Macrothink Institute™

Journal of Biology and Life Science

2010, Vol. 1, No. 1: E1

Sequence Kernel for Z' Factor

Karol Kozak (Corresponding author)
LMC-RISC, ETH Zurich
Schafmattstr. 18, CH-8093 Zurich, Switzerland
Tel: 41-44-633-34-94 E-mail: karol.kozak@lmc.biol.ethz.ch

Gabor Csucs
LMC-RISC, ETH Zurich
Schafmattstr. 18, CH-8093 Zurich, Switzerland
Tel: 41-44-633-34-94 E-mail: csucs@lmc.ethz.ch

Abstract

RNA interference (RNAi) high-content screening (HCS) enables massive parallel gene silencing and is increasingly being used to reveal novel connections between genes and disease-relevant phenotypes. The application of genome-scale RNAi relies on the development of high quality HCS assays. The Z' factor statistic provides a way to evaluate whether or not screening run conditions (reagents, protocols, instrumentation, kinetics, and other conditions not directly related to the test compounds) are optimized. Z' factor, introduced by Zhang et al., is a dimensionless value that represents both the variability and the dynamic range between two sets of sample control data. This paper describe a new extension of the Z′ factor, which integrates multiple readouts for screening quality assessment. Currently presented multivariate Z' factor is based on linear projection, which may not be suitable for data with nonlinear structure. This paper proposes an algorithm which extends existing algorithm to deal with nonlinear data by using the sequence kernel function. Using sequence kernel methods for projections, multiple readouts are condensed to a single parameter, based on which the screening run quality is monitored. The method is based on Sequence Alignment Kernel, a function reflecting the quantitative measure of match between two sequences.

Keywords: Statistics, Assay Development, Cell Biology, High Content Screening

1. Introduction

Recently, RNA interference (RNAi), a natural mechanism for gene silencing (A. Fire, 1998; G.J. Hannon, 2003), has made its way as a widely used method in molecular biology in both academics and industry. Academic researchers have used RNAi to elucidate gene functions through studying a loss-of-function phenotype. Pharmaceutical and biotech companies have set up libraries for large-scale screens employing thousands of short-interfering RNA- (siRNA) or short hairpin RNA- (shRNA) encoding vectors to identify new factors involved in the molecular pathways of diseases (J. Kurreck, 2005). RNAi may lead to advances not only in drug target identifcation and validation but also in the development of a potential whole new class of RNAi-based therapeutic agents (N. Mahanthappa, 2005). The first clinical trials based on RNAi were initiated to treat patients with age-related macular degeneration (J. Whelan, 2005). RNAi has even been seen as the third class of drug targets after small molecules and proteins (Nature News, 2006). Based on siRNA or shRNA libraries, RNAi HCS enables massive parallel gene silencing to reveal the extent to which interference with the expression of specific genes alters the cell phenotype, and it is increasingly being used to reveal novel connections between genes and disease-relevant phenotypes (P. Zuck, 2004; J.P. MacKeigan, 2005; K. Nybakken, 2005; L. Pelkmans, 2005).

Zhang et al. (1999) explored statistical methods for hit selection in RNAi HCS experiments. The application of genome-scale RNAi relies on the development of high-quality RNAi HCS assays. However, despite a strong need for a theoretically based and easily interpretable quality control (QC) metric in RNAi HCS assays, such a QC metric has yet to be developed. An important QC characteristic in an HCS assay is how much the positive control, tested compounds, and negative controls differ from one another in the assay. This QC characteristic can be evaluated using the comparison of two well types in HCS assays. Signal-to-noise ratio (S/N), signal-to-background ratio (S/B), and Z' factor have been adopted to evaluate the quality of HCS assays through the comparison of two investigated types of wells. However, S/B does not take into account any information on variability; and S/N can capture the variability only in one group and hence cannot assess the quality of assay when the two groups have different variabilities. Zhang et al. (1999) proposed a screening window coefficient called "Z' factor." The advantage of Z factor over S/N and S/B is that it takes into account the variabilities in both compared groups.

High-content screening (HCS) can provide information-rich data sets containing readouts on multiple cellular parameters, such as the number of cells, their shape, and even the intracellular distribution of cellular proteins. To monitor the sustained quality of screening assays, a "classical" Z′ factor as used in screening assays with only a single readout is usually calculated using a preselected image readout (e.g., the intensity of a certain fluorescent stain). At present, there is multivariate Z' factor (A. Kümmel, 2009) to monitor assay quality based on multiple readouts simultaneously. Such a method enables the assessment of the image readouts' suitability to monitor relevant biological effects.

Here, we illustrate the shortcomings of a univariate approach in comparison to a truly multivariate quality assessment of an HCS assay. For the multivariate analysis, we propose a

multivariate Z′ factor as a means to monitor assay quality in HCS data sets. Currently presented multivariate Z' factor is based on linear projection, which may not be optimal for data with nonlinear structure. This paper proposes an algorithm which extends existing Z' factor algorithm to deal with nonlinear data by using the kernel function. Although multivariate Z' factor works well for linear problems, it may be less effective when severe nonlinearity is involved. To deal with such a limitation, nonlinear extensions through sequence kernel functions have been proposed. The main idea of sequence kernel-based methods is to map the input data to a feature space through a nonlinear mapping, where the inner products in the feature space can be computed by a kernel function without knowing the nonlinear mapping explicitly (B. Schoekopf, 2002).

2. Methods

2.1 Z' Factor

Z' Factor experiments are performed on one or more assay plates containing replicate wells designated for background subtraction, negative control samples, and positive control samples. Typically, negative control wells are those in which the cells receive an the appropriate treatment so as to elicit the lowest desired percent response (usually untreated cells); positive control wells are those in which the cells receive an appropriate treatment so as to elicit the maximum desired percent response; background wells are treated the same as the negative control wells, except primary antibody incubation is excluded.

Z' factor is proposed to measure the separation between "tested compound" wells and "negative control" wells or between "positive control" wells and "negative control" wells. Let Z' denote Z' factor. Z' factor is defined as (J.H. Zhang, 1999).

$$Z' = 1 - \frac{3(\sigma_1 + \sigma_2)}{|\mu_1 - \mu_2|} \quad (1)$$

σ_1 = Standard deviation of positive controls

σ_2 = Standard deviation of negative controls

μ_1 = Mean of positive controls

μ_2 = Mean of negative controls

Zhang et al. (1999) further use Z' factor to refer to the parameter between tested compound wells and negative control wells and Z′ factor to refer to the parameter between positive control wells and negative control wells (Fig.1, 2) . **Z' = 1** Indicates an ideal assay. As standard deviations become very small or the difference between signals for positive and negative controls approaches infinity, Z' factor approaches 1. **1 > Z' ≥0.5** Indicates a high quality assay exhibiting a wide separation between signal and background, and low data variability. **0.5 > Z' > 0** Indicates a poor quality assay with marginal distinction between signal and background, and higher data variability. **Z' ≤0** Indicates unreliable data. Assay

conditions are not optimized or the assay is not capable of generating meaningful data. **Z' = -1** There is no distinguishable difference between background signal and sample signal.

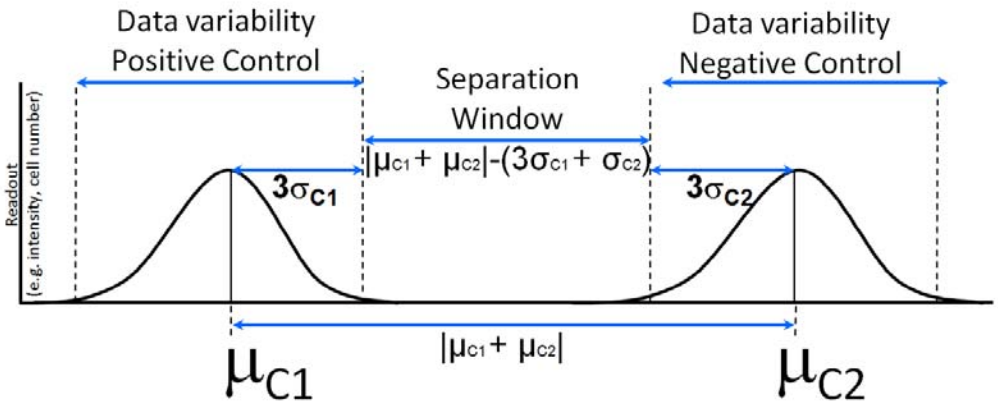

Figure 1. Illustration of data variability band for positive and negative and separation window

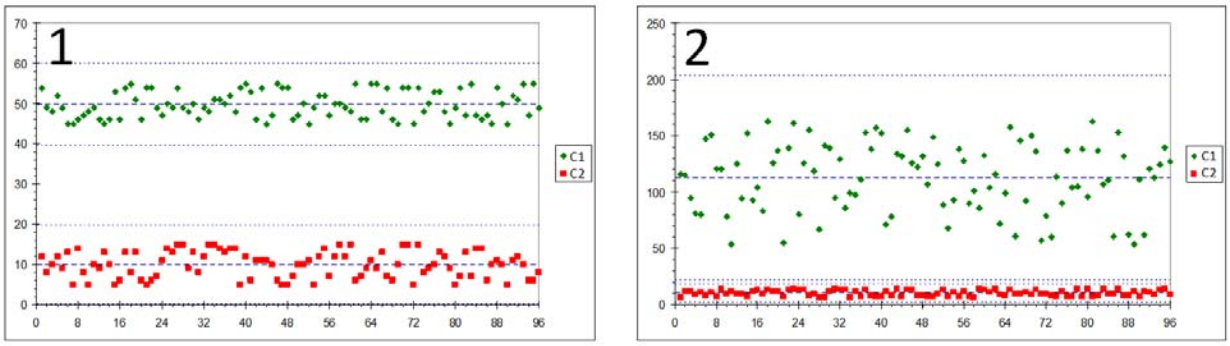

Figure 2. Quality measurement of two assays. Assay 1: C1 mean 50, C2 mean 10 S/B = 5, S/N = 13, Z' = 0.5. Assay 2: C1 mean 112, C2 mean 10 S/B = 11, S/N = 39, Z' = 0.0

2.2 Linear Projection

For the multivariate analysis, a multivariate Z′ factor has been proposed [11] as a means to monitor assay quality in HCS data sets. For multivariate Z′ factor linearly project data points from multidimensional space onto a line has been used. The values derived from these projections were then used to calculate a *multivariate* Z′ factor. The projected value of a data point, *Pn*, is calculated by a weighted sum of the original data values, x_{nj}, for all parameters *j* (with *D* as the number of assay parameters). Analog to the original calculation, the Z′ factor was determined based on the means and standard deviations of the projected values, *P*, for each group (positive and negative).

$$P_n = \sum_{j=1}^{D} w_j \cdot x_{n\,j} = \sum \quad over\ all\ paramters \tag{2}$$

(projection weight of parameter j, x value for parameter j of data point i)

μ_1 - mean (P_1) with $P_1 = \{P_n | n \in$ positive controls$\}$

μ_2 - mean (P_2) with $P_2 = \{P_2 | n \in$ negative controls$\}$

σ_{P1} = Standard deviation P_1

σ_{P2} = Standard deviation P_2

The projection weights, w_j, is calculated by linear discriminant analysis (LDA) as it turned out to robustly yield the highest possible Z' factors and is a widely used method for projection.

2.3 Kernel methods for multivariate Z' factor

Multivariate Z' factor based on linear projection, may not be optimal for data with nonlinear structure. LDA-based algorithms take the class structure into account and focus on the most discriminant feature extraction. The performance of LDA, however, is often degraded by the fact that its separability criterion is not directly related to the classification accuracy in the transformed space. Instead, the LDA optimization is based on the assumption that the intraclass distributions are all Gaussian with a common variance. In other words, the LDA assumes, aside

from the linearity of the subspace, a linear separation between classes in the low-dimensional space. (There are many generalizations of the LDA optimization principle but they all impose parametric models on the within-class distributions). The kernel trick can be utilized to form classification algorithms that are based on nonlinear subspaces (Fig. 3). The basic methodology is to (implicitly) apply a nonlinear mapping on the input image processing parameters and then apply linear methods on the resulting feature space. But, due to its limitation of linearity, LDA fails to perform well for nonlinear problems.

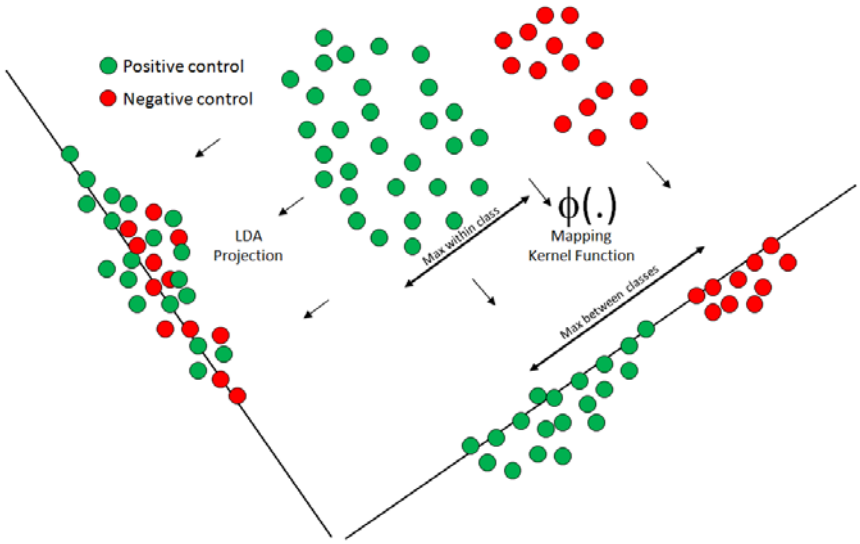

Figure 3. Kernel Projection in compare to LDA Projection finds a linear transformation of predictor variables which provides a more accurate discrimination (right). LDA find the direction to project data on so that – between class variance is maximized and – within class variance is minimized

So the question is, how do we utilize the label information in finding informative projections? Fisher-LDA Objective is to find the vector w to maximize J(w):

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \tag{3}$$

where S_B is the "between positive and negative control scatter matrix", S_W is the "within controls scatter matrix" and w^T is a vector transpose. Note that due to the fact that scatter matrices are proportional to the covariance matrices we could have defined J using covariance matrices – the proportionality constant would have no effect on the solution. The definitions of the scatter matrices are:

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \tag{4}$$

$$\mu_1 = \frac{1}{N_1} \sum_{i \in psoitive} x_i \, , \mu_2 = \frac{1}{N_2} \sum_{i \in negative} x_i \tag{5}$$

$$S_W = S_1 + S_2 \tag{6}$$

$$S_i = \sum_{x \in class\, i} (x - \mu_C)(x - \mu_C)^T \tag{7}$$

$$i \in (positive\ control,\ negative\ control)$$

Where N is a total number of compounds, N_1 is the number of compounds as positive controls and N_2 is the number of negative controls. The solution of this optimization problem is written as:

$$w = S_W^{-1}(\mu_1 - \mu_2) \tag{8}$$

In order to solve equation 8, we need to know μ_1, μ_2, and **S***w*. The class-mean vectors (μ_1 and μ_2) and the within-class variance matrix (**S***w*) can be obtained if a decision boundary vector (equivalently, a threshold value in this article) is known. In other words, we need to know a decision boundary for the purpose of obtaining a decision boundary. To address this issue, we simply chose a single threshold value to obtain two classes and then computed μ_1, μ_2, and **S***w*.

In many practical cases linear discriminants are not suitable. Fisher LDA discriminant can be extended for use in non-linear classification via the kernel methods. In Kernel methods, the original observations are effectively mapped into a higher dimensional non-linear space. For a given nonlinear mapping ϕ, the input data space X can be mapped into the feature space H:

$$\phi : X \to H \ \ where\ x \to x : \phi(x). \tag{9}$$

Linear classification in this non-linear space is then equivalent to non-linear classification in the original space. Require Fisher LDA can be rewritten in terms of dot product.

$$K(x_i, x_j) = \phi(x_i) \bullet \phi(x_j) \tag{10}$$

Unlike Support Vector Machine (SVM) it doesn't seem the dual problem reveal the kernelized problem naturally. But inspired by the SVM case we make the following key assumption,

$$w = \sum_i \alpha_i \phi(x_i) \tag{11}$$

In terms of new vektor α the objective $J(\alpha)$ becomes,

$$\arg\max_{\alpha \in R^n} J(\alpha) = \frac{\alpha^T S_B^\phi \alpha}{\alpha^T S_W^\phi \alpha} \tag{12}$$

Correspondingly, a pattern in the original input space R^n is mapped into a potentially much higher dimensional feature vector in the feature space H. The scatter matrices in kernel space can expressed in terms of the kernel only as follows:

$$S_{\mathbf{B}}^\phi = [K_1 K_1^T - K K^T] + [K_2 K_2^T - K K^T] \tag{13}$$

$$S_W^\phi = K^2 - (N_1 K_1 K_1^T + N_2 K_2 K_2^T) \tag{14}$$

$$K_1 = \frac{1}{N_1} \sum_{im \in psoitive} K_{im}, K_2 = \frac{1}{N_2} \sum_{i,m \in negative} K_{im} \tag{15}$$

$$K = \frac{1}{N} \sum_{i,j \in N} K_{ij} \tag{16}$$

In this paper we make use of Sequence Alignment Kernel (Watkins, 2000; Surkov *et al*., 2001) to define the measure of similarity between two oligo sequences. Our method is based on building the *kernel function K(*s_i, s_j*)* as a quantitative measure of similarity between two target gene mRNA of observer siRNA sequences *R* and *Q*. Suppose we are given a matrix Swap*(a, b)* which defines the score corresponding to a single point mutation of letter *a* into letter *b* or vice versa (the matrix is symmetric). We are also given a vector Gap*(a)* which defines the score corresponding to a single point deletion or insertion of letter *a*. One of the schemes for simultaneous generation of two sequences over a given alphabet was proposed by Watkins (2000). The generative model may emit either two letters (one into each sequence), only one letter into the first sequence (which corresponds to a gap into the second one), or only one letter into the second sequence (which corresponds to a gap into the first one). The model is completely defined by the probabilities for each pair it may emit. For any two nonempty sequences there are several ways (or paths) to generate them using this model.

For every such path the corresponding probability is the product of probabilities along the path. The total probability *P(a, b)* that the sequences *a* and *b* will be generated by the model is the sum of probabilities of all the paths that lead to generating the given pair. Watkins (2000) has proven that the function *P(a, b)* is symmetric and positively definite, and so may be used as a kernel for kernel-based algorithms.

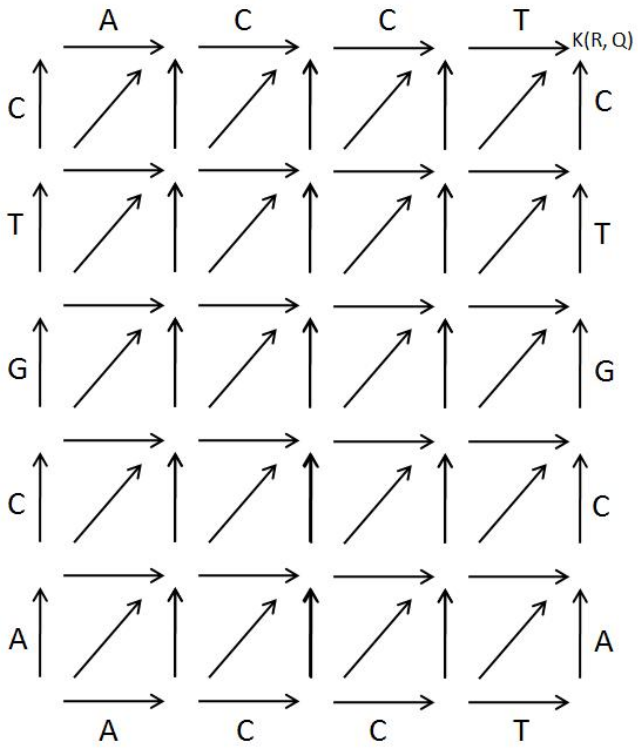

Figure 4. Version of sequence alignment kernel

Suppose we are given two sequences to align, *Q* = '*ACCT* ' and *R* = '*ACGT C*'. Let us write them along the two dimensions of an empty matrix (Fig. 4). In each cell $p_{i,j}$ of the matrix we will be keeping the probability that $Q_1 \ldots j$ aligns with $R_1 \ldots i$. It is convenient to start the calculations from the bottom left corner, which is initialized with the value of 1. Then, we fill all the other cells using the recursive formula:

$$p_{0,0} = 1,\ p_{i,0} = p_{0,j} = 0, \text{ for } i > 0 \text{ and } j > 0,$$
$$p_{i,j} \leftarrow (\text{Swap}(R_i, Q_j) \cdot p_{i-1,j-1}) + (\text{Gap}(Q_j) \cdot p_{i,j-1}) + (\text{Gap}(R_i) \cdot p_{i-1,j}) \quad (17)$$

where the Swap(a, b) matrix and the Gap(a) vector of probabilities are given as parameters to the algorithm. The kernel value we are looking for is the probability

$$K = p_{|R|,|Q|} \quad (18)$$

in the top right corner of the matrix. Note, that to calculate values on any 'backslash' diagonals of the *p* matrix ($i + j = D$) we only need to know the values on the two preceding diagonals: $i + j = D - 1$ and $i + j = D - 2$.

So, we have managed to express the problem in terms of kernels only which is what we were

after (Fig. 4). Note that since the objective in terms of α has exactly the same form as that in terms of w, we can solve it by solving the generalized eigenvalue equation. This scales as N^3 which is certainly expensive for many datasets. More efficient optimization schemes solving a slightly different problem and based on efficient quadratic programs exist in the literature. Projections of new test-points into the solution space can be computed by,

$$w^T \phi(x) = \sum_{ij \in controls} \alpha_{ij} K(x_i, x_j) \tag{19}$$

In order to classify the test point we still need to divide the space into regions which belong to one class. Alternatively, one could train any classifier in the 1-d subspace.

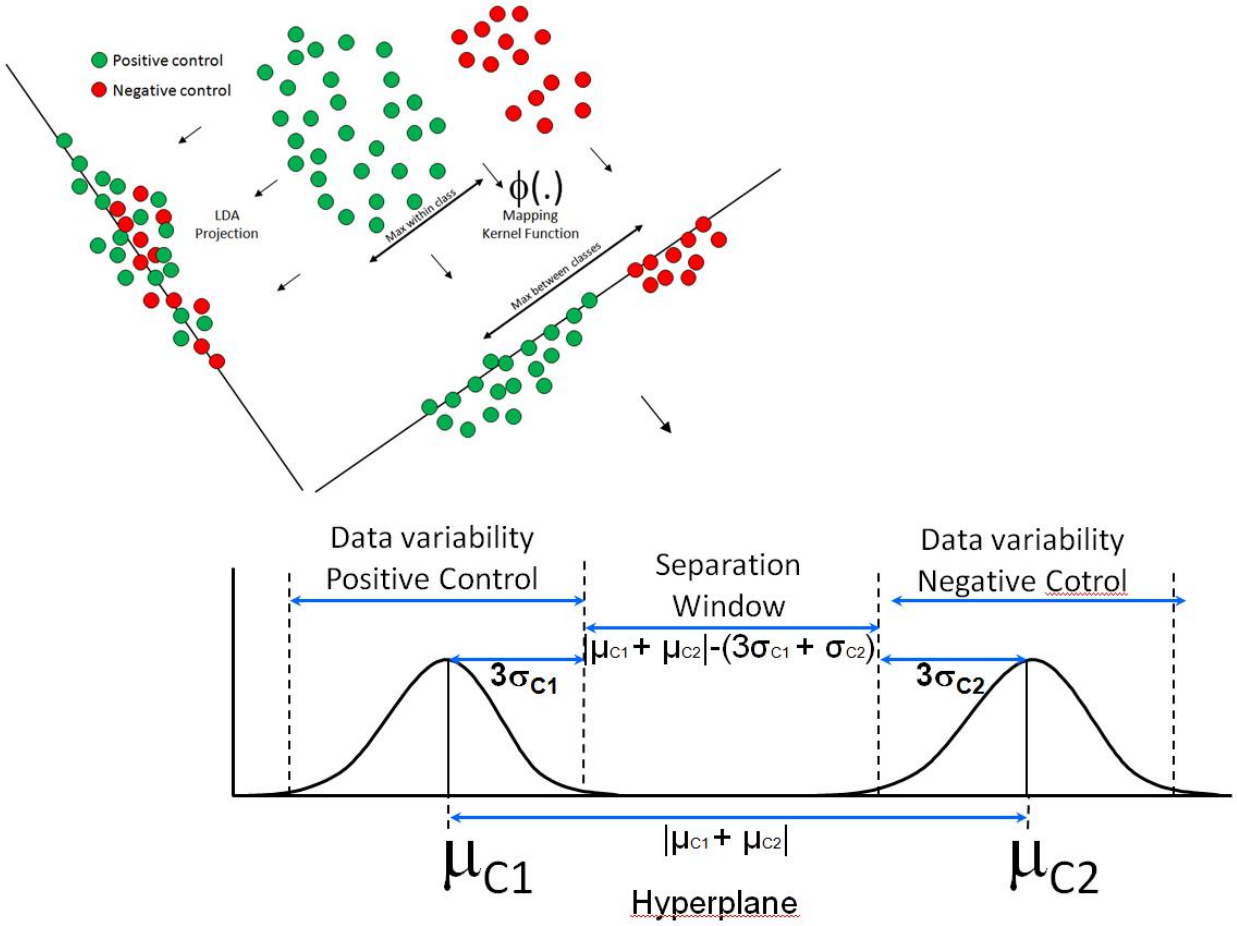

Figure 5. Z′ factors based on separate parameters and a kernelized projection

3. Results

A good quality control (QC) metric should work in a variety of experiments. Thus in this paper we concentrate on plates extracted from different RNAi HCS experiments, which may have different data ranges and different numbers of positive and negative control wells, so that we can see the impact of sample size and data range on the QC metrics.

Table 1. Z′ Factors for the Biogenesis Screen

Z′ Factor, Multivariate Z’ factor, Kernel Z’ factor for Biogenesis Screen			
Methods	Z′ – Factor	Multivariate Z’ factor	SeqKernel Z’ factor
Number of cells	0.694	x	X
Intensity nuclei	0.583	x	X
Biogenesis screen (27 parameters)	x	0.756	0.820
Biogenesis screen (15 parameters)	x	0.546	0.722
Biogenesis screen (7 parameters)	x	0.681	0.652

For the purpose of validating the new assay quality method and showing its usability in actual research projects it has been applied to experimental data from biogenesis High Content Screening. The biogenesis project is dealing with ribosomes, which are macromolecular complexes used to synthesize proteins. Ribosomes are divided into a small and a large subunit, both consisting of ribosomal proteins and ribosomal RNA (rRNA). In eukaryotes, the biogenesis of these subunits is a complex multistep process including the assembly of different component to the subunits in the nucleolus, the export of these precursors in the cytoplasm, and the final maturation and fusion of both subunits to a functional complex. In this project the biogenesis of the small ribosomal subunit (40S subunit) is studied in human cells by performing a genome-wide siRNA screen detecting 2 classes positive/negative using image processing. Follow results has been provided from experiment: 27 image features, 500 cells per well, 3 channels, 4 oligonucleotides, Total number of observations (records,-rows) = 108324. Image processing parameters: 1:green mean intensity nuclei, 2: green std intensity nuclei, 3: green mean intensity cytoplasm, 4: green std intensity cytoplasm , 5: green mean intensity cells, 6: green std intensity cells, 7: blue mean intensity nuclei, 8: blue std intensity nuclei, 9: blue mean intensity cytoplasm, 10: blue std intensity cytoplasm, 11: blue mean intensity cells, 12: blue std intensity cells, 13-27: nuclei texture green. The use of a specific assay enables the visual detection of nuclear 40S maturation defects upon depletion of a protein by RNAi. In total, 17 632 genes and 5 318 predicted genes are targeted by four different oligos.

Average Z’-score values for this screen are represented in Table 1 and calculation of single sample (multiwall plates) on Figure 6. The kernel Z’ factor was much higher than the best Z′ factor (i.e., 0.65 for “Cell number parameter”; Table 1), demonstrating the superiority of a kernel based multiparametric analysis to discriminate between controls. The bottom of Fig. 5 shows the data from plates 1–16 of genome wide biogenesis experiment. The data have the following notable features. The numbers of control wells at each plate differed in each of the three experiments: 16 positive control wells and 16 negative control wells in Experiment. The measured number of cells in −log2 scale were roughly symmetric with a few outliers Plate003, Plate 005, Plate 0014). The ability to discriminate between positive and negative

controls in the cell cycle data set was assessed with the Z′ factor based on an Kernel projection. Plate 005 had small distance between positive and negative control for cell number parameter. Respectively Z' factor for cell number was very weak for this plate. Nevertheless, considering all 37 parameters kernel Z' factor gave significant score with good quality value.

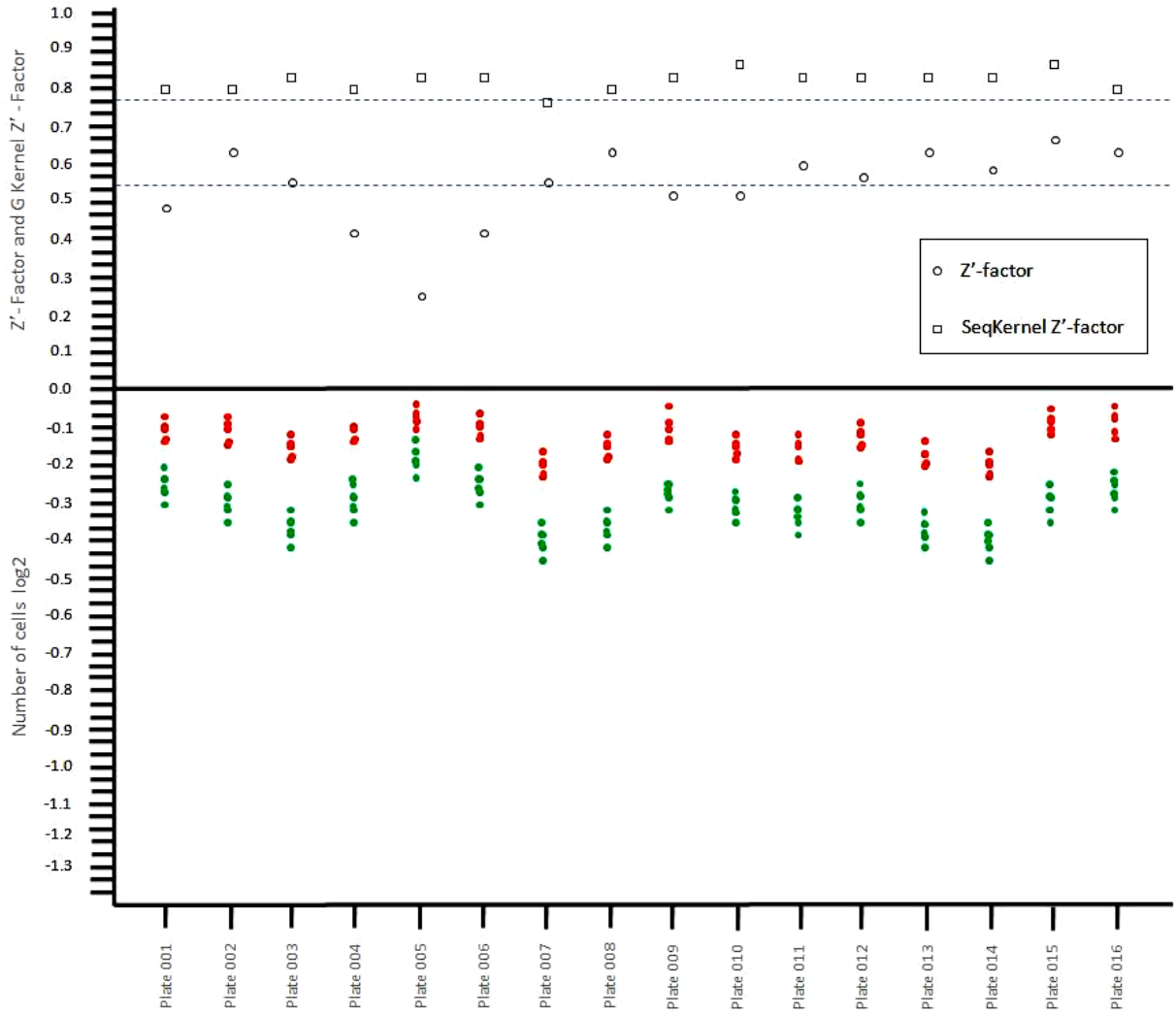

Figure 6. Data from positive (green dots) and negative (red dots) controls (shown at the bottom) and the estimated values of kernel Z factor (shown at the top) for 16 plates from biogenesis screen At the bottom, a red (or green) point represents the measured intensity of a positive (or negative) control well in a plate. At the top, small squares represent the seqkernel z' factor. A black circle represents the estimated (or robust) value of Z' factor in a plate.

4. Discussion

To obtain high-quality HCS assays, there is a strong need for a generally acceptable QC metric that can be applied to various HCS assays conducted in different labs and/or at various

times. There is a general limitation to the traditional Z'factor algorithm where it will fail to measure HCS assay quality. This occurs when we consider only one parametric analysis. For an unbiased hit selection, other data-mining approaches that do not rely on predefined sample classes as positive controls should be used to fully exploit the multidimensional response/parameter space. In these cases, the Kernel based projection can still be used in assay development for studying assay robustness and for exploring the effect of positive controls via the projection weights. The Z′ factor obviously only reflects (1) whether the selected positive controls can be discriminated from negative control and (2) based on which effect(s) they are discriminated. This metric should have a solid theoretical basis and clear probability meanings. The classical Z' factor for testing quality based on single parameter cannot work well as a QC metric in RNAi HCS assays as demonstrated in this paper. The currently available multivariate Z' factor have disadvantages for nonlinear data. In this paper, a sequence kernel Z' factor is proposed for measuring the magnitude of difference between two groups providing better projection and is then investigated for evaluating the quality of RNAi HCS assays. Kernel methods such as SVM achieve state-of-the-art results, in the case of sequence kernel Z' factor the performance improvement in assay quality tasks over linear methods was also found to be significant. Kernel based Z' factor turns out to be more effective than LDA in various applications; however, the sequence kernel Z' factor algorithms are not as simple and transparent as one parametric Z' Factor. It is the complicated formalization of sequence kernel Z' factor algorithms that covers the intuitive characteristics of kernel discriminant analysis. Possible extensions to the approach suggested here include using graph kernel method design for specific screening data. Furthermore, the kernel calculations are not limited to the Z′ factor but can also be used to calculate any similarity between sequences.

References

A. Fire, S. Xu, M.K. Montgomery, S.A. Kostas, S.E. Driver, C.C. Mello. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans, *Nature,* 391 806–811.

A. Kümmel, H. Gubler, P. Gehin, M. Beibel, D. Gabriel and C. N. Parker. (2009). Integration of Multiple Readouts into the Z' Factor for Assay Quality Assessment. *J Biomol Screen*, 2010; 15; 95 originally published online Nov 25.

B. Schoekopf and A. Smola. (2002). *Learning with Kernels*: Support Vector Machines, Regularization, Optimization and Beyond. MIT Press.

G.J. Hannon, RNA. (2003). *A Guide to Gene Silencing*, Cold Spring Harbor Laboratory Press, New York.

J. Kurreck. (2005). RNA interference: perspectives and caveats, RNA Interference Gene Silencing, 1 50–51.

J. Whelan. (2005). First clinical data on RNAi, Drug Discovery Today, 10 1014–1015.

J.H. Zhang, T.D.Y. Chung, K.R. Oldenburg. (1999). A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen*, 4 67-73.

J.P. MacKeigan, L.O. Murphy, J. Blenis. (2005). Sensitized RNAi screen of human kinases and phosphatases identifies new regulators of apoptosis and chemoresistance. *Nat. Cell Biol*., 7, 591–600.

K. Nybakken, S. Vokes, T.Y. Lin, A.P. McMahon, N.A. Perrimon. (2005). Genome-wide RNA interference screen in Drosophila melanogaster cells for new components of the HH signalling pathway. *Nat. Genet*., 37, 1323–1332.

L. Pelkmans, et al. (2005). Genome-wide analysis of human kinases in clathrinand caveolae/raft-mediated endocytosis. *Nature*, 436, 78–86.

N. Mahanthappa. (2005). Translating RNA interference into therapies for human diseases. *Pharmacogenomics,* 6, 879–883.

Nature News. (2006). Silent running: the race to the clinic. *Nature,* 442, 614–615.

P. Zuck, et al. (2004). A cell-based β-lactamase reporter gene assay for the identification of inhibitors of hepatitis C virus replication. *Anal. Biochem*. 334, 344–355.

Surkov, D., Chervonenkis, A. and Gammerman, A. (2001). Kernel for protein sequences classification. Technical Report CLRC-TR-01-08, Computer Learning Research Centre, Dept. of Computer Science, Royal Holloway, University of London.

Watkins,C. (2000). Dynamic alignment kernels. In Smola, A. Bartlett, P. Schölkopf, B. and Schuurmans, D. (eds.), Advances in Large Margin Classifiers, MIT Press, Cambridge, MA, pp. 39–50.