Published in final edited form as:
Proc Int Conf Intell Syst Mol Biol. 2000 ; 8: 286–295.

Pattern recognition of genomic features with microarrays: site typing of *Mycobacterium tuberculosis* strains

Soumya Raychaudhuri, **Joshua M. Stuart**, **Xuemin Liu**, **Peter M. Small**, and **Russ B. Altman**
Department of Medicine (Stanford Medical Informatics and Division of Infectious Disease), Stanford University 251 Campus Drive, Medical School Office Building, X-215, Stanford, CA 94305-5479

Abstract

Mycobacterium tuberculosis (*M. tb.*) strains differ in the number and locations of a transposon-like insertion sequence known as IS6110. Accurate detection of this sequence can be used as a fingerprint for individual strains, but can be difficult because of noisy data. In this paper, we propose a non-parametric discriminant analysis method for predicting the locations of the IS6110 sequence from microarray data.

Polymerase chain reaction extension products generated from primers specific for the insertion sequence are hybridized to a microarray containing targets corresponding to each open reading frame in *M. tb*. To test for insertion sites, we use microarray intensity values extracted from small windows of contiguous open reading frames. Rank-transformation of spot intensities and first-order differences in local windows provide enough information to reliably determine the presence of an insertion sequence. The non-parametric approach outperforms all other methods tested in this study.

Gene arrays measure the concentration of thousands of oligonucleotide species in parallel (Chee 1996, Lander 1999, Schena 1995). They have been used principally in the study of mRNA expression levels (Chu 1998, Duggan 1999, Spellman 1998), but can also be used to assess the presence of DNA or RNA for other reasons. For example, arrays can rapidly assess patterns in genomic DNA (as compared with a reference set of genes) in the search for genomic patterns, such as deletions, point mutations, duplications, and single nucleotide polymorphisms (Behr 1999, Gingeras 1998, Halushka 1999, Pollack 1999). Searching for genomic patterns in the context of the inaccuracies of gene array data can be difficult and requires robust algorithmic techniques.

The quality and reproducibility of array experiments is still improving. Since the resolution of the data tends to be coarse, methods built to extract information from arrays must be engineered to handle extreme amounts of noise, missing observations, and outlier points. Hybridization is a function of many experimental variables and intensity measurements have multiple sources of error. The targets on a spotted array are of different sizes, are spotted in different concentrations and may even be entirely missing. Variation in G-C content and secondary structural features can give rise to differences in hybridization affinities of target sequences for their respective probes. All of these (and perhaps other factors) contribute to the uncertainty of the intensity signal collected from microarray experiments.

One particular problem for microarray data is the difficulty in comparing the absolute intensities that are observed within and between experiments. Any assumptions about straightforward distributions of values from which these intensities are drawn are difficult to defend. We therefore introduce a general distribution-free technique for identifying genomic

sxr@smi.stanford.edu,stuart@smi.stanford.edu, rba@smi.stanford.edu.

patterns with appropriate microarray experiments. We demonstrate it with a specific application to finding the positions of insertion sequences in genomes.

Mycobacterium tuberculosis (*M. tb.*) is an infectious pathogen reported in 8 million clinical cases and claiming 2–3 million lives per annum (Kumar 1997, Sherris 1990). It is an illness that displays considerable heterogeneity dependent on both host and pathogen; some strains are extremely infectious, while others are relatively benign factors (Bloom 1998, Martinez 2000, Sherris 1990, Sreevatsan 1998, Valway 1997). Methods for typing particular strains have already proven beneficial in pathogen control (Small 1993). Associating pathogenic phenotypes to particular genomic features, such as insertions, may link pathogenesis with the molecular constituents and may allow us to predict the clinical behavior of a particular strain before it results in a malignant clinical presentation.

The IS6110 insertion element is a well known transposable element found to have variable numbers and locations in genomes across different strains (de Boer 1999). It occurs anywhere from 0 to 25 times depending on the particular *M. tb.* strain. The pattern of insertions has been used in the past to type clinical isolates of the mycobacterium in epidemiological studies (van Embden 1993, Bradford 1997). The current typing method, Restriction Fragment Length Polymorphism fingerprinting is limited in its ability to distinguish between a significant number of isolates. A robust analytical method for rapidly identifying the number and exact genomic location of insertion elements combined with a well-established array protocol could be the key to rapidly typing hundreds of clinical strains and developing an understanding of the epidemiology, biology, and evolution of this pathogen.

Discriminant analysis is a standard statistical approach described in most textbooks of multivariate analysis (Anderson 1984, Mardia 1982). This technique determines to which of two well-defined populations a given test example belongs. The technique typically involves estimating a multivariate normal distribution over a set of predefined features to describe each of the two populations from known training cases. Subsequently, the log likelihood ratio is estimated for each test case; values above a certain threshold are classified into one population while those below the threshold are classified into the other.

However, in microarray data analysis, assuming a normally distributed feature set can be problematic – the data contains outliers and complexities that can reduce the effectiveness of such an assumption. One extreme way to deal with this issue is to estimate probability densities from features of the training set directly -- however this may be *ad hoc* or computationally intensive. An alternative approach for coping with outliers is to transform the data into ranks. Thus, the absolute value of a measurement becomes less important than its value relative to other measurements. In this study, all variables taken from microarray data were first converted to their corresponding ranks and the rank values were used for the analysis. This mitigates the influence of the outliers and results in behavior that can be approximated with normal distributions.

The choice of features extracted from the data also heavily affects any method's ability to find a useful discriminant function. Here we present our empirical findings for defining a set of features appropriate for the task of distinguishing IS6110 sites from non-sites.

Method

Summary of Experimental Protocol

The details of the microarray experimental protocol will be published elsewhere, but can be summarized here. Taking advantage of the fact that all insertion sequences are the same, we use two separate primers designed to recognize the 5′ and 3′ end regions of IS6110 conserved

sequences in a genomic primer extension assay to determine the locations of the insertion sequences (Figure 1). Both primers point out from the conserved sequence so that only the flanking genes are copied. Primer extension originating from one annealing site will generate a population of products that are heterogeneous in length. No polymerase is perfectly processive and extension will cease arbitrarily at different lengths (Stryer 1995). During any given extension event, downstream sequences closer to the primer annealing site are more likely to be copied by the polymerase than sequences far from the annealing site. Therefore, in the pool of products generated from multiple rounds of polymerization, sites neighboring the primer annealing site will be over-represented whereas distant sites will be under-represented.

The products of the 5′ primer extension are labeled with one fluor (solid gray in Figure 1) and the 3′ primer products are labeled with a different fluor (dashed gray), so that each color corresponds to only one of the two directions of DNA synthesis. Downstream target sites on the microarray lying closer to an IS6110 site should light up more brightly on the array because there is a relatively large amount of extension product with overlapping sequence complementarity. Sites further from IS6110 sites look dim on the array since few extension products incorporate complementary sequence.

The microarray used for these assays (Behr 1999) contains a target site corresponding to each Open Reading Frame (ORF) listed in (Cole 1998). We use the Sanger numbering over the ORFs which order them from 1 to 3924 starting from the origin of replication on the single circular chromosome, keeping in mind that ORF 1 and ORF 3924 are neighbors (Cole 1998).

Computational Approach

Since each insertion site may be oriented 5′ to 3′ or 3′ to 5′, there are 7848 (=2 orientations × 3924 intergenic regions) candidate positions that we must screen for the presence of IS6110 sites.

The inputs to our procedure are ordered vectors of intensity values **s3** and **s5**. The intensity value for the 3′ fluor at the i^{th} ORF is recorded in $s3(i)$ while the intensity value for the 5′ fluor at the same site is in $s5(i)$. The program outputs two ordered vectors of predictions, $\pi_{\mathbf{f}}$ and $\pi_{\mathbf{r}}$; the former contains predicted insertion sites in the forward direction, while the latter in the reverse direction. If an insertion sequence is predicted to occur between ORF i and ORF i+1 oriented with ORF i closest to the 5′ end of the insertion site then $\pi_f(i) = 1$. If the insertion site is predicted to be oriented in the reverse direction $\pi_r(i+1) = 1$, where ORF i+1 is now closest to the 5′ end. All other situations assign 0 to the prediction vector.

Before conducting any analysis, we normalize intensities so that the resulting data set has a mean equal to 0 and a variance equal to 1. Intensity values collected from two different microarrays are not directly comparable since the number of DNA fragments generated in each experiment may differ. Normalizing intensity values allows combination of multiple data sets in a straightforward manner. Also, typically we average the results of two microarray experiments together to produce a more reliable profile.

The highest levels of intensity are at the two positions immediately downstream of the insertion site (ORF −1 for 5′ fluor and ORF +1 for 3′ fluor). In order to reduce our search space we eliminate candidate sites if either of these intensities are less than a threshold; which we set (after testing) to 0, the mean intensity. This +1/−1 cutoff procedure eliminates 92% of candidate sites without loss of sensitivity.

We find the features most effective in distinguishing insertion sites from non-insertion sites to be local intensities of candidate sites from both the 3′ and 5′ series and first order neighbor-to-

neighbor difference information between these local intensities. The difference information helps to preserve the form associated with insertion sites, by approximating first derivative information.

We obtain optimal results when we include intensities 4 ORFs downstream and 2 ORFs upstream of the candidate site. For example, if we are attempting to predict the presence of the insertion site between position *i* and *i+1* oriented with its 5′ side closest to ORF *i*, we would include intensities from the 5′ series, **s5**, at ORFs {*i*−3, *i*−2, *i*−1, *i*, *i*+1, *i*+2} and from the 3′ series, **s3**, {*i* −1, *i*, *i*+1, *i*+2, *i* +3, *i*+4} (Figure 2). We use first order differences between intensities from adjacent ORFs to augment the feature set. This corresponds to a feature set of 12 intensities (6 ORFs/series × 2 series) and 10 differences (5 differences/series × 2 series); each candidate site is associated with a feature vector of size *f*=22.

Figure 3 plots the profiles that we attempt to distinguish between. This is a plot over the insertion site (position 0) and the surrounding intensities in the 5′ series and the 3′ series.

Discrimination Procedure

In order to discriminate there must be two sets of feature training examples: *n* insertion sites and *m* non-insertion sites that have surpassed the +1/−1 cutoff elimination described in the last paragraph. We construct insertion site profile examples by looking at known insertion sites across multiple experiments to generate many examples. Typical values are $n = 540$, $m = 2000$. The *n* examples where IS6110 is “present” are configured in a $n \times f$ matrix **P** and the *m* examples where it is “absent” are configured in a $m \times f$ matrix **A**.

In parametric linear discrimination we use training examples to estimate normal population distributions for both cases assuming an identical feature covariance matrix **S**, but differing in feature means.

$$S = 1/2\mathrm{cov}(P) + 1/2\mathrm{cov}(A)$$

Different covariance matrices can also be treated without much difficulty (Anderson 1984). We define $\bar{p}$ as the $f \times 1$ column feature vector mean of **P**, and $\bar{a}$ as the $f \times 1$ column feature vector mean of **A**. Given a particular feature column vector *x* from a test case associated with a candidate site, we calculate a log-likelihood classification score:

$$f(x) = \log \frac{p(x|x \in P)}{p(x|x \in A)}$$

Test cases that receive high scores are likely to be insertions; cases with low scores are likely to be non-insertions. In a parametric discrimination approach we assume a normal distribution to estimate the log ratio:

$$f(x) = \log \left(\frac{e^{-1/2^*(x-\bar{p})S^{-1}(x-\bar{p})}}{e^{-1/2^*(x-\bar{a})S^{-1}(x-\bar{a})}} \right)$$

This reduces to the linear discriminant function:

$$f(x)=(\overline{p}-\overline{a})S^{-1}(x-\frac{1}{2}(\overline{p}+\overline{a}))$$

We score each test case with the above function; all test cases with a value above a certain threshold are classified as insertion elements.

Our actual approach is identical to the above, except for substitution of the feature values with rank values before applying the discrimination strategy. In the standard discrimination procedure the mean and variance of a given feature, and its covariance with other features are calculated from actual values. Instead, for each feature we assign ranks ranging form 1 to $n + m$ for all training examples in **P** and **A**. We then calculate mean and variance with respect to the rank values. Whenever a test example is presented, each feature value is converted to a rank relative to the training examples to construct a rank feature vector. We use this vector instead of an intensity vector to score the test example.

Procedure Evaluation

As our gold standard we use the H37Rv strain whose single chromosome genome was sequenced by Sanger in 1998 (Cole). At the time of this communication a number of microarray experiments were available to us: 10 sets of 5′ primer intensity series and 4 sets of 3′ primer intensity series. The experiments were conducted under variable conditions and protocols. The H37Rv strain contains 16 insertion sites with known positions oriented in both directions along the genome. Of these, two are contiguous in the same intergenic region and are oriented in the same direction; we treat these as a single insertion. Another insertion site is located between two ORFs with corresponding targets on the microarray that were defective; we eliminate this site from our analysis.

For the matrix **P** of positive training examples there are 560 instances we can choose from (14 sites × 10 5′series × 4 3′ series). For negative training examples **A** we choose 2000 sites oriented in either direction that are not insertion sites, but pass +1/−1 cutoff elimination. We do not average two sets of data to construct our training examples; averaging training (not testing) examples did not improve performance.

Specificity of our method was studied by analyzing 200,000 randomly picked intergenic positions with replacement that were not insertion sites. For each of the 200,000 positions picked, two of the 5′ series and two of the 3′ were drawn with replacement, and averaged. We used these averaged profiles to construct a feature set for each of our false test cases. All 560 examples were included in **P**; the set **A** was constructed as described above. All of these sites were scored with our procedure to establish specificities at particular thresholds (see Figure 5).

Sensitivity of our method was studied with cross-validation. We iterated through each insertion site, attempting to score profiles from it with positive examples created from the other 13 insertion sites. For each site we constructed 500 test cases by randomly drawing two 5′ series and two 3′ series with replacement and building a feature set vector. The positive examples, **P**, were constructed with the remaining 520 examples from the other 13 insertion sites; **A** was constructed as above. The pooled scores were used to calculate sensitivity levels at different threshold values.

We implemented the above approach and the alternative variants described below in Matlab.

A sensitivity-specificity plot of our algorithm compared with other variants are presented in Figure 5.

Results

Profiles of the standardized intensity values for windows focused at IS6110 sites are shown in Figure 3; this figure depicts the patterns that we discriminate between. The profiles for both the 5′ and 3′ primers are depicted in the figure. Primer extension from both ends of the insertion sequence yields similar amounts of complementary information useful for detecting insertion sequences. The error bars indicate the overlap in position intensities between the two profiles.

The means and variances between the {−1, −2} 5′ positions and the {+1, +2} 3′ positions are comparable. However, the 5′ positions {−3, −4, −5} all have larger means than their 3′ counterparts at {+3, +4, +5} suggesting a more detectable 5′ signal further away from the insertion site. The most informative positions lie near the insertion site, occurring approximately within 4 to 5 downstream ORFs of an IS6110 site. A number of variations on the discrimination procedure were performed to test the size and prediction position of the window. Window sizes approximately 6 ORFs in width out-performed windows including greater or fewer numbers of ORFs (data not shown). Also, windows asymmetrically positioned around the predicted position were superior to centered windows; incorporating more ORFs downstream rather than upstream of the predicted sites gave better results (data not shown). Large variances are associated with large mean intensities, suggesting there may be a relationship between the measurement error and the spot intensity on these microarrays.

A histogram of 5′ series intensity and rank intensity for the −1 positions in both sites and non-sites obtained from microarray experiments is shown in Figure 4. The distribution of site intensities is much flatter and wider than the non-site intensity distribution (part A of the figure). Approximating either distribution with a Gaussian would not be appropriate given the shape of the actual distributions. The non-site intensity distribution appears more exponential and the site distribution is slightly bimodal. On the other hand, the rank distributions appear to have a more normal shape (part B of figure). Also, both populations have similar variance.

Figure 5 showcases performance in sensitivity-specificity plots (or ROC curves); it contains several of the algorithmic variants that we tested. All algorithmic variants are identical except in the indicated aspect. Part A compares the results of using a space of feature ranks (our procedure) versus a space of feature values (parametric discrimination). We display both linear discrimination (pooled covariance matrix) and the quadratic discrimination (two distinct covariance matrices) when feature values are used. Note the differences in prediction performance between the discriminators using values versus ranks; for example, at a threshold yielding a sensitivity level of 65%, the rank discriminator produces 1 false positive out of the 5000 tested positions whereas the parametric discriminator produces approximately 5 false positives.

Including first-order differences in the feature vectors improves the performance of the method, especially at high specificity ranges (Figure 5B). For example, at the 99.98% specificity level, first-order differences increase the sensitivity from 54% to 65%. The +1/−1 cutoff elimination gives performance increases of the same order as first-order differences with an effect over a slightly broader specificity range.

Figure 5C illustrates the performance gains associated with combining multiple experiments into test cases. The sensitivity-specificity curve for the predictor given a single experiment lies completely beneath the curve for the same predictor given the averaged results of two experiments. The curve for the predictor given doublet averages (our procedure) in turn lies beneath the curve for the same predictor given quadruplet averages. The most dramatic improvement occurs when a second experiment is taken into consideration -- at the 85% sensitivity level, the predictor using two averaged experiments produces 8 false positives while producing around 250 false positives given only one experiment.

Discussion

Successful design of an experimental and analytic method for typing *M. tb.* will greatly enable studies of how different strains have different clinical phenotypes. When analyzing these results it is important to keep in mind that, during genotyping of one strain, thousands of sites are scanned for a comparatively few number of positive sites. Maintaining an extremely low false-positive rate is therefore key to any algorithm's success.

Parametric discriminant analysis on feature values performs poorly (Figure 5A). At specificity levels realistic for typing strains (>99.95%), parametric prediction using feature values detects <50% of insertion sites. It is a well known phenomenon that linear discriminant analysis performs poorly when the underlying distributions are far from normal or have wildly different covariance structures (both of which are indicated by the distribution plot in Figure 4A). Correcting for covariance differences does not improve the discrimination (Figure 5A); this may be a consequence of introducing too many parameters to fit a few (at most 560) positive training examples. On the other hand, a multivariate normal distribution approximates the components of the rank feature vector well when the number of training examples is large. The normal theory applies to the rank data better, so we can bring the power of the linear discriminator to bear on the problem.

The single most important factor affecting the method's accuracy in detecting insertion sequences is the number of microarray experiments used to perform the typing. Averaging intensities across multiple primer extension experiments greatly improves the performance of the method compared to using only a single experiment. In practice, however, it may be desirable to minimize the number of microarray experiments required to site type per clinical isolate, which motivates the improvement of the algorithm's performance on non-averaged data sets.

The success of our approach is dependant on a filtering step that eliminates from consideration sites where the +1 and −1 ORF positions are both below the mean intensity level. All such windows are classified as non-sites by our method before using the discriminator. This step not only eliminates from consideration many non-sites that would be mis-called by the algorithm, but implicitly eliminates many cases where missing values occur at these positions due to a defect in the microarray.

While the experimental protocols for detecting IS6110 locations are improving, the algorithm presented demonstrates that the current technique is already viable for site-typing. At 99.99% specificity it detects over half of the insertion sequences present in the genome. One run of the algorithm at this level of specificity will predict 8 of the approximately 8000 examples as putative insertions out of which 1 will be erroneously called. The method does not exhibit systematic error in that false positives and false negatives do not seem to be correlated across experiments.

The experimental data on which we conducted our analysis on was obtained under heterogeneous conditions. These conditions are now being optimized to include a better choice of hybridization procedure and replacement of defective targets on the microarray with new ones. The results presented here should only improve as the quality of the input data increases. However, it is unlikely that the data will ever be so homogenous and noise-free to render our methods unnecessary.

The discovery that the performance of the algorithm improves when ORFs upstream of the IS6110 primer are included in the feature set indicates there is some signal in the opposite direction of extension (Figure 2). We will not conjecture about the biological phenomenon

producing this curiosity but merely note that our algorithm takes advantage of some decrease or lack of signal in this region.

The success of the ranked version versus the non-ranked version of the approach underscores the importance of treating the distribution of microarray measurements in an unbiased manner. Converting microarray measurements into ranked data has two main advantages with respect to the robustness of the method --(1) it moderates the effects of outliers common in typical microarray datasets, and (2) the results are invariant to any monotonic transformation of the data (such as a logarithm). Global shifts in intensity values from one experiment to the next will not typically alter the accuracy of the method since ranks are indifferent to scaling of the data. This is a desirable property of the method given the high degree of variability across microarray experiments. The method requires the relative order of intensities remain fairly reproducible instead of the stronger requirement that the intensities themselves remain reproducible.

We believe that a similar algorithmic approach may be adapted to identify other genomic patterns such as deletions, single nucleotide polymorphism, point mutations, and gene duplications on microarray data as well as expression patterns. The same qualities that permit our methodology to perform well under the circumstances described in this study should apply in these other circumstances.

We recognize that there are other pattern recognition algorithms, such as neural networks or genetic algorithms, which may be useful for these types of problems. We used discrimination analysis as the simplest first step. Other types of non-parametric methods may prove useful and perhaps necessary for identifying patterns in and analyzing hybridization experiments.

Acknowledgments

The authors wish to thank Michael Cantor, Mary Lu, and Olga Troyanskaya for assistance in manuscript preparation. S.R. is supported by NIH training grant GM-07365; J.M.S. is supported by NIH training grant LM-07033. This work was also supported by NIHLM06244, NSF DBI-9600637 and a grant from the Burroughs-Wellcome Foundation.

References

Anderson, TW. An Introduction to Multivariate Statistical Analysis. New York, N.Y: John Wiley & Sons; 1984.

Behr MA, Wilson MA, Gill WP, Salamon H, Schoolnik GK, Rane S, Small PM. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science 1999;284:1520–1523. [PubMed: 10348738]

Bloom BR, Small PM. The evolving relation between humans and *Mycobacterium tuberculosis*. N Eng J Med 1998;338:677–678.

Bradford WZ, Koehler J, El-Hajj H, Hopewell PC, Reingold AL, Agasino CB, Cave DM, Rane S, Yang Z, Crane CM, Small PM. Dissemination of *Mycobacterium tuberculosis* across the San Francisco Bay Area. J Inf Dis 1997;177:1104–1107. [PubMed: 9534993]

de Boer AS, Borgdorff MW, de Haas PEW, Nagelkereke NJD, van Embden JDA, van Soolingen DV. Analysis of rate of change of IS6110 RFLP patterns based on serial patient isolates. The Journal of Infectious Diseases 1999;180:1238–1244. [PubMed: 10479153]

Chee M, Yang R, Hubbel E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SPA. Accessing genetic information with high-density DNA arrays. Science 1996;274:610–614. [PubMed: 8849452]

Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I. The transcriptional program of sporulation in budding yeast. Science 1998;282:699–705. [PubMed: 9784122]

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Conner R, Davies R, Devlin K,

Feltwell T, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 1998;393:537–544. [PubMed: 9634230]

Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. Nature Genetics 1999;21:10–14. [PubMed: 9915494]

Gingeras TR, Ghandour G, Wang E, Berno A, Small PM, Drobniewski F, Alland D, Desmond E, Drenkow J. Simultaneous genotyping and species identification using hybridization pattern recognition analysis of generic *Mycobacterium* DNA Arrays. Genome Research 1998;8:435–448. [PubMed: 9582189]

Halushka MA, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nature Genetics 1999;22:239–247. [PubMed: 10391210]

Kumar, V.; Cotran, RS.; Robbins, SL. Basic Pathology. London: Saunders; 1997.

Lander ES. Array of hope. Nature Genetics 1999;21:3–4. [PubMed: 9915492]

Mardia, KV.; Kent, JT.; Bibby, JM. Multivariate Analysis. New York, N.Y: Academic Press; 1982.

Martinez AN, Rhee JT, Small PM, Behr MA. Sex differences in the epidemiology of tuberculosis in San Francisco. Int J Tuberc Lung Dis 2000;4:26–31. [PubMed: 10654640]

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffery SS, Botstein D, Brown PO. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nature Genetics 1999;23:41–46. [PubMed: 10471496]

Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270:467–470. [PubMed: 7569999]

Sherris, JC., editor. Medical Microbiology. New York, N.Y: Elsevier; 1990.

Small PM, McClenny NB, Singh SP, Schoolnik GK, Tompkins LS, Mickelsen PA. Molecular strain typing of *Mycobacterium tuberculosis* to confirm cross-contamination in the AFB laboratory and modification of procedures to minimize occurrence of false positive cultures. J Clin Micro 1993;31:1677–1682.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Fucher B. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. Molecular Biology of the Cell 1998;9:3273–3297. [PubMed: 9843569]

Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, Musser JM. Restricted structural gene polymorphism in *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci USA 1997;94:9869–9874. [PubMed: 9275218]

Stryer, L. Biochemistry. New York, N.Y: W.H. Freeman; 1995.

Valway SE, Sanchez MP, Schinnick TF, Orme I, Agerton T, Hoy D, Jones JS, Westmoreland H, Onorato IM. An outbreak involving extensive transmission of a virulent strain of *Mycobacterium tuberculosis*. N Engl J Med 1998;338:633–639. [PubMed: 9486991]

Van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenarch KD, Gicquel B, Hermans P, Martin C, McAdam R, Schinnick TM, et al. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol 1993;31:406–409. [PubMed: 8381814]

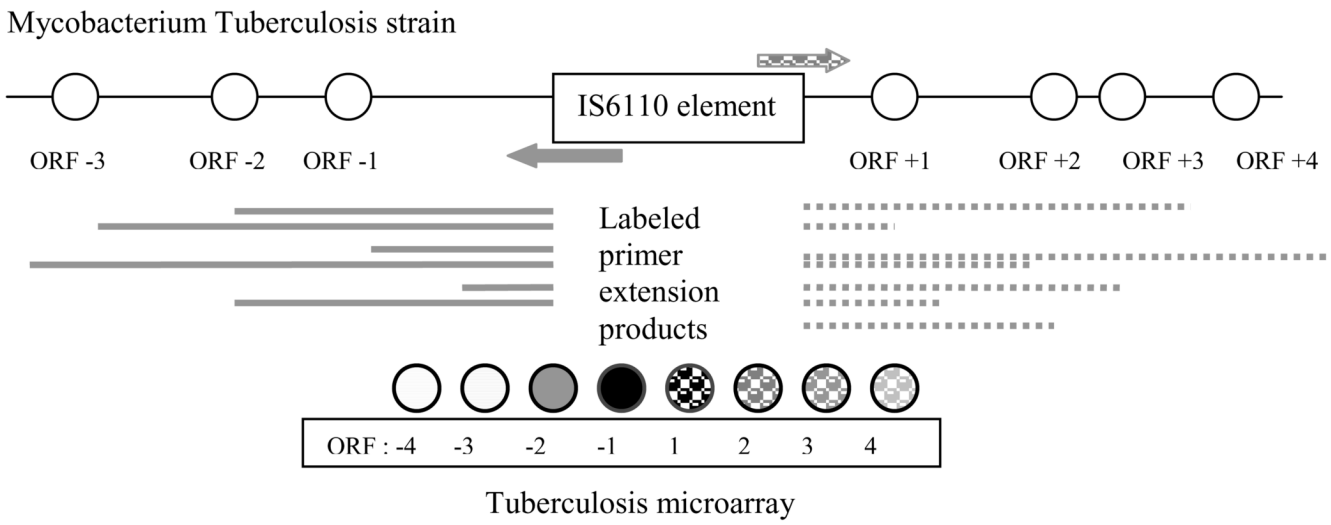

Figure 1.

Schematic description of the biological procedure. We attempt to discern the location of an IS6110 element with straightforward application of microarray technology. Primers specific for the 3′ (right arrow) and 5′ (left arrow) end of the insertion sequence are applied to the genome. Primer extension generates labeled DNA fragments. Upon application of these fragments to the microarray spots representing ORFs closest to the IS6110 element appear intense (black), while spots further from the IS6110 element appear dim (light gray). In the experiment the 5′ and 3′ fragments are labeled with two different dyes so that spots −4 to −1 will be labeled with one dye (solid) and spots 1 to 4 will be labeled with a different dye (dashed).

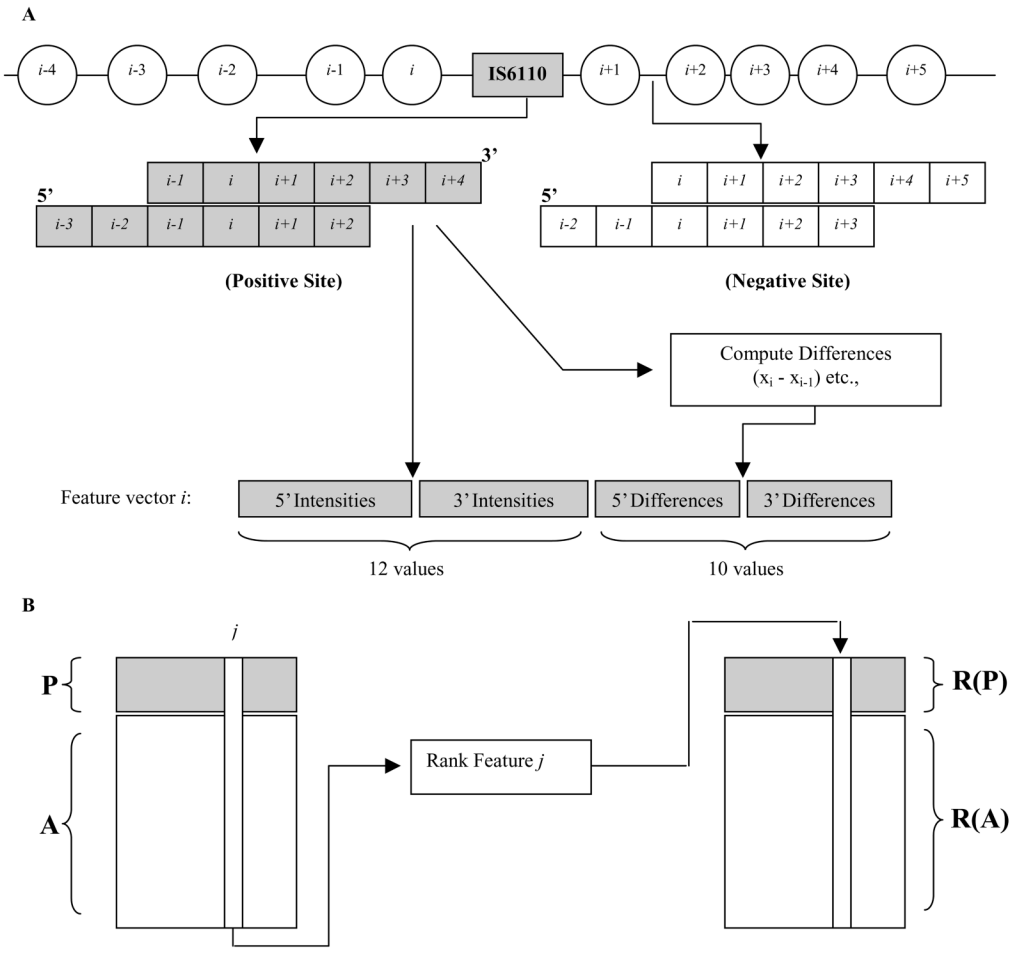

Figure 2.

Illustration of the feature extraction method. **A.** Features from the data are constructed by combining standardized intensities and first-order differences from two extension reactions into one 22-length feature vector. Two ORFs upstream and 4 ORFs downstream of the predicted position are included from both the 5′ and the 3′ reactions. Feature vector *i* is labeled as a positive site since an IS6110 sequence occurs between ORF *i* and ORF *i*+1. Feature vector *i*+1 on the other hand is labeled negative since no insertion sequence occurs between ORF *i*+1 and ORF *i*+2. **B.** The block matrix of training examples, [**P**/**A**] is constructed from the positive feature vectors, **P**, and the negative feature vectors, **A**. A column in [**P**/**A**] corresponds to one feature's values across the entire training set. A column of ranks is computed from the combined vector of negative and positive sites. The column vector for feature *j* is then replaced with the computed rank vector and the resulting **R(P)** and **R(A)** matrices are used in the discrimination procedure.

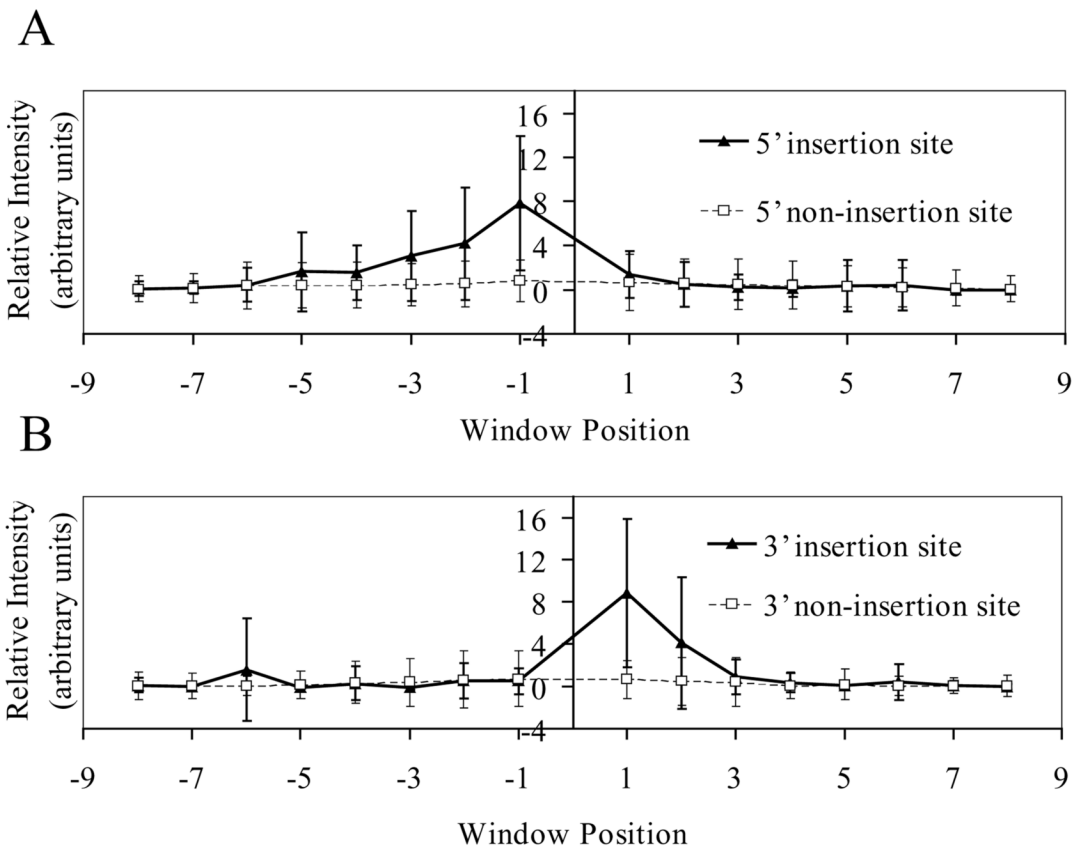

Figure 3.

Plot of the profiles that we are trying to distinguish: true insertion sites (▲) vs. non-insertion sites that are not eliminated by the +1/−1 cutoff (□). (Intensity values have been normalized to mean=0, variance=1; error bars are +/− 1 SD.) The solid lines represent experimental microarray intensity values at ORFs surrounding the insertion site at position 0. Dotted lines represent randomly chosen examples from the rest of the genome that do not have an insertion site at position 0 but pass the cutoff. The error bars indicate substantial overlap at most positions. **A.** Intensity values from microarray experiment conducted with primer specific for 5′ side of insertion sequence. **B.** Intensity values for 3′ primer microarray experiment.

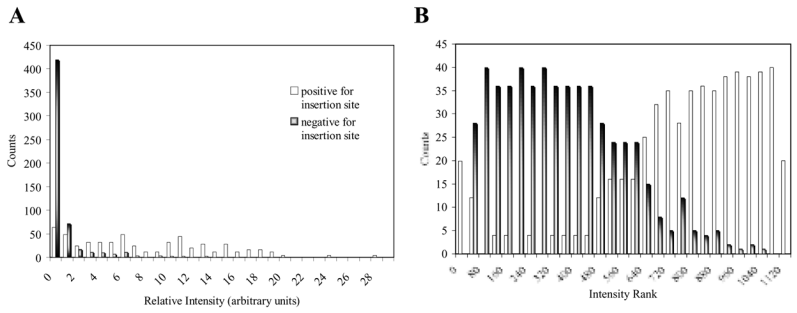

Figure 4.

A. As an example of the non-normal nature of the features examined in our discrimination procedure, we plot histograms of the intensity of the ORF closest to candidate insertion site on the 5 prime side from the 5′ primer experiment. These values correspond to the summarized values at ORF −1 in Figure 3A. We expect, for actual insertion sites that these values should be relatively high. (The above plots were generated from 560 positive examples and 560 negative examples where the +1/−1 cutoff was exceeded; intensity values have been normalized to mean=0, variance =1 across each experiment.) Notice both histograms appear to have different functional forms; both of which would be inaccurately estimated by a normal distribution. **B.** After replacing raw intensity data with ranks (highest intensity value has rank1120, lowest intensity value has rank 1) we recreate the histogram. Notice that these distributions lack the extreme nature of those in **A**; they are better approximated by a normal distribution.

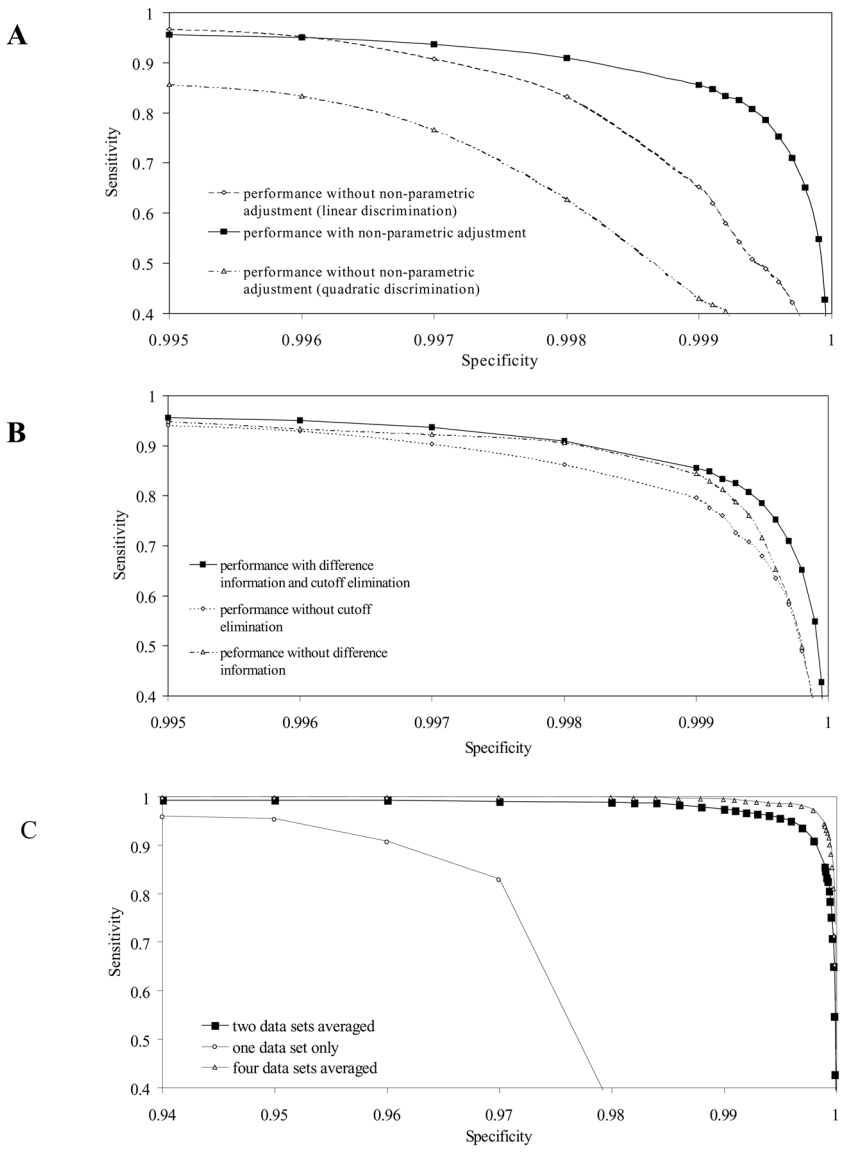

Figure 5.

Sensitivity-specificity plots illustrating performance of our approach and the critical nature of each of the features. The curves with the solid boxes (■) are identical in each plot. A. Our approach (upper trace) is compared to identical approaches in all aspects except for the non-parametric adjustment. One trace demonstrates performance when parametric estimation is attempted with one pooled covariance matrix for both distributions of insertion elements and non-insertion elements (linear discrimination); the other demonstrates estimation with two separate covariance matrices (quadratic discrimination). B. Success of our approach (upper trace) relies on utilizing difference information as well as eliminating many cases with adjacent intensities below a threshold at high specificity values. Success of identical procedure conducted without these features is depicted individually in this figure. C. Attempting discrimination on a profile created from multiple averaged experiments improves performance. Most of the results generated assume averaging two experiments.