	percentage composition of "A+T"
f_5	percentage composition of "A+T" of its flanking sequence
f_6	the number of short tandem repeats found in its flanking sequence
f_7	the length of the longest contiguous "C+G" content in its flanking sequence
f_8	the length of the longest contiguous "A+T" content in its flanking sequence

Table 2: Features suggested to affect the polymorphism of a microsatellite

Target class	Selected features	Highest prediction accuracy (%)	K value
Class 1	$\{f_1, f_2\}$	98.5	50
Class 3	$\{f_1, f_3, f_5\}$	62.7	15
Overall	$\{f_1, f_2, f_3, f_5, f_7\}$	51.5	25

Table 3: Prediction accuracy of K-NN classification with selected features on testing data

among the features. Specifically, we are most interested in those features that best predict high heterozygosity in Genethon markers and those that best predict low heterozygosity in Genethon markers. In addition, we want to ensure that the heterozygosities of our high-polymorphic DTR candidates are as large as possible. Hence, we only find features that can best predict very high-polymorphic DTRs (i.e., class 3 in G) and low-polymorphic ones (class 1 in G). In our approach, we select these two sets of features separately since they are meant for separate tasks. A number of feature selection algorithms have been proposed in the field of machine learning ([17, 21]). We used a modified version of the sequential forward selection method (SFS) ([19]).

To select the features for predicting high heterozygosity in Genethon markers, the algorithm basically runs as follows: We randomly choose 20% markers of class 3 as testing examples. For the training examples, we collect an equal number (as many as possible) of markers from each of the other two classes as well as the remaining 80% of class 3. Then, we iteratively test the prediction accuracy (the ratio of the number of testing examples in class 3 that are correctly classified) of a K-NN classifier (K value can be adjusted) with a candidate feature vector decided by us. The classifier uses the standard Euclidean distance as its distance measure. The algorithm starts with a single candidate feature f_1 since we are certain that this feature affects the polymorphism of STRs ([23, 5, 30]). The candidate feature vector in the next iteration is decided by adding one more feature to the previous candidate feature vector such that the prediction accuracy acquired by the corresponding K-NN classifier is the highest. The algorithm stops if no more features can be added into the candidate feature vector such that the prediction accuracy can be improved. In this way, the final candidate feature vector contains all selected features that may predict high heterozygosity of DTRs. The algorithm works similarly to select the features for predicting low heterozygosity of DTRs. Furthermore, we also run the algorithm to select the features that can best discriminate the 3 classes overall (The prediction accuracy is the ratio of the total number of the testing examples in all 3 classes that are correctly classified). We present the selected features with the highest classification accuracies in Table 3.

The results in Table 3 show that 98.5% of the markers with low heterozygosity can be distinguished by two features: repeat length and percentage of matches between adjacent repeat copies (which can be taken as one measure of repeat complexity) whereas 62.7% of the markers with

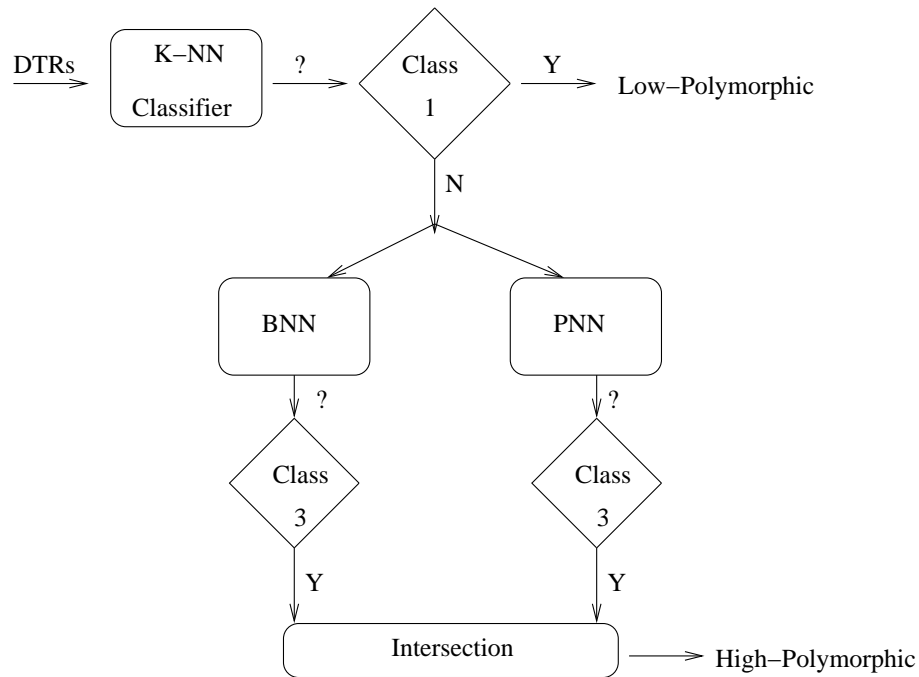


Figure 4.1: Machine learning algorithm for prediction of polymorphic DTRs

high heterozygosity can be identified by using three features: repeat length, alignment score of repeat copies (which can be another measure of repeat complexity) and AT content of flanking sequence. This suggests that the flanking sequence may play a significant role in the high variability of microsatellites. For the “overall” target class, one more factor from flanking sequence, length of the longest contiguous CG content in flanking sequence, may affect the polymorphism of microsatellites.

4 Prediction of Polymorphism of Novel DTRs

Artificial neural network (ANN) ([28]) is one of the widely used machine learning approaches in bioinformatics and it is also the earliest technique applied to the field of biological analysis ([39]). In this section, we apply a hybrid approach of K-NN and ANN to predict the polymorphism of novel DTRs using the polymorphism features determined in the previous section. The candidate high-polymorphic or low-polymorphic DTRs are then verified in wet lab.

4.1 Prediction algorithm

We propose a machine learning algorithm for the prediction of high-polymorphic DTRs, denoted as MLAP, shown in Figure 4.1. Given a set of novel DTRs, as K-NN method achieves a high accuracy for prediction of low-polymorphic DTRs according to Table 3, MLAP first filters out the low or possible non-polymorphic ones from the input set using a K-NN classifier. Those DTRs which are classified as class 2 and 3 by K-NN classifier are fed to ANNs for further identification. To increase the confidence of the prediction, MLAP employs two ANN classifiers: backpropagation neural network (BNN) ([31]) and probabilistic neural network (PNN) ([38]), which differ in the adopted learning algorithms. Only those DTRs that are classified as class 3 by both networks are reported as high-polymorphic ones. Again, this is to ensure that the heterozygosities of our high-polymorphic DTR candidates are as large as possible. Before we use the ANNs to do practical prediction, we first train and test them using the same training and

Target class	Input features	ANNs	Test accuracy rate (%)
class 3	$\{f_1, f_3, f_5\}$	BNN	64.7
		PNN	60.3

Table 4: Prediction accuracy of ANN classification with determined features on testing data

Chromosome	Number of contigs	Total sequence length (kb)	Number of DTRs	Density of DTRs (/kb)
21	5	34704	2640	0.076
22	10	34232	1927	0.056
12	18	10468	789	0.075

Table 5: Number of DTRs detected in source data by TRF program

testing examples as for K-NN method in Section 3. Details about the structure and algorithm used for two ANNs are presented below where the ANNs were implemented using MATLAB Neural Network toolbox.

BNN We adopt a two-layer backpropagation neural network which is trained by the steepest descent algorithm with adaptive learning rate and momentum training. Neurons in both the hidden layer and the output layer use the sigmoid transfer function. In our experiments, we used MATLAB’s default training parameters except for setting the momentum coefficient to 0.9. The highest prediction accuracy is achieved when the hidden layer contains 20 neurons.

PNN Probabilistic neural network is radial basis network and can be used for classification problems. The performance of a PNN varies with its parameter *spread*. In our experiments, we set all training parameters as MATLAB’s default. The highest prediction accuracy is achieved when spread is set to 1.

The test accuracy result for prediction of very high-polymorphic Genethon markers using the two networks is presented in Table 4.

4.2 Experimental Dataset

From the Genbank website ([2]), we downloaded the completed draft DNA sequences (build 26) of human chromosome 21 and 22. As our research involves finding myopia-related genes, we also downloaded one DNA segment of chromosome 12 (build 29) ranging from 83 Mb to 95 Mb which has been considered as a critical region where candidate genes responsible for myopia may be found. (A suspected gene decorin associated with myopia has been reported in [45] and is located in this region.) Then we performed TRF (with the input parameters as 2, 7, 7, 80, 10, 3, 2) to detect DTRs within these DNA sequences. The respective number of novel DTRs detected in each chromosome is summarized in Table 5. Note that due to the size limitation of input sequence (< 5 Mb) for TRF, we have splitted those contigs larger than 5 Mb into shorter and overlapped segments. Our task is now to apply MLAP to predict the polymorphism of 3 set of DTRs.

4.3 Results and Validation

As described above, we have taken 3 sets of DTRs found in chromosomes 21, 22 and 12 respectively as input to the MLAP. The number of predicted high-polymorphic and low-polymorphic candidate DTRs is summarized in Table 6 where “HP” represents high-polymorphic while “LP” represents low-polymorphic. Due to resource limitation, we are unable to verify all candidates through experiments in wet lab. Thus, we randomly chose 9, 8, and 25 DTRs from 3 “HP” candidate sets and 10, 7 and 15 DTRs from 3 “LP” candidate sets for wet lab testing.

Data source	Number of predicted candidates of MLAP		Number of tested candidates		Number of validated prediction		Prediction accuracy of MLAP (%)	
	HP	LP	HP	LP	HP	LP	HP	LP
Ch21	94	1077	9	10	9	10	100	100
Ch22	74	756	8	7	8	3	100	43
Ch12	31	304	25	15	22	7	88	47
Total	199	2137	42	32	39	20	93	63

Table 6: Prediction results of MLAP and the wet lab validation results

Oligonucleotide primers for these selected candidates were designed using the Primer3 program ([3]). PCR products were double-strand labeled by incorporation of fluorescent nucleotides. DTR sizing on an automated capillary sequencer (Megabase, MolecularDynamics) was performed. For a Chinese population of 24 individuals, we obtained the heterozygosity of each tested DTR. From Table 6, MLAP achieves a high accuracy of 93% in predicting high-polymorphic DTRs (this is the ratio of the number of DTRs with heterozygosity higher than 50% over the total number of tested candidates in “HP”). The average heterozygosity of the 39 high-polymorphic DTRs predicted by MLAP is 74% ($SD \pm 11.5\%$) (see Appendix A—Table 7(a) and 7(b)). In addition, MLAP’s accuracy in predicting low-polymorphic DTRs is 63% (this is the ratio of the number of DTRs with heterozygosity lower than 50% over the total number of tested candidates in “LP”). The average heterozygosity of these 20 low-polymorphic DTRs is 13% ($SD \pm 14.4\%$) (see Appendix B—Table 8).

From the wet lab validation results, we observe that MLAP achieves a much lower accuracy for low-polymorphic DTRs compared to the KNN results in Table 3. This could be due to the following reasons:

- (i) Instead of learning from both polymorphic and non-polymorphic (with a heterozygosity of 0%) DTRs, MLAP learns from the Genethon markers which are all polymorphic (the average heterozygosity is 70%) while a large number of novel DTRs are non-polymorphic in the real world.
- (ii) The Genethon markers are derived from a Caucasian population while our wet lab tests are based on a Chinese population. It is known that the degree of polymorphism of a STR may vary with different populations [40].

We believe that the prediction of MLAP would become more precise with the availability of more diverse training data in the future.

5 Conclusion and Future Work

STRs are widely used as genetic markers for DNA profiling of individuals and for mapping genes for diseases and traits. DTRs in particular are useful for genetic linkage and association studies because of their potentially high polymorphic content, as well as their relative abundance. Currently only about 3% of all DTRs have been documented for use as markers. We have applied a machine learning approach, MLAP, to predict the polymorphism of a DTR. This approach can be directly applied to prioritize among the DTRs to be experimentally validated as genetic markers, and to help construct a denser genetic map faster and more efficiently. With the fully sequenced human genome, this would be useful for haplotype mapping and genomewide linkage disequilibrium studies of complex diseases. In our future work, we may improve MLAP in the following aspects:

- (i) Currently, MLAP is based on a hybrid approach where one K-NN classifier and two ANNs are utilized. We may add more classifiers such as support vector machines (SVMs) ([42]) and other ANNs with various learning algorithms and parameters to improve the prediction accuracy of MLAP.
- (ii) Instead of using the intersection results from two ANNs to predict the high-polymorphic DTRs, we may use soft voting from multiple classifiers to determine whether a given DTR is polymorphic. In the soft voting method, the prediction result of each classifier contains a voting value (say, from 0 to 1) associated with each class rather than a single class label.
- (iii) The training data sets for classifiers can be expanded to contain the validated true non-polymorphic DTRs so as to improve the accuracy for predicting non-polymorphic DTRs.

We have also used a feature selection method to identify the factors that may affect the polymorphism of microsatellites. Our work suggests that one additional factor, AT content of the flanking sequence of a repeat, which has never been found in the previous research, affects the high polymorphism of a microsatellite. We may study further to investigate its effect. On the other hand, while the current feature selection method SFS is simple and straightforward, we may use more sophisticated methods such as the sequential forward/backward floating search methods (SFFS/SBFS) ([29]), and the genetic algorithm (GA) ([15]) in the future. Hopefully, the machine learning approach may help biologists to understand the real mutation mechanism for microsatellites.

Currently, our approach only applies to DTRs. However, with the public data available, it can be easily extended to predict the polymorphism of other types of microsatellites such as tri-nucleotide tandem repeats and tetra-nucleotide tandem repeats.

References

- [1] <http://www.genethon.fr/index.html>.
- [2] <http://www.ncbi.nlm.nih.gov/>.
- [3] http://www-genome.wi.mit.edu/genome_software.
- [4] BENSON, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 27(2):573–580, 1999.
- [5] BOWCOCK, A.M., RUIZ-LINARES, A., TOMFOHRDE, J., MINCH, E., KIDD, J. R., CAVALLI-SFORZAK, L. L. High Resolution of Human Evolutionary Trees with Polymorphic Microsatellites. *Nature*, 368:445–457, 1994.

- [6] BULL, L.N., PABON-PENA, C.R., FREIMER, N.B. Compound microsatellites repeats: practical and theoretical features. *Genome Research*, 9:830–838, 1999.
- [7] DIB, C., FAURE, S., FIZAMES, C., SAMSON, D., DROUOT, N., VIGNAL, A., MILLASSEAU, P., ET AL. A comprehensive genetic map of the human genome based on 5264 microsatellites. *Nature*, 380:152–154, 1996.
- [8] DOKHOLYAN, N.V., BULDYREV, S.V., HAVLIN, S., STANLEY, H.E. Distributions of dimeric tandem repeats in non-coding and coding DNA sequences. *J. theor. Biol.*, 202:273–282, 2000.
- [9] EDWARDS, A., HAMMOND, H.A., JIN, L., CASKEY, C.T., CHAKRABORTY, R. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics*, 12:241–253, 1996.
- [10] ELLEGREN, H. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genet.*, 16:551–558, 2000.
- [11] FONDON III, J., MELE, G., BREZINSCHKE, R., CUMMININGS, D., PANDE, A., WREN, J., O'BRIEN, K., KUPFER, K., WEI, M.H., LERMAN, M., MINNA, J., GARNER, H. Computerized polymorphic marker identification: Experimental validation and a predicted human polymorphism catalog. *Proc. Natl. Acad. Sci. USA*, 95:7514–7519, June 1999.
- [12] GUR-ARIE, R., COHEN, C.J., EITNA, Y., SHELEF, L., HALLERMAN, E.M., KASHI, Y. Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism. *Genome Research*, 10:62–71, 2000.
- [13] GYAPAY, G., MORISSETTE, J., VIGNAL, A., DIB, C., FIZAMES, C., MILLASSEAU, P., MARC, S., ET AL. The 1993-1994 Genethon human genetic linkage map. *Nat Genet*, 7:246–339, 1994.
- [14] HAUTH, A.M., JOSEPH, D.A. Beyond tandem repeats: complex pattern structures and distant regions of similarity. In *Proceedings of the Tenth International Conference on Computational Molecular Biology*, pages S31–S37, 2002.
- [15] HOLLAND, J. *Adaptation in natural and artificial systems*. University of Michigan Press, 1975.
- [16] HUANG, Q.Y., XU, F.H., SHEN, H., DENG, H.Y., ET AL. Mutation patterns at dinucleotide microsatellite loci in humans. *Am. J. Hum. Genet.*, 70:625–634, 2002.
- [17] JAIN, A., ZONKER, D. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, 1997.
- [18] JEFFERYS A.J., TAMAKI, K., MACLEOD, A., MONCKTON, D.G., NEIL, D.L., ARMOUR, J.A.L. Complex gene conversion events in germline mutation at human minisatellites. *Nature Genetics*, 6:136–145, 1994.
- [19] KITTLER, J. Feature set search algorithms. In *Pattern Recognition and Signal Processing*, C.H.Chen, Ed., pages 41–60, Sijthoff and Noordhoff, Netherlands, 1978.
- [20] KRUGLYUAK, S., DURRETT, R.T., SCHUG, M.D., AQUADRO, C.F. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl. Acad. Sci. USA*, 95:10774–10778, 1998.
- [21] KUDO, M., SKLANSKY, J. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33:25–41, 2000.
- [22] LANDAU, G.M., SCHMIDT, J.P. AND SOKOL, D. An algorithm for approximate tandem repeats. *Journal of Computational Biology*, 8(1):1–18, 2001.
- [23] LEVINSON, G., GUTMAN, G.A. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Mol. Biol. Evol.*, 4:203–221, 1987.
- [24] LITT, M., LUTY, J.A. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.*, 44:397–401, 1989.
- [25] MACAUBAS, G., JIN, L., HALLMAYER, J., KIMURA, A., MIGNOT, E. The complex mutation pattern of a microsatellites. *Genome Research*, 7:635–641, 1997.

- [26] MAHTANI, M.M., WILLARD, H.F. A polymorphic X-linked tetranucleotide repeat locus displaying a high rate of new mutation: Implications for mechanisms of mutation at short tandem repeat. *Hum. Mol. Genet*, 2:431–437, 1993.
- [27] MCPHERSON, J.D., MARRA, M., HILLIER, L., WATERSTON, R.H., CHINWALLA, A., WALLIS, J., ET AL. A physical map of the human genome. *Nature*, 409:934–941, 2001.
- [28] MITCHELL, T.M. *Machine Learning*. The McGraw-Hill Companies, Inc., 1997.
- [29] PUDIL, P., NOVOVICOVA, J., KITTLER, J. Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125, 1994.
- [30] RUBINSZTEIN, D.C., AMOS, W., LEGGO, J., GOODBURN, S., JAIN, S., LI, S.H., MARGOLIS, R.L., ROSS, C.A., FERGUSON-SMITH, M.A. Microsatellite evolution-evidence for directionality and variation in rate between species. *Nat Genet*, 11(4):360–2, Dec 1995.
- [31] RUMELHART, D.E., HINTON, G.E., WILLIAMS, R.J. Learning internal representation by error propagation. *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, 1, 1986.
- [32] SAGOT, M., MYERS, E. Identifying Satellites and Periodic Repetitions in Biological Sequences. In *Proceedings of the Second Annual International Conference on Computational Molecular Biology*, pages 234–242, 1999.
- [33] SANTIBANEZ-KOREF, M., GANGESWARAN, R., HANCOCK, J.M. A relationship between lengths of microsatellites and nearby substitution rates in mammalian genomes. *Mol. Biol. Evol*, 18:2119–2123, 2001.
- [34] SANTIBANEZ-KOREF, M., HANCOCK, J.M. Conservation of sequences flanking microsatellites. *Am J Hum Genet*, 63:A258, 1998.
- [35] SCHLOTTERER, C. Evolutionary dynamics of microsatellite DNA. *Chromosoma*, 109:365–371, 2000.
- [36] SCHLOTTERER, D., TAUTZ, D. Slippage synthesis of simple sequence DAN. *Nucleic Acid Research*, 20:211–215, 1992.
- [37] SMITH, G.P. Evolution of repeated DNA sequences by unequal crossover. *Science*, 191:528–535, 1976.
- [38] SPECHT, D.F. Probabilistic neural networks and the polynomial adaline as complementary techniques for classification. *IEEE Transactions on Neural Networks*, 1(1):111–121, 1990.
- [39] STORMO, G., SCHNEIDER, T., GOLD, L., EHRENFEUCHT, A. Use of the perceptron algorithm to distinguish translational initiation in E.coli. *Nucleic Acids Research*, 10(10):2997–3011, 1982.
- [40] TAN, E., WU, H., YONG, R., TAN, S., CHANG, J., GAN, L., YAP, E. short communication: Heterozygosities and allelic frequencies of a set of microsatellite markers used for genome-wide scans in a Chinese population. *Journal of Human Genetics*, 47(11):623–631, Nov 2002.
- [41] TOTH, G., GASPARI, Z., JURKA, J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research*, 10:967–981, 2000.
- [42] VAPNIK, V.N. *The Nature of Statistical Learning Theory*. New York: Springer, 1995.
- [43] WEBER, J.L., MAY, P.E. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet*, 44:388–396, 1989.
- [44] WEISSENBACH, J., GYAPAY, G., DIB, C., VIGNAL, A., MORISSETTE, J., MILLASSEAU, P., VAYSSEIX, G., ET AL. A second-generation linkage map of the human genome. *Nature*, 359:794–801, 1992.
- [45] YOUNG, T.L., RONAN, S.M., ALVEAR, A.B., WILDENBERG, S.C., OETTING, W.S., ATWOOD, L.D., WILKIN, D.J., KING, R.A. A second locus for famsial high myopia maps to chromosome 12q. *Am. J. Hum. Genet*, 63:1419–1424, 1998.

Appendix A - Validated High-Polymorphic DTRs in Wet Lab

All predicted high-polymorphic DTRs by MLAP that have been validated in wet lab are listed in Tables 7(a) and 7(b). Note that in Table 7(a), ch22M8 has been found to be a known Genethon marker.

Marker	Location (chromosome kb)	Type	Heterozygosity (%)
ch21M1	ch21 11770	GA	76
ch21M2	ch21 16279	CA	62
ch21M3	ch21 16120	CA	79
ch21M4	ch21 27505	GA	76
ch21M5	ch21 30406	CA	90
ch21M6	ch21 33367	GA	83
ch21M7	ch21 30046	CA	85
ch21M8	ch21 36914	AT	88
ch21M9	ch21 39887	CA	62
ch22M1	ch22 40982	AT	89
ch22M2	ch22 47103	GT	71
ch22M3	ch22 21203	CA	85
ch22M4	ch22 33814	GA	58
ch22M5	ch22 34236	GA	90
ch22M6	ch22 32572	CA	83
ch22M7	ch22 34260	GA	79
ch22M8	ch22 37031	CA	52

(a) chromosome 21 and 22 DTRs

Marker	Location (chromosome kb)	Type	Heterozygosity (%)
ch12M1	ch12 82338	CA	77
ch12M2	ch12 82138	CA	82
ch12M3	ch12 82424	CA	63
ch12M4	ch12 83182	GA	51
ch12M5	ch12 83159	CA	72
ch12M6	ch12 83152	GA	81
ch12M7	ch12 83578	CA	78
ch12M8	ch12 85643	CA	54
ch12M9	ch12 85775	CA	79
ch12M10	ch12 85660	CA	79
ch12M11	ch12 84944	CA	64
ch12M12	ch12 85433	CA	57
ch12M13	ch12 85262	CA	72
ch12M14	ch12 85338	AT	84
ch12M15	ch12 86690	CA	52
ch12M16	ch12 86795	CA	72
ch12M17	ch12 86284	CA	73
ch12M18	ch12 93464	CA	76
ch12M19	ch12 82281	AT	87
ch12M20	ch12 85767	GA	93
ch12M21	ch12 84915	CA	68
ch12M22	ch12 86317	GA	71

(b) chromosome 12 DTRs

Table 7: Validated high-polymorphic DTRs predicted by MLAP

Appendix B - Validated Low-Polymorphic DTRs in Wet Lab

All predicted low-polymorphic DTRs by MLAP that have been validated in wet lab are listed in Tables 8.

Marker	Location (chromosome kb)	Type	Heterozygosity (%)
ch21LP_M1	ch21 19215	GA	23
ch21LP_M2	ch21 17388	CA	35
ch21LP_M3	ch21 29275	CA	43
ch21LP_M4	ch21 44383	CA	0
ch21LP_M5	ch21 21266	CA	0
ch21LP_M6	ch21 25328	CA	10
ch21LP_M7	ch21 25534	CA	0
ch21LP_M8	ch21 25576	CA	4
ch21LP_M9	ch21 36055	CA	0
ch21LP_M10	ch21 39725	CA	0
ch22LP_M1	ch22 42519	CA	21
ch22LP_M2	ch22 46663	CA	26
ch22LP_M3	ch22 40097	CA	0
ch12LP_M1	ch12 90523	CA	35
ch12LP_M2	ch12 85960	CA	17
ch12LP_M3	ch12 85827	CA	29
ch12LP_M4	ch12 87095	CA	4
ch12LP_M5	ch12 89485	GA	4
ch12LP_M6	ch12 90130	CA	0
ch12LP_M7	ch12 93386	CA	0

Table 8: Validated low-polymorphic DTRs predicted by MLAP