Supervised Classification of Genetic Sequences for Population Analysis

Ljubomir Buturović[*] Sarah Cohen[†] Zhihong He
Martin Eggenberger Diane Nacci[‡]
Dragutin Petković
Department of Computer Science, Department of Biology[†], Romberg Tiburon Center[†]
San Francisco State University
US EPA, Office of Research and Development, NHEERL, Atlantic Ecology Division[‡]

Abstract

We used a Support Vector Machine (SVM) algorithm for analysis of biological populations through supervised classification of sequence data. We describe an open-source software which implements the method, and apply the method and the program to analyze environmentally challenged populations of the estuarine fish Fundulus heteroclitus.

Specifically, we investigate whether the genetic composition (DNA sequence) of a particular detoxification locus predicts population assignment of fish to chemically contaminated versus clean estuaries. The analysis method uses an SVM algorithm to assign individual fish, characterized by their allelic composition, into a toxic-resistant or non-resistant group. We employed classification error in assignment as a measure of population similarity. The results validate the proposed method by providing supporting evidence for the previously suggested role of AHR1 (aryl hydrocarbon receptor) locus in the toxic response pathway of Fundulus heteroclitus.

Keywords - haplotype analysis, population genetics, machine learning, Support Vector Machine

[*] The corresponding author. Mailing address: San Francisco State University, 1600 Holloway Avenue, San Francisco, CA 94132. E-mail: ljubomir@sfsu.edu.

1 Introduction

Population genetics seeks to analyze genetic differences within and among species and make inferences about the evolutionary process. It has utilized a wide variety of standard and specialized statistical and sequence analysis methods to explain genetic diversity [6]. Recently, classic meachine learning methods (artificial neural networks, decision trees and k-nearest neighbor classifiers) have been introduced to solve the problem of assigning individuals to their populations of origin [16]. We developed a Support Vector Machine (SVM)-based method and a software package for population analysis.

The motivation for the use of SVM in the analysis of genetic populations is based on the following properties of the algorithm:

- demonstrated ability of SVM to discover highly non-linear relationships among input variables [3], [15], with potential relevance for the analysis of complex traits
- in contrast to other statistical pattern classification algorithms (k-nearest neighbors, artificial neural networks, linear discriminants), there is no requirement that the input data must be embedded in a vector space [20]. This permits the use of domain-specific similarity measures between input samples, e.g. Smith-Waterman scores, which may be better able to detect subtle sequence signals.
- recent advances in the application of SVM in computational biology [17].

The specific problem we studied concerns adaptation of the estuarine fish *Fundulus heteroclitus* to adverse environmental conditions in the form of severe chemical contamination in estuarine waters. *Fundulus heteroclitus* is a widely used model organism in population genetics and toxicology [1], [12], [18]. Previous work [12], [2], [9] has implicated the role of the *Aryl hydrocarbon receptor (AHR1)*, a transcription control factor in the detoxification pathway, as a possible agent in the evolution of extreme chemical tolerance in this species. Thus, an interesting and relevant population genetics problem is to analyze the relationship between AHR1 mutations and ability of fish to survive in an adverse environment. We propose to analyze and quantify the relationship using supervised classification of the individual fish into two categories (classes), corresponding to the two phenotypes being studied: resistant, and susceptible to extreme chemical contamination.

We chose to represent the fish samples using DNA sequences consisting of nucleotides at selected AHR1 SNP locations, and to apply the SVM supervised classification algorithm to predict fish population assignment (i.e., assign fish specimen into one of the populations). We reason that the error rate of this assignment may reflect the overall similarity of local populations in terms of AHR1 SNP sequence information, with the expectation that similar populations will be harder to classify (i.e., will exhibit higher classification error rate). Consequently, we decided to use the estimated cross-validation error rate of the assignment to characterize the relationship among populations.

The paper is organized as follows. Section 2 provides a background on the applications of SVM algorithms in sequence analysis, and the proposed use of SVM in the study of genetic populations. Section 3 contains a detailed statement of the problem and results.

2 Methods

2.1 Machine Learning and Support Vector Machine Algorithm for Sequence Analysis

First we provide a background on the Support Vector Machine (SVM) pattern classification algorithm and its applications in the analysis of biological sequences.

The Support Vector Machine algorithm is a pattern classifier which predicts classification of an input sample x according to the following equation (for a two-class problem):

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{n} y_i \alpha_i k(x, x_i) + b) \qquad (1)$$

where x_i are the training set samples, y_i is $+1$ for class 1 samples, -1 for class 2 samples, and $k(x, x_i)$ is a *kernel* function, which generally has the meaning of similarity among x and x_i. Sample x is assigned to class 1 if $f(x)$ equals 1, and to class 2 otherwise. In this equation, the coefficients $\alpha_i \geq 0$ are positive for a subset of the training data set samples called *Support Vectors*. Thus, only the Support Vectors are used to classify an unlabeled sample x. The coefficients are result of a quadratic optimization procedure of a suitable criterion defined as ([15]):

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j), \qquad (2)$$

subject to $0 \leq \alpha_i \leq \frac{C}{n}$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$. Eq. (2) is a *dual* representation of the criterion that the separable training samples be separated by as wide a margin as possible, and that the non-separable ones be as few as possible. The relative weight of these two requirements is controlled by the user-defined parameter C.

A widely used kernel function for data embedded in a vector space is the *Radial Basis Function* (RBF):

$$k(x, y) = \exp(-\gamma \|x - y\|^2) \qquad (3)$$

A key property of the algorithm, as can be seen from Eq. (1), is that the input data x and x_i are used in the classifier *only* through the kernel function $k(x, x_i)$. Therefore, as long as a meaningful and computable function measuring input point similarity can be devised, the inputs can be arbitrary objects. This is in contrast with the majority of other pattern classification methods which require that the inputs be embedded in a vector space, thus forcing a sometimes artificial encoding of the input data.

In order to apply SVM to the classification of DNA or protein sequences, it is sufficient to define a kernel function. At first, it may be tempting to employ a natural measure of similarity for biological sequences such as a local alignment score. It is however easy to verify that the measure is not a positive-definite kernel, and thus search for alternatives has been focus of intense research in recent years. In this paper, we used the *empirical kernel map* ([17]) to convert the input sequence data into vector format suitable for use in standard kernels such as (3).

2.2 Analysis of populations using SVM

In this section we describe the method developed to compare fish populations experiencing different levels of exposure to chemical contaminants.

As indicated earlier, we assign each input specimen into one of the two populations on the basis of its genotype sequence, using the SVM classifier. We propose to quantify the similarity of populations by the prediction error rate of the classifier, and make population inferences using this measure. The error rate estimate is the value achieved by the optimal choice of SVM parameters C and γ. The optimum is found by estimating the SVM cross-validation error rate for a range of values of the parameters, and using the lowest error rate.

In this paper, we analyze data obtained from diplotype sequences produced by direct sequencing of PCR products. To reconstruct alleles required for SVM analyses, we used the haplotype reconstruction program PHASE [8]. Note that the principle of quantifying population relationships using classifier prediction accuracy is not limited to analysis of reconstructed haplotypes; indeed, it could be used for the analysis of experimentally obtained allele sequences.

Each fish specimen has originally been represented by the DNA sequence of 12 segregating sites (SNP locations) per allele from a portion of the AHR1 locus (originally 361 bp long). The population allele sequences have subsequently been processed by program `PHASE` (version 2.0.2, [8]) to obtain haplotype estimates. The two haplotype estimates per specimen were concatenated into a 24-nucleotide long sequence, which was then used as input to the machine learning software.

We used the RBF kernel (3) for all analyses in this paper. We used two approaches to encode input sequences: decimal encoding, and empirical kernel map.

Decimal encoding assigns numeric values $\{1, 2, 3, 4\}$ to DNA nucleotides `A, T, C, G`, respectively.

Empirical kernel map represents each sequence with a vector of local alignment scores against all sequences in the learning data set. Thus, for input sequence S and learning data set of size N, the empirical kernel map representation of S is numeric vector x defined as:

$$x = [d_1 d_2 \dots d_N] \tag{4}$$

where d_i is local alignment score between S and sequence S_i in the training set. Note that the kernel map, unlike decimal encoding, permits analysis of sequences of unequal lengths (i.e, sequences with insertions/deletions).

2.3 Software

We developed an open-source C++ program named `PhaseMachine` for supervised classification of the haplotype estimates produced by `PHASE`, using the Support Vector Machine algorithm. The software is a command-line program for GNU/Linux and Windows operating systems, and incorporates a widely used open-source SVM implementation LIBSVM [19]. It accepts files produced by `PHASE` as input, builds SVM models, and produces classification performance report. The program supports modeling, prediction, and cross-validation modes, and incorporates decimal and empirical kernel map encoding of input sequences.

3 Results

We analyzed eight populations of *Fundulus heteroclitus* collected in geographical areas shown in Fig. 1. The population characteristics are shown in Table 1. The subset of 8 populations was chosen according to geographical location, number of clean replicate populations sampled in close proximity, and type of contamination. We sought to control, as much as possible, variation due to mixtures of contaminants which might produce conflicting or varying signals of genetic variation in AHR1, exon 10. We analyzed populations from two distinct geographical areas, Massachusetts and Long Island Sound.

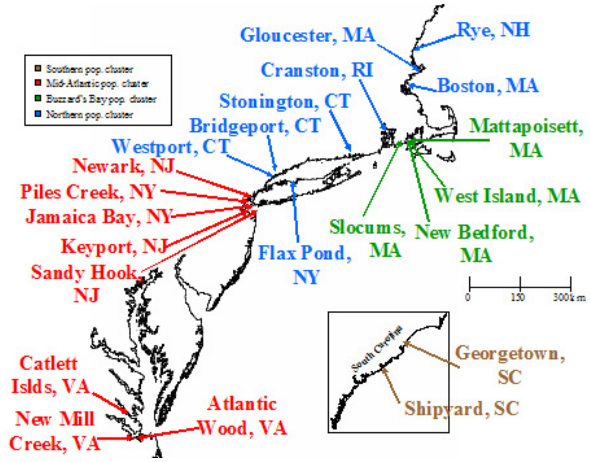

Figure 1: Geographical distribution of fish populations sampled for AHR1 DNA sequence variation ([12], [21]). A subset of the populations, described in Table 1, was used in the present study.

All experiments were conducted using `PhaseMachine` software described in Section 2.3.

First, we calculated within-cluster pairwise cross-validation error rates of the C-SVM classifier, using RBF kernel and decimal encoding of the nucleotide sequences. The results are shown in Table 2.

The upper section of Table 2 contains results for the Long Island Sound cluster, while the lower section has results for Massachusetts. The misclassification error rate between two clean populations (WP, FLAX) is much higher (34.7%) than the misclassification error rate between a dirty population (BP) and either of the two clean populations (9.6% and 15.2%, respectively).

The Massachusetts cluster presents a more complicated analysis because of a greater geographic signal within this cluster. Glouces-

Population	*Location*	*Contaminated*	*Number of individual fish used in this study*
BP	Bridgeport, CT	Yes	26
FLAX	Flax Pond, NY	No	15
WP	Westport, CT	No	27
NBH	New Bedford, MA	Yes	26
GLO	Gloucester, MA	No	28
MAT	Mattapoisett, MA	No	29
SLO	Slocums, MA	No	27
WIC	West Island, MA	No	29

Table 1: Fish samples used for SVM test in this study. The first 3 populations are the Long Island cluster and the remaining 5 are a Massachusetts set. Contamination refers to measures of PCB contamination or evolved resistance measured by the US EPA NHEERL AED ([4], [11], [13], [7]). Haplotypes used in this study were defined using a PHASE 2.0 haplotype inference probability threshold of 0.6.

Population 1	*Population 2*	*Error rate*	*Population 1 error rate*	*Population 2 error rate*
BP	WP	9.6	17.9	1.1
BP	FLAX	15.2	17.9	11.1
FLAX	WP	34.7	67.2	13.0
NBH	GLO	9.1	7.3	10.7
NBH	WIC	26.5	39.2	15.2
NBH	MAT	24.7	33.5	16.9
NBH	SLO	27.9	33.8	22.2
SLO	GLO	16.5	15.9	17.1
WIC	GLO	18.8	15.2	22.5
MAT	GLO	15.8	6.9	25.0
SLO	WIC	32.9	30.7	34.8
SLO	MAT	37.3	38.5	36.2
WIC	MAT	47.4	56.6	38.3

Table 2: Within-cluster pairwise cross-validation classification error rates for populations defined in Table 1. All experiments were performed using RBF kernel. The values are percentage values averaged over 10 experiments for optimal choice of SVM parameters.

ter is located north of Cape Cod, a frequently cited potential biogeographic barrier [5], whereas the rest of the Massachusetts populations are all located south of the Cape and within Buzzard's Bay in much closer proximity. Except for the Gloucester population, the remaining clean populations (WIC, MAT, SLO) exhibit higher error rates among themselves than when compared with the dirty population (NBH), although the differences are less pronounced than in the Long Island Sound cluster. Excluding GLO, the highest dirty vs. clean error rate is 24.7% (NBH vs. MAT), while the lowest pairwise error rate among clean populations is 32.9% (SLO vs. WIC).

Our goal is to test for a relationship between AHR1 partial exon 10 SNP genotype and extreme chronic contaminant exposure. In case of physically distant populations, it is expected that this effect may be attenuated by the geographical signal, i.e. genetic divergence due to distance. To verify this, we computed cross-validation error rates for a set of populations from different clusters. The results are shown in Table 3.

As expected, the geographical isolation among the populations overshadows genetic changes related to environmental factors, resulting in uniformly low error rates among all pairs studied in this experiment.

Next we examined the effect of a different sequence encoding principle. Specifically, we used an *empirical kernel map* to convert input sequence strings into vector representation, as elaborated in Section 2.2. The results of this experiment are shown in Table 4.

The rationale for this experiment is that the domain-specific similarity scores built into the Smith-Waterman algorithm may be better able to highlight subtle sequence signals than the simple decimal encoding. First we computed cross-validation error rates for two clean/dirty population pairs (NBH/SLO and SLO/WIC) which appear to show the least clear distinction in the Massachusetts cluster. The results were somewhat inconclusive. In the original experiment, using decimal encoding of the DNA sequences, the gap between the two error rates is 5.0%; using the empirical kernel map, the gap rose insignificantly to 6.1%. Similarly, the differences in error rates between Long Island Sound cluster populations (BP/FLAX and FLAX/WP) for decimal and local alignment encoding were 19.5% and 17.1%, respectively.

4 Discussion

The goal of the paper is to examine the applicability of supervised learning using a Support Vector Machine algorithm for inferring population assignments for individuals based on their allelic composition. The results demonstrate that, for the set of populations we have analyzed, the pattern of the supervised classification cross-validation error of the population sequences generally reflects the expected trends, and is consistent with the hypothesized role of AHR1 in mediating toxicity.

To analyze populations of *Fundulus heteroclitus*, we first compared classification error rates between clean and dirty populations using sequence from a portion of the AHR1 lo-

Population 1	*Population 2*	*Error rate*	*Population 1 error rate*	*Population 2 error rate*
BP	NBH	8.1	5.0	11.5
BP	MAT	3.7	0.4	6.9
FLAX	NBH	9.1	5.6	11.5
FLAX	MAT	6.4	5.6	11.5
WP	NBH	7.7	4.1	11.5
WP	MAT	5.5	4.1	6.9

Table 3: Selected between-cluster pairwise cross-validation classification error rates.

Population 1	*Population 2*	*Error rate*	*Population 1 error rate*	*Population 2 error rate*
NBH	SLO	26.4	30.4	22.6
SLO	WIC	32.5	29.6	35.2
BP	FLAX	15.6	18.6	11.1
FLAX	WP	32.7	62.2	13.0

Table 4: Selected within-cluster error rates using local alignment encoding.

cus for the comparison. We observed that in all cases except one particularly distinct population (GLO, Gloucester, MA), the error rates between dirty and clean populations are lower than error rates between clean populations. The GLO population appears to be sufficiently remote from the other populations that the strength of geographical signal made it impossible to make other inferences using the proposed method.

Next, we compared classification error rates among populations from different geographical clusters. We found that, consistent with our expectations, there is a strong signal of genetic differentiation between regions that complicates detection of genetic correlations related to chemical stress tolerance.

Finally, we used a domain-specific sequence encoding method, based on the Smith-Waterman local alignment score, to enhance discrimination between genotypes. These experiments did not produce conclusive results and warrant further analyses.

In summary, we have demonstrated the presence of a discrimination signal among environmentally challenged populations of *Fundulus heteroclitus*, by applying supervised classification to the nucleotide sequences of a portion of the AHR1 locus. These results support the previously proposed role of the AHR1 gene in evolved tolerance of some populations to extreme chemical contamination. They also validate the proposed use of Support Vector Machine algorithm classification in population genetics.

Code Availability: The code used to obtain results reported in this paper has been released as an open-source project `PhaseMachine`. The program runs on Windows and Linux operating systems, and is available for download at `http://phasemachine.sourceforge.net`.

Acknowledgements

Fish samples and AHR1 sequence data were collected with support from The Hudson River Foundation (SC) and the US EPA STAR program (grant R82902201, SC), and with assistance from Denise Champlin (US EPA NHEERL Atlantic Ecology Division), Glen Yang (supported by an EPA UNEMS summer fellowship). Saro Jayaraman (US EPA NHEERL AED) carried out assays to characterize sediment contamination levels.

Although the research described in this contribution has been funded partially by the U.S. EPA, it has not been subjected to Agency-level review. Therefore, it does not necessarily reflect the views of the agency. Mention of trade names, products, or services does not constitute endorsement or recommendation for use.

References

[1] Brown, B. and R. Chapman, "Gene flow and mitochondrial DNA variation in the killifish, Fundulus heteroclitus," *Evolution* 45: 1147, 1991.

[2] Hahn, M.E., "Mechanisms of innate and acquired resistance to dioxin-like compounds," *Reviews in Toxicology* 2, 395-443, 1998.

[3] V. Cherkassky and F. Mulier, *Learning From Data: Concepts, Theory and Methods*. John Wiley & Sons, 1998.

[4] Nacci, D.E., Coiro, L., Champlin, D., Jayaraman, S., McKinney, R., Gleason, T.R., Munns, W.R., Specker, J.L., and Cooper, K.R., "Adaptations of wild populations of the estuarine fish Fundulus heteroclitus to persistent environmental contaminants," *Marine Biology*, 134:9-17, 1999.

[5] Engle,, V and J. Summers, "Latitudinal gradients in benthic community composition in Western Atlantic estuaries," *J. Biogeography* 26: 1007-1023, 1999.

[6] D. L. Hartl, *A Primer Of Population Genetics*. Sinauer Associates, Inc., 2000.

[7] Champlin, D., Nacci, D., Serbst, J., Jayaraman, S., Gleason, T., Rocha, K., Rego, S., and Coiro, L., "Characterizing Sites With Differing Environmental Quality That Serve As Habitats For The Estuarine Fish Fundulus heteroclitus," Society of Environmental Toxicology and Chemistry (SETAC) 21st annual meeting, Nashville Tennessee, Abstract Book. 12-16 November, 2000.

[8] M. Stephens, N. J. Smith, and P. Donnelly, "A New Statistical Method for

Haplotype Reconstruction from Population Data, " *Am. J. Hum. Genet.* 68:978 989, 2001.

[9] Bello, S.M., Franks, D.G., Stegeman, and J.J., Hahn, M.E., "Acquired resistance to Ah receptor agonists in a population of Atlantic killifish (Fundulus heteroclitus) inhabiting a marine Superfund site: in vivo and in vivo studies on the inducibility of xenobiotic metabolizing enzymes," *Toxicological Sciences,* 60:77-91, 2001.

[10] Cohen, S., "Strong positive selection and habitat-specific amino acid substitution patterns in Mhc from and estuarine fish under intense pollution stress," *Molecular Biology and Evolution,* 19:1870-1880, 2002.

[11] Nacci, D.E., Champlin, D., Coiro, L., McKinney, R., and Jayaraman, S., "Predicting the occurrence of genetic adaptation to dioxinlike compounds in populations of the estuarine fish Fundulus heteroclitus," *Environ. Tox. and Chem.*, 21(7):1525-1532, 2002.

[12] S. Cohen and D. Nacci, "Effects of dioxin-like compound (DLC) contamination on an estuarine fish species: adaptive changes at specific loci," Proc. US/Vietnam Scientific Conference on Agent Orange/Dioxins, March 3-6, 2002, Hanoi, Vietnam.

[13] G. Yang, "Analysis of evolved resistance and population genetic structure of the estuarine teleost Fundulus heteroclitus (Atlantic killifish) using AHR1, " B. A. thesis in Environmental Science and Public Policy, Harvard College, Cambridge, MA, 2003.

[14] Z. He, "Supervised classification of DNA sequence data from estuarine fish populations," M.Sc. thesis, Department of Computer Science, San Francisco State University, 2003.

[15] B. Schölkopf and A. Smola, *Learning with Kernels*. MIT Press, 2002.

[16] B. Guinand, A. Topchy, K. S. Page, M. K. Burnham-Curtis, W. F. Punch, and K. T. Scribner, "Comparisons of Likelihood and Machine Learning Methods of Individual Classification," *Journal of Heredity* 93(4), pp. 260-269, 2002.

[17] W. Noble, "Support Vector Machine Applications in Molecular Biology." in B. Schölkopf, K. Tsuda and J.-P. Vert (eds.), *Kernel Methods in Computational Biology*. MIT Press, 2004, pp. 71-92.

[18] Wirgin, I and J. Waldman, "Resistance to contaminants in North American fish populations." *Mutat. Res.* 552:73-100, 2004.

[19] C.-C. Chang and C.-J. Lin, "LIBSVM: a Library for Support Vector Machines," `http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf`, August 12, 2004.

[20] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[21] Cohen, S. et al, unpublished data.