BIOINFORMATICS

Vol. 22 no. 14 2006, pages e49–e57
doi:10.1093/bioinformatics/btl242

Integrating structured biological data by Kernel Maximum Mean Discrepancy

Karsten M. Borgwardt[1,*], Arthur Gretton[2], Malte J. Rasch[3], Hans-Peter Kriegel[1], Bernhard Schölkopf[2] and Alex J. Smola[4]

[1]Institute for Computer Science, Ludwig-Maximilians-University Munich, Germany, [2]Max Planck Institute for Biological Cybernetics, Tübingen, Germany, [3]Graz University of Technology, Austria and [4]National ICT Australia, Canberra, Australia

ABSTRACT

Motivation: Many problems in data integration in bioinformatics can be posed as one common question: Are two sets of observations generated by the same distribution? We propose a kernel-based statistical test for this problem, based on the fact that two distributions are different if and only if there exists at least one function having different expectation on the two distributions. Consequently we use the maximum discrepancy between function means as the basis of a test statistic.

The Maximum Mean Discrepancy (MMD) can take advantage of the kernel trick, which allows us to apply it not only to vectors, but strings, sequences, graphs, and other common structured data types arising in molecular biology.

Results: We study the practical feasibility of an MMD-based test on three central data integration tasks: Testing cross-platform comparability of microarray data, cancer diagnosis, and data-content based schema matching for two different protein function classification schemas. In all of these experiments, including high-dimensional ones, MMD is very accurate in finding samples that were generated from the same distribution, and outperforms its best competitors.

Conclusions: We have defined a novel statistical test of whether two samples are from the same distribution, compatible with both multivariate and structured data, that is fast, easy to implement, and works well, as confirmed by our experiments.

Availability: http://www.dbs.ifi.lmu.de/~borgward/MMD

Contact: kb@dbs.ifi.lmu.de

1 INTRODUCTION

1.1 Data integration in bioinformatics

The ultimate need for bioinformatics is founded on the wealth of data generated by modern molecular biology. The purpose of bioinformatics is to structure and analyze this data. A central preprocessing step is the integration of datasets that were generated by different laboratories and techniques. If we know how to combine data produced in different labs, we can exploit the results jointly, not only individually. In some cases, the larger datasets thus constructed may support biologically relevant conclusions which were not possible using the original smaller datasets, a hypothetical example being the problem of reliable gene selection from high-dimensional small microarray datasets.

*To whom correspondence should be addressed.

1.2. Distribution testing in data integration

The questions arising in data integration essentially boil down to the following problem of distribution testing: Were two samples X and Y generated by the same distribution? In data integration terms, are these two samples part of the same larger dataset, or should these data be treated as originating from two different sources?

This is a fundamental question when two laboratories are studying the same biological subject. If they use identical techniques on identical subjects but obtain results that are not generated by the same distribution, then this might indicate that there is a difference in the way they generate data, and that their results should not be integrated directly. If the data were integrated without recalibration, differences or patterns within the joint data might be caused by experimental discrepancies between laboratories, rather than by biological processes.

As microarray data are produced by a multitude of different platforms, techniques and laboratories, they are the most prominent data source in bioinformatics for which distribution testing is indispensable. Recently, Marshall (2004) gave an extremely negative picture of cross-platform comparability—and hence the reliability and reproducibility—of microarray results, due to the various platforms and data analysis methods employed (Shi *et al*., 2005). It is therefore crucial for bioinformatics to develop computational methods that allow us to determine whether results achieved across platforms are comparable. In this article, we present a novel statistical test to tackle this problem.

What distinguishes bioinformatics is that it has produced a wealth of complex data, from protein sequences to protein interaction networks, i.e. from strings to graphs. Consequently any practically relevant distribution test needs to be *easily* applicable in all these cases. To the best of our knowledge, the statistical test proposed in our paper is the first method that can handle this wide range of different domains.

To summarize our goals, we will present a novel statistical test for differences in distribution, based on the Maximum Mean Discrepancy (MMD). We will show that it can take advantage of the kernel trick. Hence it is applicable to all data types, from high-dimensional vectors to strings and graphs, arising in bioinformatics. In experiments, we will apply this test to microarray cross-platform comparability testing and cancer diagnosis. Furthermore, we will show how to perform schema matching on complex data by considering a data integration problem on two molecular graph datasets.

Outline of this article In Section 2, we present MMD and its properties. In Section 3, we test the applicability of MMD in cross-platform microarray comparability analysis and cancer diagnosis, and evaluate it on a schema matching problem. We discuss our findings in Section 4.

2 MMD AND THE TWO-SAMPLE PROBLEM

In statistics, the central question of data integration described above is often referred to as the *two-sample* or *homogeneity problem*. The principle underlying the maximum mean discrepancy is that we want to find a function that assumes different expectations on two different distributions. The hope then is that if we evaluate this function on empirical samples from the distributions, it will tell us whether the distributions they have been drawn from are likely to differ. This leads to the following statistic, which is closely related to a proposal by [Fortet and Mourier (1953)]. Here and below, $\mathcal{X}$ denotes our input domain and is assumed to be a nonempty compact set.

DEFINITION 2.1. *Let $\mathcal{F}$ be a class of functions $f:\mathcal{X}\to\mathbb{R}$. Let p and q be Borel probability distributions, and let $X = (x_1, \ldots, x_m)$ and $Y = (y_1, \ldots, y_n)$ be samples composed of independent and identically distributed observations drawn from p and q, respectively. We define the maximum mean discrepancy (MMD) and its empirical estimate as*

$$\mathrm{MMD}[\mathcal{F}, p, q] := \sup_{f\in\mathrm{F}} \left(\mathbf{E}_p[f(x)] - \mathbf{E}_q[f(y)]\right)$$

$$\mathrm{MMD}[\mathcal{F}, X, Y] := \sup_{f\in\mathrm{F}} \left(\frac{1}{m}\sum_{i=1}^{m} f(x_i) - \frac{1}{n}\sum_{i=1}^{n} f(y_i)\right)$$

Intuitively it is clear that if $\mathcal{F}$ is 'rich enough', MMD $[\mathcal{F}, p, q]$ will vanish if and only if $p = q$. Too rich an $\mathcal{F}$, however, will result in a statistic that differs significantly from zero for most finite samples X, Y. For instance, if $\mathcal{F}$ is the class of *all* real valued functions on $\mathcal{X}$, and if X and Y are disjoint, then it is trivial to construct arbitrarily large values of MMD$[\mathcal{F}, X, Y]$, for instance by ensuring that $f|_X$ is large and $f|_Y = 0$. This phenomenon of *overfitting* can be avoided by placing restrictions on the function class. That said, these restrictions ought not to prevent the MMD from detecting differences between p and q when these are legitimately to be found. As we shall see, one way to accomplish this tradeoff is by choosing $\mathcal{F}$ to be the unit ball in a universal reproducing kernel Hilbert space, RKHS for short.

We will propose a test of $p = q$, based on an unbiased variant of MMD $[\mathcal{F}, X, Y]$[1] which relies on the asymptotic Gaussianity of this test statistic and on the guaranteed rapid convergence to this asymptotic regime. Thus, the performance guarantees provided by the test apply in the case of a large sample size. The test has a computational cost of $O((m + n)^2)$, although randomization techniques could be employed to reduce the cost to essentially linear time-complexity (at the expense of a somewhat reduced sensitivity).

2.1 MMD for kernel function classes

We now introduce a class of functions for which MMD may easily be computed, while retaining the ability to detect all discrepancies between p and q without making any simplifying assumptions. To this end, let $\mathcal{H}$ be a complete inner product space (i.e., a Hilbert space) of functions $f:\mathcal{X}\to\mathbb{R}$, where $\mathcal{X}$ is a nonempty compact set. Then $\mathcal{H}$ is termed a reproducing kernel Hilbert space if for all $x\in\mathcal{X}$, the linear point evaluation functional mapping $f\to f(x)$ exists and is continuous. In this case, $f(x)$ can be expressed as an *inner product* via

$$f(x) = \langle f\phi(x)\rangle_{\mathrm{H}} \tag{1}$$

where $\phi:\mathcal{X}\to\mathcal{H}$ is known as the *feature space map* from x to $\mathcal{H}$. Moreover, the inner product between two feature maps is called the *(positive definite) kernel*, $k(x, x') := \langle\phi(x), \phi(x')\rangle_{\mathrm{H}}$. Of particular interest are cases where we have an analytic expression for k that can be computed quickly, despite $\mathcal{H}$ being high- or even infinite-dimensional. An example of an infinite-dimensional $\mathcal{H}$ is that corresponding to the Gaussian kernel $k(x, x') = \exp(-\|x - x'\|^2/(2\sigma^2))$.

We will consider universal reproducing kernel Hilbert spaces in the sense defined by Steinwart (2002). Although we do not go into technical detail here, we are guaranteed that RKHSs based on Gaussian kernels are universal, as are string kernels (Section 2.3). See also (Schölkopf *et al*., 2004) for an extensive list of further kernels.

When $\mathcal{F}$ is the unit ball in a universal RKHS, the following theorem (Smola *et al*., 2006) guarantees that MMD$[\mathcal{F}, p, q]$ will detect any discrepancy between p and q.

THEOREM 2.2. *Let p, q be Borel probability measures on $\mathcal{X}$ a compact subset of a metric space, and let $\mathcal{H}$ be a universal reproducing kernel Hilbert space with unit ball $\mathcal{F}$. Then MMD$[\mathcal{F}, p, q] = 0$ if and only if $p = q$.*

Moreover, denote by $\mu_p := \mathbf{E}_p[\phi(x)]$ the expectation of $\phi(x)$ in feature space (assuming that it exists).[2] *Then one may rewrite MMD as*

$$\mathrm{MMD}[\mathcal{F}, p, q] = \|\mu_p - \mu_q\|_{\mathrm{H}}.$$

The main ideas for the proof can be summarized as follows. It is known from probability theory (Dudley, 2002, Lemma 9.3.2) that under the stated conditions, a sufficient condition for $p = q$ is that for all continuous functions f, we have $\int f\,dp = \int f\,dq$. Such functions f, however, can be arbitrarily well approximated using functions in a universal RKHS (Steinwart, 2002). For the second part of the result, observe that due to (1), we may rewrite the MMD as

$$\begin{aligned}\mathrm{MMD}[\mathcal{F}, p, q] &= \sup_{\|f\|_{\mathcal{H}}\le 1} \mathbf{E}_p[f(x)] - \mathbf{E}_q[f(y)] \\ &= \sup_{\|f\|_{\mathcal{H}}\le 1} \mathbf{E}_p[\langle\phi(x), f\rangle_{\mathcal{H}}] - \mathbf{E}_q[\langle\phi(y), f\rangle_{\mathcal{H}}] \\ &= \sup_{\|f\|_{\mathcal{H}}\le 1} \langle\mu_p - \mu_q, f\rangle_{\mathcal{H}} = \|\mu_p - \mu_q\|_{\mathcal{H}}.\end{aligned}$$

The finite sample computation of MMD is greatly simplified by (2), as shown in the corollary below:

COROLLARY 2.3. *Under the assumptions of theorem* 2.2 *the following is an unbiased estimator of* MMD$^2[\mathcal{F}, p, q]$:

$$\mathrm{MMD}^2[\mathcal{F}, X, Y] = \frac{1}{m(m-1)}\sum_{i\neq j}^{m} k(x_i, x_j) + \frac{1}{n(n-1)}\sum_{i\neq j}^{n} k(y_i, y_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i, y_j).$$

[1] Note that MMD$[\mathcal{F}, X, Y]$ as defined above is biased: even when $p = q$, it will tend to give strictly positive results for finite sample sizes.

[2] A sufficient condition for this is $\|\mu p\|_{\mathrm{H}}^2 < \infty$, which is rearranged as $\mathbf{E}_p[k(x, x')] < \infty$, where x and x' are independent random variables drawn according to p.

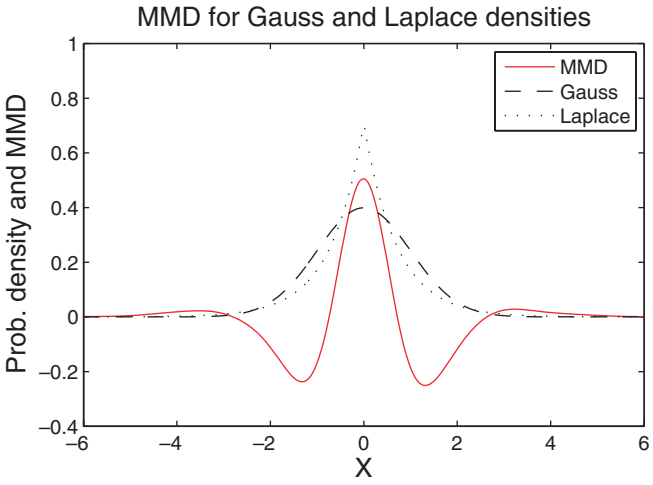

Fig. 1. Illustration of the function maximizing the mean discrepancy in the case where a Gaussian is being compared with a Laplace distribution. Both distributions have zero mean and unit variance. The maximizer of the MMD has been scaled for plotting purposes, and was computed empirically on the basis of 2×10^4 samples, using a Gaussian kernel with $\sigma = 0.5$.

Proof

We compute

$$\begin{aligned}\mathrm{MMD}^2[\mathcal{F}, p, q] &:= \langle \mu_p - \mu_q, \mu_p - \mu_q \rangle_{\mathcal{H}} \\ &= \langle \mu_p, \mu_p \rangle_{\mathcal{H}} + \langle \mu_q, \mu_q \rangle_{\mathcal{H}} - 2\langle \mu_p, \mu_q \rangle_{\mathcal{H}} \\ &= \mathbf{E}_p \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} + \mathbf{E}_q \langle \phi(y), \phi(y') \rangle_{\mathcal{H}} \\ &\quad - 2\mathbf{E}_{p,q} \langle \phi(x), \phi(y) \rangle_{\mathcal{H}},\end{aligned}$$

where x' is a random variable independent of x with distribution p, and y' is a random variable independent of y with distribution q. The proof is completed by applying $\langle \phi(x), \phi(x') \rangle_{\mathrm{H}} = k(x, x')$, and replacing the expectations with their empirical counterparts.

We illustrate the behavior of MMD in Figure 1 using a one-dimensional example: the data X and Y are generated from distributions p and q with equal means and variances, however p is Gaussian and q is Laplacian. For the application of MMD we pick $\mathcal{H}$ to be an RKHS using the Gaussian kernel. We observe that the function f that witnesses the MMD (in other words, the function maximizing the mean discrepancy) is smooth, positive where the Laplace density exceeds the Gaussian density (at the center and tails), and negative where the Gaussian density is larger. Moreover, the magnitude of f is a direct reflection of the amount by which one density exceeds the other, insofar as the smoothness constraint permits it.[3]

Although the expression of $\mathrm{MMD}^2(\mathcal{F}, X, Y)$ in Corollary 2.3 is the minimum variance unbiased estimate (Serfling, 1980), a more tractable unbiased expression can be found in the case where $m = n$, with a slightly higher variance (the distinction is in practice irrelevant, since the terms that differ decay much faster than the variance). It is obtained by dropping the cross-terms $i = j$ from the sum over $k(x_i, y_i)$:

LEMMA 2.4. *Assuming the samples X and Y both have size m, define $z_i = (x_i, y_j)$, and let*

$$h(z_i, z_j) := k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i).$$

An unbiased estimate of $\mathrm{MMD}^2[\mathcal{F}, p, q]$ *is given by*

$$\mathrm{MMD}^2[\mathcal{F}, X, Y] := \frac{1}{m(m-1)} \sum_{i \neq j}^{m} h(z_i, z_j).$$

Note that with some abuse of notation we used the *same symbol* as in Corollary 2.3 for a slightly different estimator. However there should be no ambiguity in that we use only the present version for the remainder of the paper.

An important property of the new statistic is that its kernel $h(z_i, z_j)$ is a positive definite kernel in its own right, since

$$h(z_i, z_j) = \langle \phi(x_i) - \phi(y_i), \phi(x_j) - \phi(y_j) \rangle.$$

Thus $z = (x, y) \to \phi(x) - \phi(y)$ is a valid feature map for h. This gives another interpretation of MMD: it is the expected inner product between vectors obtained by connecting a point from one distribution to a point from the other. For detailed discussions of the problem of defining kernels between distributions and sets, see (Cuturi *et al*., 2005; Hein and Bousquet, 2005).

2.2 MMD tests

We now propose a two-sample test based on the asymptotic distribution of an unbiased estimate of MMD^2, which applies in the case where $\mathcal{F}$ is a unit ball in a RKHS, and $m = n$. This uses the following theorem, due to Hoeffding (1948). See also Serfling (1980, Section 5.5.1). For a proof and further details see Smola *et al*. (2006).

THEOREM 2.5. *Let z_i and $h(z_i, z_j)$ be specified as in Definition* 2.4 *and assume that* $\mathbf{E}_{p,q}[\mathrm{MMD}^4[\mathcal{F}, X, Y]] < \infty$. *Then for $m \to \infty$, the statistic* $\mathrm{MMD}^2(\mathcal{F}, X, Y)$ *converges in distribution to a Gaussian with mean* $\mathrm{MMD}^2[\mathcal{F}, p, q]$ *and variance*

$$\sigma^2_{\mathrm{MMD}} = \frac{2^2}{m}\left(\mathbf{E}_z[(\mathbf{E}_{z'} h(z, z'))^2] - [\mathbf{E}_{z,z'}(h(z, z'))]^2\right).$$

The convergence to the normal occurs rapidly: according to Serfling (1980, Theorem B,p. 193), the CDF of the U-statistic converges uniformly to the asymptotic CDF at rate $1/\sqrt{m}$.

Our goal is to test whether the above normal distribution has zero mean (the null hypothesis), as opposed to a mean that is positive. Since we need not concern ourselves with negative deviations from the mean ($\mathrm{MMD}[\mathcal{F}, p, q] \geq 0$ may never become negative), it suffices to test whether $\mathrm{MMD}^2[\mathcal{F}, X, Y] \leq \varepsilon$ for some threshold ε. Thus, we obtain the two-sample test below as a corollary to Theorem 2.5, following the principles outlined by Casella and Berger (2002, Section 10.3.2).

[3] One may show that the maximizer of $\mathrm{MMD}[\mathcal{F}, p, q]$ is given by $f(x) = \langle \mu_p - \mu_q, \phi(x) \rangle$. The same holds true for the maximizer of the empirical quantity, with the means being replaced by empirical means. See (Smola *et al*., 2006) for further details and a proof.

Algorithm 1 MMD test using asymptotic normality

Input: positive definite kernel k, level of test $\alpha \in (0, 1)$, samples X and Y of size m drawn from p and q respectively

$\text{MMD}^2 \leftarrow 0$ and $\sigma^2 \leftarrow 0$

for $i = 1$ to m **do**

$t \leftarrow 0$

for $j = 1$ to m **do**

if $j \neq i$ **then**

$t \leftarrow t + k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_i, y_i)$

end if

end for

$\text{MMD}^2 \leftarrow \text{MMD}^2 + \frac{1}{m(m-1)} t$ and $\sigma^2 \leftarrow \sigma^2 + t^2$

end for

$\sigma^2 \leftarrow \frac{4}{(m^2(\text{m}-1)^2)}\sigma^2 - \frac{4}{m}(MMD^2)^2$

$\epsilon \leftarrow \sqrt{2\sigma^2}$ erfinv $(1-2\alpha)$

Output: If $\text{MMD}^2 \leq \epsilon$ return $p = q$ accepted. Otherwise return $p = q$ rejected.

COROLLARY 2.6. *A test of the null hypothesis $p = q$ with asymptotic size*[4] *α, and asymptotic Type II error zero, has the acceptance region*

$$\text{MMD}^2[\mathcal{F}, X, Y] \leq \hat{\sigma}_{\text{MMD}} z_\alpha$$

where

$$\hat{\sigma}^2_{\text{MMD}} = \frac{4}{m^2(m-1)^2} \sum_{i=1}^{m} \Big(\sum_{j \neq i}^{m} h(z_i, z_j) \Big)^2 - \frac{4}{m} \text{MMD}^4[\mathcal{F}, X, Y]$$

or any empirical estimate of σ_{MMD} that converges in probability. Here z_α satisfies $\Pr(z > z_\alpha) = \alpha$ when $z \sim \mathcal{N}(0,1)$.

It is also of interest to estimate the p-value of the test. We first describe a sample-based heuristic. We draw randomly without replacement from the aggregated data $Z = \{X, Y\}$ to get two new m-samples X^* and Y^*, and compute the test statistic $\text{MMD}^2_*(\mathcal{F}, X^*, Y^*)$ between these new samples (bear in mind that under the null hypothesis $p = q$, this aggregation is over data drawn from a single distribution). We repeat this procedure t times to obtain a set of test statistics under the null hypothesis (conditioned on the observations). We then add the original statistic $\text{MMD}^2(\mathcal{F}, X, Y)$ to this set, and sort the set in ascending order. Define as r the rank of the original test statistic within this ordering. Then our estimated p-value is $p = (t + 1 - r)/(t + 1)$. Alternatively, we can find an upper bound on p using the distribution-free large deviation result of Hoeffding (1963, p. 25) (see Smola *et al.*, 2006, Section 6), which is exact at finite sample sizes. This bound is only tight when m is large, however, and may be too conservative at small sample sizes.

We give the complete pseudocode for the above MMD-based test in Algorithm 1. We emphasize that the computational cost is $O(m^2)$, and that the method is easily parallelized (the kernel matrix can be broken up into submatrices, and the relevant sums computed independently before being combined). In addition, the kernel matrix needs never be stored in memory, but only a running sum must be kept, which makes the analysis of very large data sets feasible. Randomized methods could also be used to speed up the double-loop required for evaluating Algorithm 6, by only computing parts of the sum. This procedure would reduce the quality of the test, however.

Finally, we note that other approaches are also possible in determining the acceptance region of the test. For instance, Smola *et al.* (2006) describe two tests based on large deviation bounds: the first uses Rademacher averages to obtain a bound that explicitly accounts for the variation in the test statistic, the second uses a distribution-independent upper bound on the test statistic variation due to Hoeffding (1963, p. 25). These approaches have the advantage of giving an exact, distribution-free test of level α that holds for finite samples, and not just in the asymptotic regime. In addition, they provide a finite sample upper bound on the p-value, which is again distribution-free. A disadvantage of these approaches is that they require a larger sample size than the test in Corollary 6 before they can detect a given disparity between the distributions p and q, i.e. they have a higher Type II error. For this reason, we do not use these tests in Section 3.

2.3 Universal kernels for discrete data

While many examples of universal kernels on compact subsets of $\mathbb{R}^d$ are known (Steinwart, 2002), little attention has been given to finite domains. It turns out that the issue is considerably easier in this case: the weaker notion of *strict positive definiteness* (kernels inducing nonsingular Gram matrices $(k(x_i, x_j))_{ij}$ for arbitrary sets of distinct points x_i) ensures that every function on a discrete domain $x = \{x_1, \ldots, x_n\}$ lies in the corresponding RKHS (and hence that the kernel is universal). To see this, let $f \in \mathbb{R}^n$ be an arbitrary function on $\mathcal{X}$. Then $\alpha = K^{-1}f$ ensures that the function $f = \sum_j k(.,x_j)$ satisfies $f(x_i) = f_i$ for all i.

It turns out that string kernels fall in this class:

THEOREM 2.7. *Let $\mathcal{X}$ be a finite set of strings, and let $\#_s(x)$ denote the number of times substring s occurs in x. Then any string kernel of the form $k(x, x') = \sum_{s \in X} w_s \#_s(x) \#_s(x')$ with $w_s > 0$ for all $s \in \mathcal{X}$ is strictly positive definite.*

Proof. We will show that the vectors $\{\phi(x) \,|\, x \in \mathcal{X}\}$ obtained by the feature map are linearly independent, implying that all Gram matrices are nonsingular. The feature map is given by $\phi(x) = (\sqrt{w_s}\#_s(x), \sqrt{w_{s'}}\#_{s'}(x), \ldots)$ where we assume for the purpose of the proof that all substrings s are ordered by nondecreasing length. Now for a given set X of size m consider the matrix with columns $\phi(x_1), \ldots, \phi(x_m)$, where the entries in X are assumed to be ordered in the same manner as the substrings (i.e. by nondecreasing length). By construction, the upper triangle of this matrix is zero, with the highest nonzero entry of each row being $\sqrt{w_x}$, which implies linear independence of its rows.

For graphs unfortunately no strictly positive definite kernels exist which are efficiently computable. Note first that it is necessary for strict positive definiteness that $\phi(x)$ be injective, for otherwise we would have $\phi(x) = \phi(x')$ for some $x \neq x'$, implying that the kernel matrix obtained from $X = \{x, x'\}$ is singular. However, as Gärtner *et al.* (2003) show, an injective $\phi(x)$ allows one to match graphs by computing $\|\phi(x) - \phi(x')\|^2 = k(x,x) + k(x',x') - 2k(x,x')$. Graph matching, however, is NP-hard, hence no such

[4] Size and level are defined following Casella and Berger (2002, Section 8.3).

kernel can exist. That said, there exists a number of useful graph kernels. See e.g. (Borgwardt *et al*., 2005) for further details.

2.4 Kernel choice

So far, we have focused on the case of universal kernels. These kernels have various favorable properties, including that

- universal kernels are strictly positive definite, making the kernel matrix invertible and avoiding non-uniqueness in the dual solutions of SVMs,
- Continuous functions on $\mathcal{X}$ can be arbitrarily well approximated (in the $\|\cdot\|_\infty$-norm) using an expansion in terms of universal kernels, and SVMs using universal kernels are consistent in the sense that (subject to certain conditions) their solutions converge to the Bayes optimal solution (Steinwart, 2002).
- MMD using universal kernels is a test for identity of arbitrary Borel probability distributions.

However, note that for instance in pattern recognition, there might well be situations where the best kernel for a given problem is not universal. In fact, the kernel corresponds to the choice of a prior, and thus using a kernel which does *not* afford approximations of arbitrary continuous functions can be very useful—provided that the functions it does approximate are known to be solutions of the given problem.

The situation is similar for MMD. Consider the following example: suppose we knew that the two distributions we are testing are both Gaussians (with unknown mean vectors and covariance matrices). Since the empirical means of products of input variables up to order two are sufficient statistics for the family of Gaussians, we should thus work in an RKHS spanned by products of order up to two—any higher order products contain no information about the underlying Gaussians and can therefore mislead us. It is straightforward to see that for $c > 0$, the polynomial kernel $k(x,x') = (\langle x,x'+c)^2$, with c > 0, does the job: it equals

$$\sum_{i,j=1}^{d} x_i x_j x'_i x'_j + 2c \sum_{i=1}^{d} x_i x'_i + c^2 = \langle \phi(x), \phi(x') \rangle,$$

where $\phi(x) = (c, \sqrt{2c}x_1, \ldots, \pm\sqrt{2c}x_d, x_i x_j \,|\, i,j = 1,\ldots,d)^\top$. If we want to test for differences in higher order moments, we use a higher order kernel[5] $k(x,x') = (\langle x,x\rangle + c)^p$.

Note, however, that this does not tell us how to choose c. With additional prior knowledge, we could further improve the odds of our test working well on small sample sizes. For instance, if we knew that the Gaussians differ mainly in their covariance structures, then we could incorporate this by choosing a small c. If the available prior knowledge is less specific, we could also sum up several MMDs by using summed kernels.

2.5 Related methods

Various empirical methods have been proposed to determine whether two distributions are different. The first test we consider, and the simplest, is a multivariate generalization of the t-test (Hotelling, 1951), which assumes both distributions are multivariate Gaussian with unknown, identical covariance structure. This test is not model-free in the sense of MMD (and the tests described below)—indeed, it is easy to construct examples in which it fails completely (Figure 1).

Two well-established model-free univariate tests are the Kolmogorov-Smirnov statistic and the Wald-Wolfowitz runs test. Both tests are powerful in that the distribution of the test statistic is known independently of p and q for finite sample sizes, under the null hypothesis $p = q$. A generalization of the Wald-Wolfowitz runs test to the multivariate domain was proposed by Friedman and Rafsky (1979). It involves counting the number of edges in the minimum spanning tree over the aggregated data that connect points in X to points in Y. The resulting test relies on the asymptotic normality of the test statistic. The computational cost of this method using Kruskal's algorithm is $O((m+n)^2\log(m+n))$, although more modern methods improve on the $\log(m+n)$ term. Two possible generalizations of the Kolmogorov-Smirnov test to the multivariate case were studied by Bickel (1969); Friedman and Rafsky (1979). The approach of Friedman and Rafsky in this case again requires a minimal spanning tree, and thus has a similar cost to their multivariate runs test.

Hall and Tajvidi (2002) propose to aggregate the data as $Z = \{X,Y\}$, find the j points in Z closest to each point in X for all $j \in \{1,\ldots,m\}$, count how many of these are from Y, and compare this with the number of points expected under the null hypothesis (the procedure is repeated for each point in Y wrt points in X). The test statistic is costly to compute; Hall and Tajvidi (2002) consider only tens of points in their experiments.

Another approach is to use some distance (e.g. L_1 or L_2) between estimates of the densities as a test statistic (Anderson *et al*., 1994; Biau and Gyorfi, 2005), based on the asymptotic distribution of this distance given $p = q$. One problem with the approach of Biau and Gyorfi (2005), however, is that it requires the space to be partitioned into a grid of bins, which becomes difficult or impossible for high dimensional problems (such as those in Section 3).

We now illustrate these tests with a simple example. In Figure 2, we compare several alternatives to the MMD-based test in distinguishing 100 samples taken from each of two normal distributions with unit variance. Results are averaged over a series of Euclidean distances between the means of both distributions, and plotted as a function of increasing dimensionality. The t-test has the highest chance of correctly rejecting the null hypothesis for low dimensions. However, for high dimensions the estimation of the sample covariance matrices is poor due to the limited sample sizes. Note that we do *not* apply the Biau & Györfi test for high dimensionalities, since memory requirements force the number of partitions per dimension to be too low.

MMD performs very well and outperforms all other model-free approaches, namely the multivariate Kolmogorov-Smirnov test (FR Smirnov), the multivariate Wald-Wolfowitz runs test (FR Wolf), and the Biau & Györfi test (Biau). The comparison becomes harder for increasing dimensionality, since the sample size is fixed to 100 random vectors per distribution for all dimensions. Moreover, MMD also yields a very low rejection rate of the null hypothesis, when it is true (see figure legend).

Finally, we mention that the connection between means in RKHSs and distributions has, in a less general setting, been observed before in the field of kernel machines. Schölkopf and Smola (2002) point out that the empirical mean of a set of points

[5] Kernels with infinite-dimensional RKHS can be viewed as a nonparametric generalization where we have infinitely many sufficient statistics.

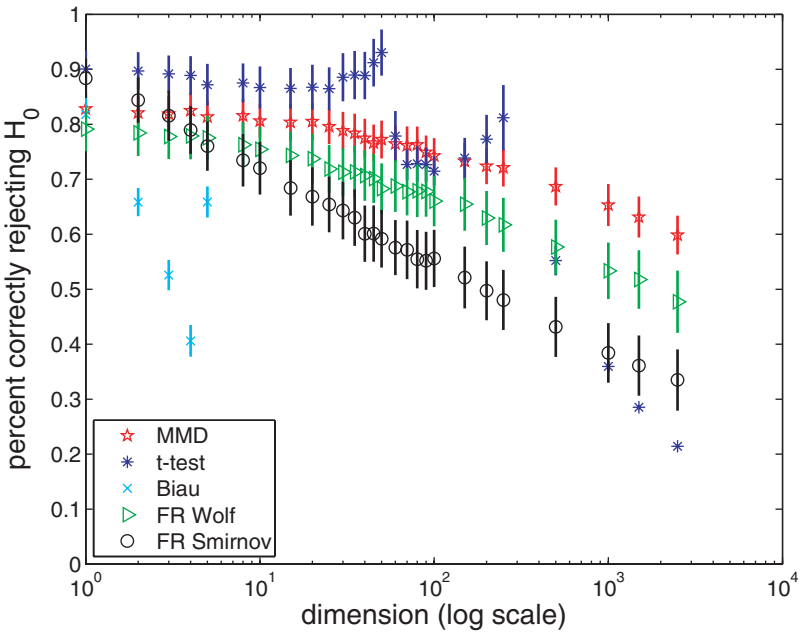

Fig. 2. Test of samples from two normal distributions with different means and unit variance, based on a significance level $\alpha = 0.05$. The cumulative percentage of times the null hypothesis was correctly rejected over the set (0.1,0.2,0.25,0.3,0.4,0.5,0.6,0.7,0.8,1,2,5,10,15) of Euclidean distances between the distribution means, was computed as a function of the dimensionality of the normal distributions. Its average and standard error in 333 repetitions is shown for each of the five tests employed. The sample size was 100 for each distribution. The MMD used a Gaussian kernel, with kernel size σ obtained by maximizing MMD (for σ values within 0.25 and 20) to get the most conservative test. In case of the t-test, a ridge was added to the covariance estimate, to avoid singularity (the ridge was incremented in steps of 0.01 until the 2-norm condition number was below 10). For the Biau test, equal partitions per dimension were used, although this becomes intractable for high dimensions. When samples from distributions with equal mean were compared, the tests wrongly rejected the null hypothesis in the following number of trials out of 8991 (summed over all dimensions in the plot, with 333 runs each): 112 (MMD), 960 (t-test), 379 (FR Wolf), 441 (FR Smirnov). For the Biau test: 4 out of 1665 trials.

in an RKHS can be viewed as a Parzen windows estimate of the density underlying the data; and Shawe-Taylor and Cristianini (2004) propose to use the distance to the mean as a novelty detection criterion, and provide a statistical analysis.

3 EXPERIMENTS

In this section, we present applications of MMD in data integration for bioinformatics, namely microarray cross-platform comparability, cancer (subtype) diagnosis, and schema matching for enzyme protein structures.

3.1 Microarray cross-platform comparability

Experimental scenario Microarrays as a large-scale gene expression observation tool offer a unique possibility for molecular biologists to study gene activity at a cellular level. In recent years, there have been a great number of developments in different microarray platforms, techniques and protocols, advances in these techniques, and biological and medical studies making use of these approaches. As a result, microarray data for a given problem, and the results derived from it (e.g. marker genes for a certain subtype of cancer), may vary greatly (Carter *et al*., 2005), both between labs and platforms. Even for the subsequent step of data processing, e.g. missing value imputation, a large battery of different techniques is available. Consequently, despite an avalanche of microarray data being generated nowadays, it remains to be determined if and how to combine microarray data from different studies on the same biological subject.

Therefore, it is necessary to establish a statistical test of whether two microarray measurements on the same subject, obtained by two different labs or on two different platforms, can be regarded as comparable and can be used for joint data analysis. We define such a test using MMD as a statistic: if an MMD-based test rejects the null hypothesis that the microarray measurements are generated from the same distribution, then we deem them not comparable.

We test this approach on published microarray datasets from two different platforms. If our criterion is useful in practice and able to detect the limited cross-platform comparability of microarray data, then MMD should judge microarray data achieved on different platforms as being less often comparable than those found on the same platform.

Data For our first experiment, we obtained 2 datasets from Warnat *et al*. (2005), from two studies on breast cancer by Gruvberger *et al*. (2001) and West *et al*. (2001). Both comprise gene expression levels for a common set of 2,166 genes. Different microarray platforms were used in these studies: while Gruvberger *et al*. (2001) achieved their results on a c-DNA platform, West *et al*. (2001) utilized oligonucleotide microarrays.

We tried to find out via MMD if there is any statistically significant difference between the microarray results achieved on these different platforms. Samples were scaled to zero mean and unit variance beforehand, although not for the t-test. We compared

Table 1. Microarray cross-platform comparability

Platforms	H_0	MMD	t-test	FR Wolf	FR Smirnov
Same	accepted	100	100	93	95
Same	rejected	0	0	7	5
Different	accepted	0	95	0	29
Different	rejected	100	5	100	71

Cross-platform comparability tests on microarray level for cDNA and oligonucleotide platforms. Repetitions 100, sample size (each) 25, dimension of sample vectors: 2,116

Table 2. Cancer diagnosis

Health status	H_0	MMD	t-test	FR Wolf	FR Smirnov
Same	accepted	100	100	97	98
Same	rejected	0	0	3	2
Different	accepted	0	100	0	38
Different	rejected	100	0	100	62

Comparing samples from normal and prostate tumor tissues (Singh *et al*., 2002). H_0 is hypothesis that $p = q$. Repetitions 100, sample size (each) 25, dimension of sample vectors: 12,600

the MMD results to the multivariate t-test and the Friedman-Rafsky multivariate Kolmogorov-Smirnov and Wald-Wolfowitz tests (denoted Smirnov and Wolf, respectively). The high dimensionality of this problem, as well as of the experiments below, prevents a comparison with the Biau-Györfi test.

We chose $\alpha = 0.05$ as the level of significance for all tests. A Gaussian kernel was employed for MMD, with $\sigma = 20$. We obtained an average performance over 100 distribution tests using 50 microarray measurements from different platforms (X being 25 cDNA measurements and Y being 25 oligonucleotide measurements), and 100 distribution tests with data from 50 microarray measurements taken from only one of the two platforms. For each test, the studies were randomly selected without replacement from the relevant measurement pools. We repeated this experiment for MMD and each of the competing methods.

Results Results are reported in Table 1 , showing the number of times MMD and the other three methods deemed two samples as originating from the same distribution, on data from both identical and dissimilar platforms. In the majority of repetitions, both MMD and the Friedman-Rafsky tests recognize correctly whether two samples were generated on the same platform or not. However MMD is the only test that makes no Type I or Type II errors in all repeats of the experiment. While the FR Wolf test has no false negatives when the samples are from different platforms, it finds occasional false positives when the samples arise from the same platform. The FR Smirnov test has a slightly reduced Type I error rate compared with the FR Wolf test, but at the expense of a much larger Type II rate. Finally, the t-test appears unable to distinguish differences in platform, which is unsurprising given the high dimensionality of the data. As inter-platform comparability of microarray data is reported to be modest in many recent publications (van Ruissen *et al*., 2005; Carter *et al*., 2005; Stec *et al*., 2005), MMD is very successful in detecting these differences in our experiments. We also note that our sample sizes are relatively small, which makes problematic the assumption of both the MMD and Friedman-Rafsky tests that the associated statistic has an asymptotic distribution (this remark also holds for the experiments in the next section). That said, this approximation appears reasonable for the tasks we address, in the light of our results.

3.2 Cancer and tumor subtype diagnosis

Experimental scenario Besides microarray cross-platform comparability, it is interesting to examine whether MMD can distinguish between the gene expression profiles of groups of people who are respectively healthy or ill, or who suffer from different subtypes of a particular cancer. Alternatively, as in the previous experiment, MMD can be employed to determine whether we should integrate two sets of observations (which might arise from different subtypes of a cancer) into one joint set, or if we should treat them as distinct classes.

When using MMD for cancer diagnosis, we test whether the microarray data at hand contain a significant level of difference between ill and healthy patients. Conversely, when looking at cancer (or tumor) subtypes, MMD indicates whether two subtypes of cancer should be considered independently when designing a computational predictor of cancer, or if they can be assigned to one common super-class. In terms of classification methods, MMD can be used to choose whether binary (cancer/healthy) or multi-class (healthy, cancer subtype 1, . . . , cancer subtype n) classification will be more accurate when developing a diagnosis tool.

Data For our second microarray experiment, we obtained datasets from two cancer microarray studies. The first, by Singh *et al*. (2002), is a dataset of gene expression profiles from 52 prostate tumor and 50 normal, non-tumor samples. The second, by Monti *et al*. (2005), consists of microarray data from diffuse large B-cell lymphoma samples. In particular, we are interested in cancer diagnosis on the data of Singh *et al*. (2002), and tumor subtype diagnosis on the data of Monti *et al*. (2005). We again normalized each data sample to zero mean and unit variance, besides for the t-test.

Cancer diagnosis

We examine whether MMD can distinguish between normal and tumor tissues, using the microarray data from the prostate cancer study by Singh *et al*. (2002). Again, α was set to 0.05. Randomly choosing 100 pairs of 25 healthy and 25 cancer patients' gene expression profiles, we used MMD to test the null hypothesis that both samples were generated by the same distribution. We then did the same test for 100 randomly chosen pairs of samples of size 25, both drawn from the same tissue type (healthy or tumor). For all 200 pairs of samples, we compared our results to those of the multivariate t-test and both Friedman-Rafsky tests (Wolf and Smirnov).

Results Results are reported in Table 2. Both MMD and the Friedman-Rafsky tests are in agreement that there is a large difference between samples from cancer patients and healthy patients, and little difference within a particular class. We again see that both MMD and FR Wolf make no Type II errors, but that only MMD makes no Type I errors; and that FR Smirnov has a much higher Type II error rate than FR Wolf (while making one fewer Type I errors).

Table 3. Tumor subtype tests

Subtype	H_0	MMD	t-test	FR Wolf	FR Smirnov
Same	accepted	100	100	95	96
Same	rejected	0	0	5	4
Different	accepted	0	100	0	22
Different	rejected	100	0	100	78

Comparing samples from different and identical tumor subtypes of lymphoma (Monti *et al*., 2005). H_0 is hypothesis that $p = q$. Repetitions 100, sample size (each) 25, dimension of sample vectors: 2,118.

Table 4. Data-content based schema matching

Test	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6
EC 1	0	50	45	50	50	50
EC 2	50	0	50	50	50	50
EC 3	48	50	0	50	50	50
EC 4	50	50	50	0	50	50
EC 5	50	50	50	50	0	50
EC 6	50	50	50	50	50	0

Data-content based schema matching for $\alpha = 0.01$. Numbers indicate how often null hypothesis ($p = q$) was rejected.

Tumor subtype diagnosis

We performed the same experiment as above for tumor subtype diagnosis on data from Monti *et al*. (2005). We are interested in whether MMD is able to distinguish betweeen lymphoma of two subtypes: ''oxidative phosphorylation'' and ''B-cell receptor/proliferation''.

Results We report results in Table 3. As in the previous experiment, both MMD and the Friedman-Rafsky tests prefer to reject the null hypothesis that both samples are generated by the same distribution, when the lymphoma subtypes are different. In other words, all three tests succeed in finding discrepancies between samples from different tumor subtypes in this case. This is consistent with previous results by Monti *et al*. (2005) who discovered these different lymphoma subtypes by using a combination of several clustering algorithms. Hence MMD confirms the existence of these subtypes in our experiment. Comparing the performance of the various tests gives results consistent with the previous two experiments: MMD and FR Wolf do not make any Type II errors, but only MMD has no Type I errors; and FR Smirnov has a much worse Type II error rate than FR Wolf, but makes one fewer Type I errors.

3.3 Schema matching on molecular graph data

Experimental scenario Classifying biological data into ontologies or taxonomies is the central step in structuring and organizing the data. However, different studies may use different ontologies, resulting in the need to find correspondences between two ontologies. We employ MMD to discover matching terms in two ontologies using the data entries associated with these terms.

We study the following scenario: Two researchers have each dealt with 300 enzyme protein structures. These two sets of 300 proteins are disjunct, i.e. there is no protein studied by both researchers. They have assigned the proteins to six different classes according to their enzyme activity. However, both have used different protein function classification schemas for these six classes, and are not sure which pairs of classes correspond.

To find corresponding classes, the MMD can be employed. We obtained 600 proteins modeled as graphs from Borgwardt *et al*. (2005), and randomly split these into two subsets A and B of 300 proteins each, such that 50 enzymes in each subset belong to each one of the six EC top level classes. We then computed MMD for all pairs of EC classes from subset A and subset B to check if the null hypothesis is rejected or accepted. To compute the MMD, we employed the protein random walk kernel for protein graphs, following Borgwardt *et al*. (2005). We compared all pairs of classes via MMD, and repeated the experiment 50 times.

Results For a significance level of $\alpha = 0.05$, MMD rejected the null hypothesis that both samples are from the same distribution whenever enzymes from two different EC classes were compared. When enzymes from the same EC classes were compared, MMD accepted the null hypothesis. MMD thus achieves error-free data-based schema matching here.

We checked whether the same good results were found for a higher significance level of $\alpha = 0.01$. We report results in Table 4. This time, in 7 comparisons out of 1800 comparisons the null hypothesis is incorrectly accepted, whereas in all other cases, the correct decision is taken. Hence even for the high significance level of $\alpha = 0.01$ MMD is very accurate.

In addition to these promising results, note that although we consider the basic case of 1:1 correspondence between classes in our experiment, the fact that MMD uses the kernel trick allows for even more powerful approaches to data-content based schema matching. As kernels are closed under addition and pointwise multiplication, we can test complex correspondences between different classes as well, where one class in schema A corresponds to a combination of classes in schema B. Schema matching for complex correspondences via MMD is a topic of current research.

4 DISCUSSION AND CONCLUSIONS

In this paper, we have presented, to the best of our knowledge, the first principled statistical test for distribution testing and data integration of structured objects, using the Maximum Mean Discrepancy (MMD) as a test statistic. MMD makes use of kernels, and hence is not limited to vector data. As a consequence, MMD is not only applicable to a wide range of problems in molecular biology, but also to common data types in bioinformatics, such as strings and graphs. Kernels for biological data, which have previously been used in classification tasks, can now be employed for distribution testing. Amongst others, these include kernels on protein sequences, protein structures, and microarray time series (Schölkopf *et al*., 2004).

MMD is easy to implement, memory-efficient, and fast to compute. In all of our experiments, it outperformed competing methods (provided the latter were applicable at all, i.e., on vectorial data). We applied our MMD-based test to microarray cross-platform comparability, cancer diagnosis, and data-content based schema matching.

We believe that MMD could also be employed to validate computational simulations of biological processes. If wetlab experiments and simulations generate results and predictions that MMD deems comparable, it is likely that the simulator has produced realistic predictions. This validation procedure will become increasingly relevant as more model-based simulations of microarray data become available (den Bulcke *et al*., 2006).

MMD could also be used for keyplayer gene selection from microarray data. This type of feature selection could be employed to find genes that are involved in a cancer outbreak when looking at gene expression profiles from healthy and cancer patients. MMD would be applied to subsets of genes from two classes of microarrays to find the subset that maximizes the probability that the two classes arise from different distributions. These genes should be studied experimentally in more detail. If, however, MMD cannot find any subset of genes that results in significant differences between healthy and cancer patients, then this might serve as an indicator that the microarrays did not contain the essential genes involved in cancer progress.

ACKNOWLEDGEMENTS

The authors are grateful to Patrick Warnat (DKFZ, Heidelberg) for providing datasets for one of our MMD experiments, and to Olivier Bousquet and Matthias Hein for helpful discussions. This work was supported in part by National ICT Australia and by the German Ministry for Education, Science, Research and Technology (BMBF) under grant no. 031U112F within the BFAM (Bioinformatics for the Functional Analysis of Mammalian Genomes) project which is part of the German Genome Analysis Network (NGFN). National ICT Australia is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council. This work was supported in part by the Austrian Science Fund Fonds zur Förderung der Wissenschaftlichen Forschung (FWF), project # S9102-N04, and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

REFERENCES

N. Anderson, P. Hall, and D. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50:41–54, 1994.

G. Biau and L. Gyorfi. On the asymptotic properties of a nonparametric l_1-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51(11):3965–3973, 2005.

P. Bickel. A distribution free version of the smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics*, 40(1):1–23, 1969.

K. M. Borgwardt, C. S. Ong, S. Schonauer, S. V. N. Vishwanathan, A. J. Smola, and H. P. Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21 (Suppl 1):i47–i56, Jun 2005.

S. L. Carter, A. C. Eklund, B. H. Mecham, I. S. Kohane, and Z. Szallasi. Redefinition of affymetrix probe sets by sequence overlap with cdna microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. *BMC Bioinformatics*, 6(1):107, 2005.

G. Casella and R. Berger. *Statistical Inference*. Duxbury, Pacific Grove, CA, 2nd edition, 2002.

M. Cuturi, K. Fukumizu, and J.-P. Vert. Semigroup kernels on measures. *JMLR*, 6: 1169–1198, 2005.

T. Van den Bulcke, K. Van Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. De Moor, and K. Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7(1):43, 2006.

R. M. Dudley. *Real analysis and probability*. Cambridge University Press, Cambridge, UK, 2002.

R. Fortet and E. Mourier. Convergence de la réparation empirique vers la réparation théorique. *Ann. Scient. École Norm. Sup*., 70:266–285, 1953.

J. Friedman and L. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717, 1979.

T. Gärtner, P.A. Flach, and S. Wrobel. On graph kernels: Hardness results and efficient alternatives. In B. Schölkopf and M. K. Warmuth, editors, *Proc. Annual Conf. Computational Learning Theory*. Springer, 2003.

S. Gruvberger, M. Ringner, Y. Chen, S. Panavally, L. H. Saal, A. Borg, M. Ferno, C. Peterson, and P. S. Meltzer. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res*, 61 (16): 5979–5984, Aug 2001.

P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002.

M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In Z. Ghahramani and R. Cowell, editors, *Proc. of AI & Statistics*, volume 10, 2005.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

H. Hotelling. A generalized t test and measure of multivariate dispersion. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 23–41, 1951.

E. Marshall. Getting the noise out of gene arrays. *Science*, 306(5696):630–631, Oct 2004.

S. Monti, K. J. Savage, J. L. Kutok, F. Feuerhake, P. Kurtin, M. Mihm, B. Wu, L. Pasqualucci, D. Neuberg, R. C. Aguiar, P. Dal Cin, C. Ladd, G. S. Pinkus, G. Salles, N. L. Harris, R. Dalla-Favera, T. M. Habermann, J. C. Aster, T. R. Golub, and M. A. Shipp. Molecular profiling of diffuse large b-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood*, 105(5):1851–1861, Mar 2005.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

B. Schölkopf, K. Tsuda, and J.-P. Vert. *Kernel Methods in Computational Biology*. MIT Press, Cambridge, MA, 2004.

R. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK, 2004.

L. Shi, W. Tong, H. Fang, U. Scherf, J. Han, R. K. Puri, F. W. Frueh, F. M. Goodsaid, L. Guo, Z. Su, T. Han, J. C. Fuscoe, Z. A. Xu, T. A. Patterson, H. Hong, Q. Xie, R. G. Perkins, J. J. Chen, and D. A. Casciano. Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics*, 6 Suppl 2:S12, Jul 2005.

D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1 (2): 203–209, Mar 2002.

A.J. Smola, A. Gretton, and K. Borgwardt. Maximum mean discrepancy. Technical Report NICTA-SML-06-001, National ICT Australia, 2006. URL http://sml.nicta.com.au/smola/papers/SmoGreBor06tr.pdf.

J. Stec, J. Wang, K. Coombes, M. Ayers, S. Hoersch, D. L. Gold, J. S. Ross, K. R. Hess, S. Tirrell, G. Linette, G. N. Hortobagyi, W. Fraser Symmans, and L. Pusztai. Comparison of the predictive accuracy of dna array-based multigene classifiers across cdna arrays and affymetrix genechips. *J Mol Diagn*, 7(3):357–367, Aug 2005.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2002.

F. van Ruissen, J. M. Ruijter, G. J. Schaaf, L. Asgharnegad, D. A. Zwijnenburg, M. Kool, and F. Baas. Evaluation of the similarity of gene expression data estimated with sage and affymetrix genechips. *BMC Genomics*, 6:91, Jun 2005.

P. Warnat, R. Eils, and B. Brors. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, 6: 265, Nov 2005.

M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. r. Olson JA, J. R. Marks, and J. R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA*, 98(20):11462–11467, Sep 2001.