BMC Bioinformatics

Methodology article

Open Access

Evaluating different methods of microarray data normalization

André Fujita[1,2], João Ricardo Sato[1], Leonardo de Oliveira Rodrigues[2], Carlos Eduardo Ferreira[1] and Mari Cleide Sogayar*[2]

Address: [1]Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010 – São Paulo, 05508-090 SP, Brazil and [2]Chemistry Institute, University of São Paulo, Av. Lineu Prestes, 748 – São Paulo, 05513-970 SP, Brazil

Email: André Fujita - fujita@ime.usp.br; João Ricardo Sato - jsato@ime.usp.br; Leonardo de Oliveira Rodrigues - leonardo@iq.usp.br; Carlos Eduardo Ferreira - cef@ime.usp.br; Mari Cleide Sogayar* - mcsoga@iq.usp.br

* Corresponding author

Published: 23 October 2006

BMC Bioinformatics 2006, **7**:469 doi:10.1186/1471-2105-7-469

Received: 12 May 2006
Accepted: 23 October 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/469

Abstract

Background: With the development of DNA hybridization microarray technologies, nowadays it is possible to simultaneously assess the expression levels of thousands to tens of thousands of genes. Quantitative comparison of microarrays uncovers distinct patterns of gene expression, which define different cellular phenotypes or cellular responses to drugs. Due to technical biases, normalization of the intensity levels is a pre-requisite to performing further statistical analyses. Therefore, choosing a suitable approach for normalization can be critical, deserving judicious consideration.

Results: Here, we considered three commonly used normalization approaches, namely: Loess, Splines and Wavelets, and two non-parametric regression methods, which have yet to be used for normalization, namely, the Kernel smoothing and Support Vector Regression. The results obtained were compared using artificial microarray data and benchmark studies. The results indicate that the Support Vector Regression is the most robust to outliers and that Kernel is the worst normalization technique, while no practical differences were observed between Loess, Splines and Wavelets.

Conclusion: In face of our results, the Support Vector Regression is favored for microarray normalization due to its superiority when compared to the other methods for its robustness in estimating the normalization curve.

Background

DNA microarray technology is a powerful approach for genomic research, playing an increasingly important role in biomedical research. This technology yields simultaneous measurement of gene expression levels of thousands of genes, allowing the analysis of differential gene expression patterns under different conditions such as disease (pathological) states or treatment with different chemotherapeutic drugs. Due to small differences in RNA quantities and fluctuations generated by the technique, the intensity levels may vary from one replicate to the other due to effects which are unrelated to the genes, requiring data normalization before they can be compared.

Therefore, normalization is an important step for microarray data analysis. The purpose of data normalization is to minimize the effects caused by technical variations and, as a result, allow the data to be comparable in order to

find actual biological changes. Several normalization approaches have been proposed, most of which derive from studies using two-color spotted microarrays. Some authors proposed normalization of the hybridization intensity ratios; others use global, linear methods, while others use local, non-linear methods. Several authors suggested using the spike-in controls, housekeeping genes, or invariant genes [1-7].

Recently, some authors suggested the use of non-linear normalization methods [8-10] which are believed to be superior to the above mentioned approaches. The locally weighed regression Lowess procedure [11] has been widely used for this purpose and implemented by several microarray analysis software packages [12,13], but similar methods are suggested such as Splines [14,15] and Wavelets [16].

Here, we compare three different well-known microarray data normalization methods, namely: Loess Regression (LR), Splines Smoothing (SS) and Wavelets Smoothing (WS). In addition, we propose two different normalization approaches, called Kernel Regression (KR) [17,18] and Support Vector Regression (SVR) [19], which, to the best of our knowledge, have yet to be applied for microarray normalization. In order to assess the most appropriate normalization technique, benchmark studies were carried out using data derived from CodeLink™ mouse microarray experiments [20], generated at our Cell and Molecular Biology Laboratory (Chemistry Institute, University of São Paulo).

Results

We sought to highlight the performance of five different methods of microarray normalization, namely: Loess, Splines, Wavelets, Kernel and Support Vector Regression in a simulated microarray and in an actual CodeLink™ microarray platform, which comprised ten thousand mouse genes. Although we have focused on the use of simulated two-color cDNA microarray data analysis, our discussions are also applicable to the single-color oligonucleotide microarrays.

The artificial microarrays composed by ten thousand spots were generated using the model proposed by Balagurunathan et al. (2002) [21]. The parameters used were: ($a_0^1 = 0$, $a_1^1 = 100^{1/0.7}$, $a_2^1 = -0.7$, $a_3^1 = 1$) and ($a_0^2 = 0$, $a_1^2 = 100^{1/0.9}$, $a_2^2 = -0.9$, $a_3^2 = 1$) for sinusoid shape, ($a_0^1 = 0$, $a_1^1 = 500$ $a_2^1 = -1$, $a_3^1 = 1$) and ($a_0^2 = 0$, $a_1^2 = 10$, $a_2^2 = -1$, $a_3^2 = 1$) for banana shape and, ($a_0^1 = 0$, $a_1^1 = 10$, $a_2^1 = -1$, $a_3^1 = 1$) and ($a_0^2 = 0$, $a_1^2 = 100^{1/0.7}$, $a_2^2 = -0.7$, $a_3^2 = 1$) for mixed shape. Gene expression was generated by an exponential distribution with parameter $\lambda = 1/3000$ and the outliers were generated by a Beta distribution with parameters $B(1.7,4.8)$. For more details, see Balagurunathan et al. (2002).

The smoothing parameters used in each dataset are described in Table 1. For SVR, we tested a range of values and, as a result, we selected $\varepsilon = 0.01$ and $C = 4$ as the most adequate one. It is important to highlight that the parameters are arbitrary; therefore, we chose the optimum parameters for each method, i.e., the one which resulted in the lowest mean square error. In Figure 1 are described the mean square errors for each normalization method applied to three different simulated microarrays with no outliers.

In order to compare the perturbation caused by the presence of outliers and the robustness of each normalization method, we randomly inserted 5, 10, 15, 20 and 40% of outliers (genes which display very high differential expression) at three different expression levels (low, medium, high), and the respective mean square errors between the regression curve and the actual curve (the function from which the microarray was generated) was calculated. This step was repeated 100 times to estimate the average sum of the squared errors and their variance. The Wilcoxon and the Kolmogorov-Smirnov tests were performed in order to determine whether the five regression methods differ from one another in any significant manner.

A high performance normalization technique should yield unbiased corrections and corrections with the smallest standard deviation.

Comparison of the results presented in Table 2, 3 and 4 shows no important difference between LR, SS and WS. Although the non-parametric KR method has been successfully applied in econometrics data analysis [22], it displayed a poor performance for microarray normalization, probably because it is highly sensitive to outliers [23].

Upon analyzing Table 2, it is possible to observe, in the case of sinusoid shape, that when outliers are inserted in regions of low gene expression, SVR, WS, SS, LR and KR, in this order, have the lowest to the highest mean square error, being statistically different (p value < 0.001) from one another. For the banana and mixed shapes, LR and SS presented a lower MSE than WS. In Table 3, it is interesting to note that when outliers are inserted in regions of medium gene expression, i.e., high density of genes, the order of performance remains the same as in Table 2 and SVR displays a mean square error which is significantly different from the others (p value < 0.001). LR and SS showed no significant difference (p value > 0.05) and KR

Table 1: Smoothing parameters used for each microarray dataset. For Loess, it is the span value, for Splines and Wavelets it is the number of functions, for Kernel and SVR it is the maximum value minus the minimum value multiplied by the number described in the table.

	Banana	Sinusoid	Mixed
Loess	0.30	0.10	0.10
Splines	10.00	20.00	20.00
Wavelets	16.00	64.00	16.00
Kernel	0.50	0.50	0.70
SVR	0.20	0.20	0.60

is significantly worse than the other methods (p value < 0.001). In Table 4, the outliers are inserted in a high gene expression region. Once more, the trend is maintained, namely, KR is the most affected by outliers (p value < 0.001) and no differences between SS and WS (p value > 0.05) were observed for the sinusoid shape. For the other two shapes, LR and SS were better than WS (p value < 0.001).

In all three cases (outliers at low, medium and high gene expression), SVR is the affected by outliers (p value < 0.001), independently the microarray's shape. In addition, SVR yields the smallest standard deviation, followed by LR, SS, WS, with KR displaying the largest deviation. In addition, the five methods were applied to actual microarray data, with outliers inserted artificially, and the results were the same when compared to those obtained from artificial microarray experiments.

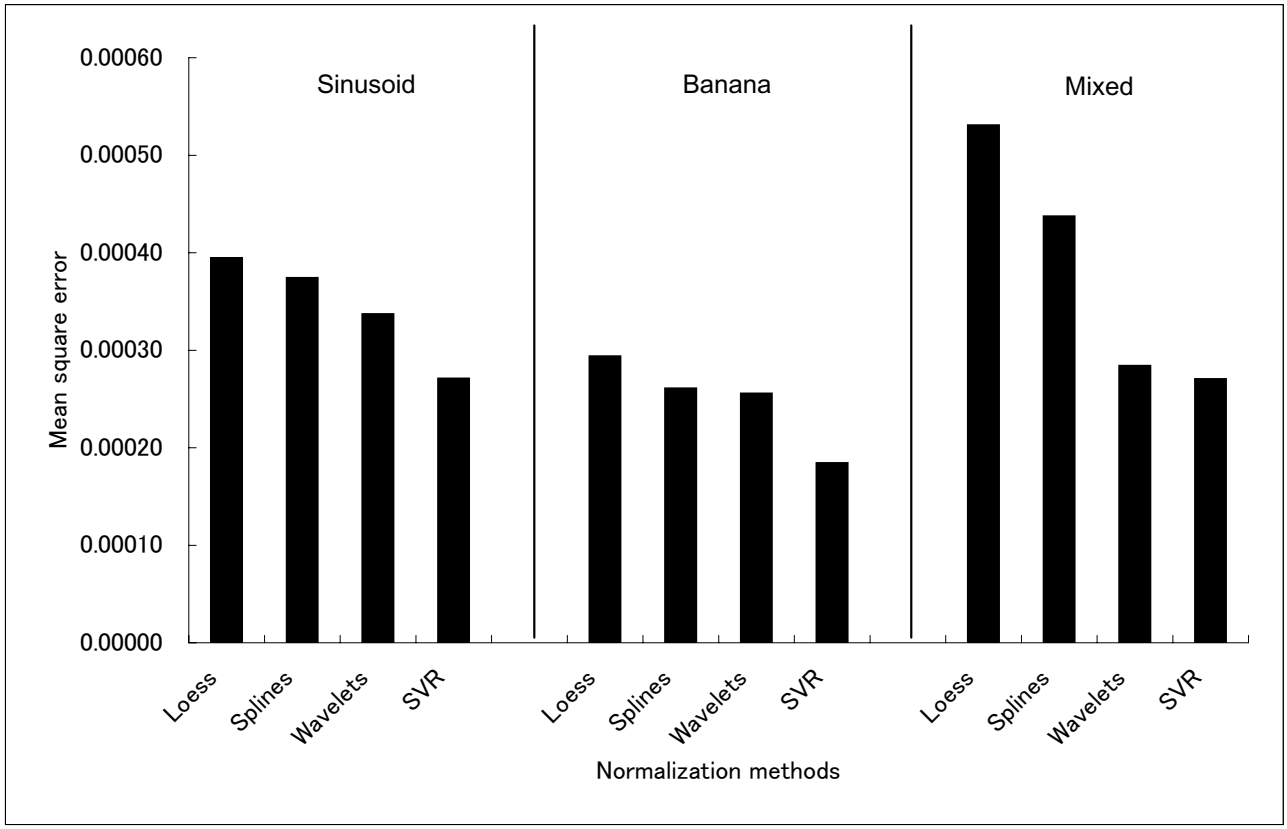

Figure 1
The minimum mean square error for three different simulated microarray datasets. From left to right: 1) sinusoid shape; 2) banana shape; 3) mix shape. The Kernel regression was not included in this figure because its MSE is 10^3 orders of magnitude greater than the other normalization methods.

Table 2: The mean square errors of estimated gene expression levels for simulated cDNA microarray data with differentially expressed genes inserted under the low expression levels condition.

Percentage of DEG		Sinusoid			Banana			Mixed		
	Method	25% Quantile	Median	75% Quantile	25% Quantile	Median	75% Quantile	25% Quantile	Median	75% Quantile
5%	Loess	0.00038	0.00039	0.00040	0.00029	0.00029	0.00029	0.04806	0.04816	0.04825
	Splines	0.00035	0.00036	0.00037	0.00027	0.00028	0.00028	0.04828	0.04839	0.04848
	Wavelets	0.00033	0.00034	0.00035	0.00127	0.00128	0.00128	0.04816	0.04827	0.04835
	Kernel	0.03781	0.03782	0.03783	0.14368	0.14404	0.14416	0.19869	0.19888	0.20098
	SVR	0.00031	0.00032	0.00033	0.00016	0.00016	0.00017	0.04729	0.04733	0.04738
10%	Loess	0.00047	0.00048	0.00050	0.00038	0.00038	0.00039	0.04631	0.04646	0.04661
	Splines	0.00044	0.00045	0.00047	0.00037	0.00038	0.00039	0.04649	0.04663	0.04678
	Wavelets	0.00042	0.00043	0.00045	0.00117	0.00118	0.00118	0.04637	0.04651	0.04666
	Kernel	0.03780	0.03781	0.03783	0.15535	0.15574	0.15602	0.19177	0.19337	0.19596
	SVR	0.00040	0.00041	0.00043	0.00031	0.00032	0.00033	0.04543	0.04548	0.04556
15%	Loess	0.00055	0.00057	0.00059	0.00037	0.00038	0.00040	0.05157	0.05177	0.05194
	Splines	0.00053	0.00055	0.00057	0.00036	0.00036	0.00038	0.05180	0.05199	0.05215
	Wavelets	0.00050	0.00053	0.00054	0.00085	0.00087	0.00088	0.05165	0.05184	0.05199
	Kernel	0.03779	0.03781	0.03784	0.16922	0.16965	0.17000	0.18852	0.18989	0.19142
	SVR	0.00048	0.00050	0.00052	0.00033	0.00034	0.00035	0.05057	0.05069	0.05078
20%	Loess	0.00064	0.00066	0.00068	0.00042	0.00043	0.00044	0.04780	0.04797	0.04819
	Splines	0.00061	0.00063	0.00066	0.00040	0.00042	0.00043	0.04799	0.04818	0.04837
	Wavelets	0.00059	0.00061	0.00064	0.00138	0.00140	0.00142	0.04786	0.04807	0.04825
	Kernel	0.03778	0.03781	0.03785	0.14796	0.14841	0.14864	0.18435	0.18606	0.18721
	SVR	0.00056	0.00058	0.00060	0.00035	0.00036	0.00037	0.04630	0.04638	0.04647
40%	Loess	0.00098	0.00102	0.00104	0.00057	0.00061	0.00065	0.07937	0.07985	0.08031
	Splines	0.00096	0.00099	0.00103	0.00060	0.00064	0.00069	0.07965	0.08014	0.08059
	Wavelets	0.00095	0.00098	0.00101	0.00178	0.00182	0.00187	0.07954	0.08003	0.08047
	Kernel	0.03771	0.03780	0.03786	0.14208	0.14235	0.14271	0.20321	0.20363	0.20426
	SVR	0.00088	0.00091	0.00094	0.00047	0.00048	0.00049	0.06863	0.06901	0.06943

DEG: Differentially expressed genes

In Figure 2, we illustrate the performance of the five normalization methods applied to actual microarray data, without the insertion of artificial outliers. A small difference could be observed in the normalization curves in which the genes displayed low and high expression, due to the low quantity of genes and the high variance.

Discussion

By analyzing the extent to which the outliers could disturb the regression curve, we observed that KR is more highly sensitive to outliers than LR, SS and WS in all three cases (outliers in low, medium and high expression). In all three cases, SVR is shown to be the least affected.

The superior performance of Splines, when compared to KR, may be explained by the degree of smoothing, which varies according to the density of points, differently from KR, which has a fixed window size. Wavelet also has a slightly better performance than KR, probably due to multi-resolution properties. In general, SS and WS presented similar performance when we compared the median of the mean square error using the Wilcoxon test. However, when we used the Kolmogorov-Smirnov test, they presented a statistically significant difference (p value < 0.001). SS and WS constitute somewhat better normalization techniques than LR when we analyzed the sinusoid shape, but, for the other two shapes, LR is better than SS and WS. For practical purposes, the differences between them in terms of disturbance by outliers are too small to be of any concern.

The SVR method is shown to be very robust to outliers presented at different gene expression levels, becoming the best normalization technique to identify actual differentially expressed genes.

One well-known problem in identifying differentially expressed genes is normalizing genes displaying low

Table 3: The mean square errors of estimated gene expression levels for simulated cDNA microarray data with differentially expressed genes inserted under the medium expression levels condition.

Percentage of DEG		Sinusoid			Banana			Mixed		
	Method	25% Quantile	Median	75% Quantile	25% Quantile	Median	75% Quantile	25% Quantile	Median	75% Quantile
5%	Loess	0.00356	0.00373	0.00389	0.00379	0.00392	0.00407	0.05214	0.05234	0.05259
	Splines	0.00354	0.00370	0.00387	0.00379	0.00393	0.00407	0.05237	0.05258	0.05283
	Wavelets	0.00351	0.00368	0.00384	0.00438	0.00450	0.00466	0.05227	0.05247	0.05272
	Kernel	0.03799	0.03816	0.03838	0.17380	0.17441	0.17487	0.18858	0.18882	0.18904
	SVR	0.00337	0.00353	0.00369	0.00357	0.00368	0.00382	0.05034	0.05049	0.05067
10%	Loess	0.00709	0.00723	0.00743	0.00758	0.00780	0.00799	0.05483	0.05506	0.05532
	Splines	0.00707	0.00721	0.00741	0.00763	0.00787	0.00805	0.05507	0.05532	0.05556
	Wavelets	0.00705	0.00718	0.00739	0.00858	0.00881	0.00898	0.05497	0.05522	0.05547
	Kernel	0.03837	0.03857	0.03886	0.15366	0.15461	0.15523	0.18801	0.18830	0.18867
	SVR	0.00672	0.00688	0.00707	0.00709	0.00731	0.00750	0.05150	0.05164	0.05183
15%	Loess	0.01041	0.01061	0.01094	0.01108	0.01136	0.01165	0.05964	0.05985	0.06012
	Splines	0.01039	0.01060	0.01091	0.01109	0.01136	0.01165	0.05990	0.06013	0.06039
	Wavelets	0.01038	0.01058	0.01089	0.01251	0.01276	0.01310	0.05978	0.06003	0.06029
	Kernel	0.03867	0.03897	0.03927	0.12923	0.13026	0.13111	0.19337	0.19367	0.19414
	SVR	0.00986	0.01006	0.01032	0.01027	0.01056	0.01081	0.05499	0.05526	0.05550
20%	Loess	0.01393	0.01418	0.01444	0.01487	0.01519	0.01542	0.06362	0.06398	0.06432
	Splines	0.01393	0.01415	0.01442	0.01486	0.01518	0.01542	0.06390	0.06425	0.06460
	Wavelets	0.01390	0.01414	0.01440	0.01631	0.01666	0.01689	0.06375	0.06410	0.06445
	Kernel	0.03915	0.03957	0.04004	0.12265	0.12366	0.12464	0.19808	0.19858	0.19909
	SVR	0.01310	0.01334	0.01365	0.01365	0.01399	0.01416	0.05809	0.05835	0.05858
40%	Loess	0.02772	0.02813	0.02862	0.02969	0.03004	0.03043	0.07856	0.07910	0.07975
	Splines	0.02774	0.02814	0.02861	0.02966	0.03002	0.03038	0.07884	0.07937	0.08002
	Wavelets	0.02771	0.02811	0.02859	0.03012	0.03049	0.03092	0.07873	0.07926	0.07995
	Kernel	0.04195	0.04261	0.04316	0.15640	0.15816	0.16012	0.20368	0.20443	0.20518
	SVR	0.02545	0.02581	0.02614	0.02656	0.02685	0.02724	0.06786	0.06830	0.06862

DEG: Differentially expressed genes

expression levels, due to the low quantity of the corresponding transcripts and the high spot intensity variance. An equivalent problem occurs with genes presenting very high expression levels due to the low frequency of these genes. Once more, under these conditions, the SVR method is shown to be better than other currently used methods.

We performed the same tests for five other pairs of CodeLink™ microarrays and the results obtained were the same: the SVR is the most robust to outliers and the KR method is the worst method, being highly sensitive to differentially expressed genes and yielding poor regression curves.

Other methods, which are also robust to outliers and are based on a new regression method called two-way semi-linear model [24-27] have also been applied for microarray data normalization. This new approach developed in the last few years, deserves further studies, which we are planning to undertake in the future.

Conclusion

We have proposed a new approach to normalize microarray data and tested this SVR method by benchmark studies and by several simulations. The results obtained with SVR were superior than those obtained with some widely used normalization techniques such as LR, SS and WS. SVR is shown to be more robust to outliers even at very low and very high gene expression levels, being useful to identify differentially expressed genes. Even tested in different microarray shapes, SVR was superior to the other methods, while LR, SS and WS presented similar performances. Therefore, we have demonstrated that SVR is feasible and very promising for microarray data normalization.

Table 4: The mean square errors of estimated gene expression levels for simulated cDNA microarray data with differentially expressed genes inserted under the high expression levels conditions.

Percentage of DEG		Sinusoid			Banana			Mixed		
	Method	25% Quantile	Median	75% Quantile	25% Quantile	Median	75% Quantile	25% Quantile	Median	75% Quantile
5%	Loess	0.00038	0.00039	0.00040	0.00081	0.00087	0.00094	0.04633	0.04639	0.04648
	Splines	0.00035	0.00036	0.00037	0.00079	0.00086	0.00092	0.04658	0.04665	0.04674
	Wavelets	0.00033	0.00034	0.00035	0.00115	0.00121	0.00129	0.04643	0.04650	0.04660
	Kernel	0.03781	0.03782	0.03783	0.16417	0.16453	0.16513	0.17428	0.17436	0.17447
	SVR	0.00031	0.00032	0.00033	0.00080	0.00087	0.00092	0.04527	0.04534	0.04544
10%	Loess	0.00146	0.00153	0.00168	0.00145	0.00156	0.00167	0.04733	0.04748	0.04758
	Splines	0.00142	0.00149	0.00160	0.00147	0.00157	0.00168	0.04756	0.04767	0.04779
	Wavelets	0.00140	0.00147	0.00159	0.00160	0.00170	0.00182	0.04725	0.04736	0.04748
	Kernel	0.02662	0.03454	0.03789	0.23086	0.23135	0.23245	0.18441	0.19003	0.19176
	SVR	0.00126	0.00133	0.00144	0.00142	0.00154	0.00166	0.04678	0.04687	0.04696
15%	Loess	0.00203	0.00217	0.00234	0.00199	0.00212	0.00224	0.04963	0.04976	0.04991
	Splines	0.00198	0.00211	0.00223	0.00200	0.00211	0.00222	0.04987	0.05001	0.05014
	Wavelets	0.00196	0.00209	0.00224	0.00219	0.00240	0.00257	0.04975	0.04989	0.05001
	Kernel	0.02318	0.02992	0.03729	0.17578	0.20066	0.22763	0.18472	0.18898	0.18923
	SVR	0.00178	0.00190	0.00200	0.00170	0.00180	0.00189	0.04885	0.04898	0.04912
20%	Loess	0.00260	0.00275	0.00293	0.00259	0.00272	0.00289	0.04917	0.04930	0.04944
	Splines	0.00254	0.00268	0.00286	0.00260	0.00270	0.00289	0.04933	0.04946	0.04961
	Wavelets	0.00253	0.00267	0.00284	0.00289	0.00304	0.00320	0.04919	0.04933	0.04947
	Kernel	0.02224	0.02819	0.03468	0.16500	0.17817	0.20385	0.18207	0.18716	0.19141
	SVR	0.00226	0.00239	0.00255	0.00247	0.00258	0.00272	0.04839	0.04850	0.04863
40%	Loess	0.00501	0.00520	0.00545	0.00518	0.00538	0.00555	0.04980	0.04999	0.05020
	Splines	0.00498	0.00519	0.00539	0.00520	0.00538	0.00558	0.05002	0.05022	0.05038
	Wavelets	0.00496	0.00517	0.00537	0.00535	0.00551	0.00572	0.04984	0.05004	0.05020
	Kernel	0.02155	0.02487	0.02810	0.18250	0.20140	0.22296	0.17524	0.18072	0.18433
	SVR	0.00446	0.00467	0.00489	0.00464	0.00483	0.00505	0.04809	0.04829	0.04860

DEG: Differentially expressed genes

Methods

Simulation

The program which generates the artificial microarray and the analyses were implemented in R, a language for statistical computing [28]. This script may be downloaded at: [29].

CodeLink™ microarray

Cell lysis and RNA extraction

Cell cultures were lysed with guanidine isocyanate and RNA was purified by of the cell lysates on a cesium chloride cushion (Chirgwin el al, 1979). Absorbance ratio at 260/280 nm was used to assess the RNA purity, a ratio of 1.8 – 2.0 indicating adequate purity.

Labeling and purification of targets

RNA samples were prepared and processed according to protocols supplied by the manufacturer (Amersham Biosciences). Briefly, cDNAs were synthesized from purified RNA (2 μg) and control bacterial mRNAs. Samples were purified using the QIAquick Spin kit (Qiagen) and concentrated by SpeedVac. Concentrated pellets were used in a biotinylated-UTP based cRNA synthesis using the CodeLink™ Expression Assay Reagent Kit (Amersham). Labeled cRNAs were purified using the RNeasy kit (Qiagen) and fragmented with supplied solution at 94°C for 20 min.

Hybridization and washing of arrays

Fragmented biotin-labeled cRNAs (10 μg) were incubated with CodeLink™ bioarrays and shaken (300 rpm) for 20 h. The bioarrays were then washed and incubated with Cy5-Streptavidin (30 min). Scanning of the bioarrays was performed in a GenePix 4000 B Array Scanner (Axon Instruments) and the data were collected using the CodeLink™ System Software (Amersham), which provided the raw data and invalidated data from irregular spots.

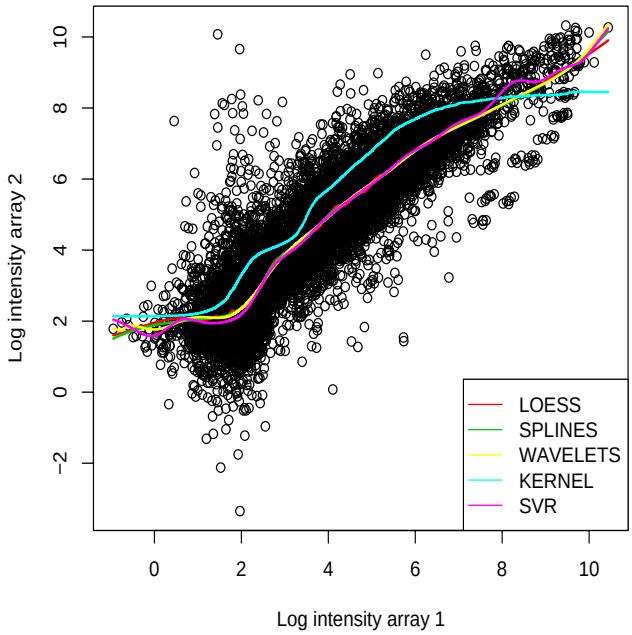

Figure 2
Fitted normalization curves for actual cDNA microarray data using the five different normalization methods (Loess, Splines, Wavelets, Kernel, SVR).

Loess regression

Consider we have n measurements, for each of which the response expected is y_i and let x_i be the predictor, where x is the log intensity of one microarray and y is the log intensity of the other one, in case we are analyzing a single-color microarrays. Whether the microarray is a two-color platform, x is the log of one dye intensity and y is the log of the other dye intensity.

In this model, they are supposed to be related by

$$y_i = g(x_i) + \varepsilon_i \quad (1)$$

where g is the regression function and ε_i is a random error. The idea of local regression is that near $x = x_0$, the regression function $g(x)$ can be locally approximated by the value of a function in some specified parametric class. Such a local approximation is obtained by fitting a regression surface to the data points within a chosen neighborhood of the point x_0.

In this method, weighed least squares are used to fit linear or quadratic functions of the predictors at the centers of the neighborhoods. The radius of each neighborhood is chosen so that the neighborhood contains a specified percentage of the data points. The fraction of the data, called the smoothing parameter, in each local neighborhood is weighted by a smooth decreasing function of their distance from the center of the neighborhood [30].

B-Splines smoothing

Due to its simple structure and good approximation properties, polynomials are widely used in practice for approximating functions [31,32]. Let x and y as defined above and

$$(x-\gamma)_+^0 = \begin{cases} 1, x \geq \gamma \\ 0, x < \gamma \end{cases} \quad (2)$$

and

$$(x-\gamma)_+^{m-1} = \begin{cases} (x-\gamma)^{m-1}, x \geq \gamma, m > 1 \\ 0, x < \gamma \end{cases} \quad (3)$$

Therefore, let

$$\ldots \leq \gamma_{-1} \leq \gamma_0 \leq \gamma_1 \leq \gamma_2 \leq \ldots \quad (4)$$

be a sequence of real numbers. Given integers i and $m > 0$, we define

$$Q_i^m(x) = \begin{cases} (-1)^m [\gamma_i, \ldots, \gamma_{i+m}](x-\gamma)_+^{m-1}, if\ \gamma_i < \gamma_{i+m} \\ 0, otherwise \end{cases} \quad (5)$$

for all real x. We call Q_i^m the m th order B-Spline associated with the knots $\gamma_i, \ldots, \gamma_{i+m}$.

For $m = 1$, the B-Spline associated with $\gamma_i < \gamma_{i+1}$ is particularly simple. It is the piecewise constant function

$$Q_i^1(x) = \begin{cases} \dfrac{1}{\gamma_{i+1} - \gamma_i}, \gamma_i \leq x \leq \gamma_{i+1} \\ 0, otherwise. \end{cases} \quad (6)$$

In our analysis, we applied the cubic Splines, i.e., Splines of order 3.

We can also give explicit formulate for Q_i^m in case either γ_i or γ_{i+m} is a knot of multiplicity m.

Wavelet smoothing

The Wavelet transform is a relatively new approach and has some similarities with the Fourier transform. Wavelets differ from Fourier methods in that they allow the localization of a signal in both time and frequency. In the wavelet theory, a function is represented by an infinite series expansion in terms of dilated and translated version of a basic function ψ called the "mother' Wavelet. A Wavelet transformation leads to an additive decomposition of a

signal into a series of different components describing smooth and rough features of the signal.

The term Wavelets means small curves, therefore, they are oscillations that rapidly decay. As the B-Splines functions system, the Wavelets functions $\psi(t)$ can be used to generate a function basis for certain spaces [33]. An ortonormal basis can be generated by dyadic dilations and translations of a mother Wavelet $\psi(t)$, by

$$\psi_{j,k}(t) = 2^{j/2}\,\psi(2^j j - k), \quad j, k \in Z \qquad (7)$$

Wavelets are functions which satisfy the following properties:

i) $\int_{-\infty}^{\infty} \psi(t)\,dt = 0$ (8).

ii) $\int_{-\infty}^{\infty} |\psi(t)|\,dt < \infty$ (9).

iii) $\int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2\,d\omega}{|\omega|} = 0$, where the function $\Psi(\omega)$ is the Fourier transform of $\psi(t)$ (10).

iv) $\int_{-\infty}^{\infty} t^j \psi(t)dt = 0$, $j = 0,1,\ldots, r - 1$ for $r \geq 1$ and $\int_{-\infty}^{\infty} \left|t^r \psi(t)\right| dt < \infty$ (11).

An important result is that any function $f(t)$ with $\int_{-\infty}^{\infty} f^2(t) < \infty$ can be expanded as

$$f(t) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} c_{j,k}\psi_{j,k}(t) \qquad (12).$$

In other words, any function $f(t)$ can be represented by a linear combination of functions $\psi_{j,k}(t)$. The smoothing procedure can be carried out by an approximation, choosing a maximum resolution $J(t)$ for $j = 1,2,\ldots, J(t)$ and $k = 1,2,\ldots, 2^{j-1}$. Here, we considered the Mexican hat Wavelet [34] defined by

$$\psi(t) = (1 - t^2)\exp\left(\frac{-t^2}{2}\right) \qquad (13)$$

rather than other functions such as Morlet or Shannon since they do not have an analytic formula.

The C_{jk} coefficients are estimated via an ordinary least square regression. An important feature in the wavelets representation is that it allows the description of functions belonging to both Sobolev and Besov spaces [35].

Kernel regression

KR is one class of modeling methods that belongs to the smoothing methods family. It is part of the non-parametric regression methods. KR allows basing the prediction of a value on passed observations, and weighing the impact of past observations depending on how similar they are, compared to the current values of the explanatory variables.

The KR is one of the most widely used procedures in non-parametric curve estimation. Nadaraya (1964) and Watson (1964) proposed an estimator for the curve g given by

$$g_h(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i)Y_i}{\sum_{j=1}^{n} K_h(x - X_j)} \qquad (14)$$

In our datasets, we used the Gaussian Kernel because it is symmetric and centralized in the mean.

In addition to being easy to compute, the Nadaraya-Watson estimator $g_h(x)$ is consistent. When $h \to 0$ the estimated curve presents a large variability and when $nh \to \infty$, we obtain an overly smooth curve [36]. The bandwidth h controls the smoothness degree of the estimated curve. It is easy to observe that this KR estimator is just a weighted sum of the observed responses Y_i. The denominator ensures that the weights sum up to 1.

Support Vector Regression

SVR generalized algorithm is a non-linear regression from the Generalized Portrait algorithm developed in Russia by Vapnik and Lerner (1963) [37] and Vapnik and Chervonenkis (1964) [38]. It is based upon the statistical learning theory which has been developed by Vapnik and Chervonenkis (1974) [39]. In Bioinformatics, and, more specifically, in microarray data analysis, to the best of our knowledge, this algorithm has previously been used only once, by Hisanori et al. (2004), to extract relations between promoter sequences and strengths [40]. Here, we propose the use of SVR to normalize microarray data.

Let $\{(x_1, y_1),\ldots, (x_1, y_1)\} \subset R \times R$ be the gene expression data derived from microarray experiments, where x is the

log intensity of one microarray and y is the log intensity of the other one, in case we are analyzing a single-color microarrays. When the microarray is a two-color platform, x is the log of one dye intensity and y is the log of the other dye intensity. In ε-SVR [41], the goal is to obtain a function $f(x)$ that has at the most ε deviation from the y_i for all the data, and is as flat as possible.

In the case of linear functions f:

$$f(x) = (w^t x) + b \text{ with } w \in R^n, b \in R \quad (15)$$

Flatness in (15) means

$$\text{Minimize } \frac{1}{2}\|w\|^2$$

$$\text{Constrained to } \begin{cases} y_i - \left(w^t x_i\right) - b \le \varepsilon \\ \left(w^t x_i\right) + b - y_i \le \varepsilon \end{cases} \quad (16)$$

In (16) there is a function f which, with ε precision, approximates all pairs (x_i, y_i). But there are cases where it is necessary to allow for some errors. To solve this problem, one can introduce slack variables ξ_i, ξ_i^* to deal with unfeasible constraints of the optimization problem (16) arriving at the formulation stated in [41]

$$\text{Minimize } \frac{1}{2}\|w\|^2 + C\sum(\xi_i + \xi_i^*)$$

$$\text{Constrained to } \begin{cases} y_i - \left(w^t x_i\right) - b \le \varepsilon + \xi_i \\ \left(w^t x_i\right) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \quad (17)$$

where the constant $C > 0$ is the trade-off between the amount up to which deviations larger than ε are tolerated, maintaining the flatness of f. This corresponds to dealing with the ε-insensitive loss function $|\xi|_\varepsilon$:

$$|\xi|_\varepsilon := \begin{cases} 0 \text{ if } |\xi| \le \varepsilon \\ |\xi| - \varepsilon \text{ otherwise} \end{cases} \quad (18)$$

It is necessary to construct a Lagrange function from the primal objective function and the corresponding constraints by introducing a dual set of variables. According to Mangasarian (1969) [42], McCormick (1983) [43], and Vanderbei (1997) [44] it follows that:

$$L := \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right) - \sum_{i=1}^{l}\left(\eta_i\xi_i + \eta_i^*\xi_i^*\right) - \sum_{i=1}^{l}\alpha_i\left(\varepsilon + \xi_i - y_i + \left(w^t x_i\right) + b\right) - \sum\alpha_i^*\left(\varepsilon + \xi_i^* + y_i - \left(w^t x_i\right) - b\right) \quad (19)$$

where L is the Lagrangian and η_i, η_i^*, α_i, α_i^* are Lagrange multipliers. Hence the dual variables in (19) have to satisfy

$$\alpha_i^{(*)}, \eta_i^{(*)} \ge 0 \quad (20)$$

Note that we refer to α_i and α_i^* as $\alpha_i^{(*)}$.

From the saddle point condition, the partial derivatives of L related to (w, b, ξ_i, ξ_i^*) have to vanish for optimality.

$$\partial_b L = \sum_{i=1}^{l}\left(\alpha_i^* - \alpha_i\right) = 0 \quad (21)$$

$$\partial_w L = w - \sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right)x_i = 0 \quad (22)$$

$$\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0 \quad (23)$$

From the substitution of (21), (22) and (23) into (19) we obtain a dual optimization problem.

$$\text{Maximize } \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{l}\left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right)\left(x_i^t x_j\right) \\ -\varepsilon\sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right) + \sum_{i=1}^{l} y_i\left(\alpha_i - \alpha_i^*\right) \end{cases} \quad (24)$$

$$\text{Subject to } \sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right) = 0 \text{ and } \alpha_i, \alpha_i^* \in [0, C]$$

Equation (22) can be rewritten as follows

$$w = \sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right)x_i, \quad \text{thus} \quad f(x) = \sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right)\left(x^t x\right) + b \quad (25)$$

This is the Support Vector expansion, i.e., the description of w as a linear combination of x_i.

To compute b, it is necessary to use Karush-Kuhn-Tucker (KKT) conditions [45,46]. These authors state that at the

point of the solution the product between dual variables and constraints has to vanish.

$$\alpha_i (\varepsilon + \xi_i - y_i + (w^t x_i) + b) = 0 \quad (26)$$

$$\alpha_i^* (\varepsilon + \xi_i^* + y_i - (w^t x_i) - b) = 0$$

and

$$(C - \alpha_i)\xi_i = 0 \quad (27)$$

$$(C - \alpha_i^*)\xi_i^* = 0$$

From (26) and (27) it follows that:

(i) Only samples (x_i, y_i) with corresponding $\partial_b L = \sum_{i=1}^{l} \left(\alpha_i^* - \alpha_i \right) = 0$ (21) = C lie outside the ε-insensitive tube;

(ii) $\alpha_i \alpha_i^* = 0$

From (i) and (ii), it is possible to conclude that

$$\varepsilon - y_i + (w^t x_i) + b \geq 0 \text{ and } \xi_i = 0 \text{ if } \alpha_i < C \quad (28)$$

$$\varepsilon - y_i + (w^t x_i) + b \leq 0 \quad \text{if } \alpha_i > 0 \quad (29)$$

In conjunction with an analogous analysis on α_i^*

$$\max\{-\varepsilon + y_i - (w^t x_i) | \alpha_i < C \text{ or } \alpha_i^* > 0\} \leq b \leq \min\{-\varepsilon + y_i - (w^t x_i) | \alpha_i > 0 \text{ or } \alpha_i^* < C\} \quad (30)$$

If some $\alpha_i^* \in (0, C)$ the inequalities become equalities.

To point out the sparsity of the SV expansion: from (26), the Lagrange multipliers may be nonzero only for $|f(x_i) - y_i| \geq \varepsilon$.

Therefore, we have a sparse expansion of w in terms of x_i [47].

Abbreviations

LR: Loess Regression

SS: Splines Smoothing

WS: Wavelets Smoothing

KR: Kernel Regression

SVR: Support Vector Regression

Authors' contributions

AF – has made substantial contributions to conception and design of the study, analysis and interpretation of data and has been involved in drafting of the manuscript.

JRS – has made substantial contributions to conception and design of the study, analysis and interpretation of data and has been involved in drafting of the manuscript.

LOR – acquisition of the benchmark data and has been involved in drafting parts of the manuscript.

CEF – has discussed the results and critically revised the manuscript for important intellectual content and has given the final approval of the version to be published.

MCS – has directed the work on differentially expressed genes using the CodeLink™ platform and critically revised the manuscript for important intellectual content and has given the final approval of the version to be published.

Acknowledgements

This research was supported by FAPESP, CAPES, CNPq, FINEP and PRP-USP.

References

1. Quackenbush J: **Microarray data normalization and transformation.** *Nat Genet* 2002, **32:**496 -501.
2. Cullane AC, Perriere G, Considine EC, Cotter TG, Higgins DG: **Between-group analysis of microarray data.** *Bioinformatics* 2002, **18:**1600-1608.
3. Durbin BP, Hardin JS, Hawkins DM, Rocke DM: **A variance-stabilizing transformation for gene-expression microarray data.** *Bioinformatics* 2002, **18:**S105-110.
4. Kepler TB, Crosby L, Morgan KT: **Normalization and analysis of DNA microarray data by self-consistency and local regression.** *Genome biol* 2002, **3:**RESEARCH0037.
5. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, Quackenbush J: **Within the fold: assessing differential expression measures and reproducibility in microarray assays.** *Genome Biol* 2002, **3:**research0062.
6. Schadt EE, Li C, Ellis B, Wong WH: **Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data.** *J Cell Biochem* 2001, **37:**120-125.
7. Hill AA, Brown EL, Whitley MZ, Tucker-Kellogg G, Hunter CP, Slonim DK: **Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls.** *Genome Biol* 2001, **2:**RESEARCH0055.
8. Yang YH, Speed T: **Design issues for cDNA microarray experiments.** *Nat Rev Genet* 2002, **3:**579-588.
9. Perou CM: **Show me the data!** *Nat Genet* 2001, **29:**373.
10. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M: **Minimum information about microarray experiment (MIAME)-toward standards for microarray data.** *Nat Genet* 2001, **29:**365-371.
11. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite**

method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 2002, **30:**e15.
12. Beheshti B, Braude I, Marrano P, Thorner P, Zielenska M, Squire JA: **Chromosomal localization of DNA amplifications in neuroblastoma tumors using cDNA microarray comparative genomic hybridization.** *Neoplasia* 2003, **5:**53-62.
13. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J: **TM4: a free, open-source system for microarray data management and analysis.** *Biotechniques* 2003, **34:**374-378.
14. Baird D, Johnstone P, Wilson T: **Normalization of microarray data using a spatial mixed model analysis which includes splines.** *Bioinformatics* 2004, **17:**3196-205.
15. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielsen HB, Saxild HH, Nielsen C, Brunak S, Knudsen S: **A new non-linear normalization method for reducing variability in DNA microarray experiments.** *Genome Biology* 2002, **3(9):**research0048.1-0048.16.
16. Wang J, Ma JZ, Li MD: **Normalization of cDNA microarray data using wavelet regressions.** *Combinatorial Chemistry & High Throughput Screening* **9:**783-791.
17. Nadaraya EA: **On estimating regression.** *Theory of probability and its applications* 1964, **10:**186-190.
18. Watson GS: **Smooth regression analysis.** *Sankya A* 1964, **26:**359-372.
19. Vapnik VN: **The Nature of Statistical Learning Theory.** *Springer* 1995.
20. Ramakrishnan R, Dorris D, Lublinsky A, Nguyen A, Domanus M, Prokhorova A, Gieser L, Touma E, Lockner R, Tata M, Zhu X, Patterson M, Shippy R, Sendera TJ, Mazumder A: **An assessment of Motorola CodeLink™ microarray performance for gene expression profiling applications.** *Nucleic Acids Research* 2002, **30:**.
21. Balagurunathan Y, Dougherty ER, Chen Y, Bittner ML, Trent JM: **Simulation of cDNA microarrays via a parameterized random signal model.** *Journal of Biomedical Optics* 2002, **7(3):**507-523.
22. Dias R: **A review of non-parametric curve estimation methods with application to Econometrics.** *Economia* 2002, **2:**31-75.
23. Archambeau C: **Probabilistic models in noisy environment – and their application to a visual prosthesis for the blind.** In *PhD thesis* Universite catholique de Louvain, Applied Sciences Faculty; 2005.
24. Fan J, Tam P, Vande WG, Ren Y: **Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine.** *PNAS* 2004, **101:**1135-1140.
25. Fan J, Peng H, Huang T: **Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency.** *J Am Stat Assoc* 2005, **100(471):**781-813.
26. Huang J, Wang D, Zhang C: **A two-way semi-linear model for normalization and analysis of cDNA microarray data.** *J Am Stat Assoc* 2005, **100(471):**814-829.
27. Wang D, Huang J, Xie H, Manzella L, Soares MB: **A robust two-way semi-linear modelo for normalization of cDNA microarray data.** *BMC Bioinformatics* 2005, **6(14):**.
28. **The R project for statistical computing** [http://www.r-project.org]
29. **Evaluating different methods of microarray data normalization** [http://mariwork.iq.usp.br/normalization/]
30. Cleveland WS, Grosse E, Shyu WM: **Local regression models.** *Chapter 8 Statistical Models in S. Wadsworth & Brooks/Cole* 1992.
31. Schumaker LL: *Spline functions basic theory* New York: John Wiley & Sons; 1981.
32. Prenter PM: *Splines and variational methods* New York: John Wiley & Sons; 1975.
33. Meyer Y: *Wavelets Algorithms and Applications* Philadelphia: SIAM; 1993.
34. Chui CK: *An introduction to wavelets* San Diego: Academic Press; 1992.
35. Härdle W: *Smoothing techniques with implementation* New York: Springer-Verlag; 1990.
36. Donoho DL, Johnstone IM: **Minimax estimation via wavelet shrinkage.** *Annals of Statistics* 1998, **26:**879-921.
37. Vapnik V, Lerner A: **Pattern recognition using generalized portrait method.** *Automatic and Remote Control* 1963, **24:**774-780.
38. Vapnik V, Chervonenkis A: **A note on one class of perceptrons.** *Automatics and Remote Control* 1964:25.
39. Vapnik V, Chervonenkis A: *Theory of pattern recognition* Moskow: Nauka; 1974.
40. Hisanori K, Oshima T, Asai K: **Extracting relations between promoter sequences and their strengths from microarray data.** *Bioinformatics* 2004, **21:**1062-1068.
41. Vapnik VN: *Statistical Learning Theory* New York: Wiley; 1998.
42. Mangasarian OL: *Nonlinear Programming* New York: McGraw-Hill; 1969.
43. McCormick GP: *Nonlinear Programming Theory Algorithms and Applications* New York: John Wiley and Sons; 1983.
44. Vanderbei RJ: **An interior point code for quadratic programming.** In *Statistics and Operations Research* Princeton Univ., NJ; 1997.
45. Karush W: **Minima of functions of several variables with inequalities as side constraints.** In *Master thesis* University of Chicago Department of Mathematics; 1939.
46. Kuhn HW, Tucher AW: *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probabilistics* Berkeley University of California Press; 1951:481-492.
47. Smola AJ, Schölkopf B: **A tutorial on support vector regression.** *Statistics and Computing* 2004, **14:**199-222.