Nucleic Acids Research, 2005, Vol. 33, No. 12 ***3699–3707***
doi:10.1093/nar/gki660

A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes

Aristotelis Tsirigos[1,2] and Isidore Rigoutsos[2,3,*]

[1]New York University, Computer Science, New York, NY 10021, USA, [2]Bioinformatics and Pattern Discovery Group, Computational Biology Center, IBM Thomas J. Watson Research Center, PO Box 218, Yorktown Heights, NY 10598, USA and [3]Department of Chemical Engineering, Massachusetts Institute of Technology, Room 56-469, Cambridge, MA 02139, USA

Received March 30, 2005; Revised and Accepted May 27, 2005

ABSTRACT

In earlier work, we introduced and discussed a generalized computational framework for identifying horizontal transfers. This framework relied on a gene's nucleotide composition, obviated the need for knowledge of codon boundaries and database searches, and was shown to perform very well across a wide range of archaeal and bacterial genomes when compared with previously published approaches, such as Codon Adaptation Index and C + G content. Nonetheless, two considerations remained outstanding: we wanted to further increase the sensitivity of detecting horizontal transfers and also to be able to apply the method to increasingly smaller genomes. In the discussion that follows, we present such a method, Wn-SVM, and show that it exhibits a very significant improvement in sensitivity compared with earlier approaches. Wn-SVM uses a one-class support-vector machine and can learn using rather small training sets. This property makes Wn-SVM particularly suitable for studying small-size genomes, similar to those of viruses, as well as the typically larger archaeal and bacterial genomes. We show experimentally that the new method results in a superior performance across a wide range of organisms and that it improves even upon our own earlier method by an average of 10% across all examined genomes. As a small-genome case study, we analyze the genome of the human cytomegalovirus and demonstrate that Wn-SVM correctly identifies regions that are known to be conserved and prototypical of all beta-herpesvirinae, regions that are known to have been acquired horizontally from the human host and, finally, regions that had not up to now been suspected to be horizontally transferred. Atypical region predictions for many eukaryotic viruses, including the α-, β- and γ-herpesvirinae, and 123 archaeal and bacterial genomes, have been made available online at http://cbcsrv.watson.ibm.com/HGT_SVM/.

INTRODUCTION

For several decades, scientists have been documenting the ability of microbes to incorporate foreign DNA into their genome (1). Even though in the early years, claims of extensive horizontal transfers were met with skepticism, the data gathered by the various genome sequencing projects have provided overwhelming evidence of how widespread this phenomenon is. For an extensive discussion of various aspects of horizontal gene transfer (HGT), the reader is referred to (2), which provides an excellent introduction to the topic.

In our earlier work on the subject of horizontal gene transfer (3), we reviewed extensively the numerous computational methods, which have been devised over the years for identifying such events. Consequently, in what follows we will only summarily describe the main categories of methods, namely, phylogenetic and composition-based methods, and outline their key characteristics.

Phylogenetic methods depend on the knowledge of orthologous sequences and are very robust if sufficient amounts of data are available. They are particularly effective in identifying transfers and this has already been demonstrated by many researchers (4). On the other hand, compositional methods rely on the premise that a given organism exhibit compositional features which remain relatively constant across its genomic

*To whom correspondence should be addressed. Tel: +1 914 945 1384; Fax: +1 914 945 4217; Email: rigoutso@us.ibm.com sequence and can thus be used to generate a 'description' of the organism. As one might expect, a great variety of approaches have been proposed in this category: the methods ranged from the use of the G + C content to dinucleotide signatures, to codon usage patterns and the Codon Adaptation Index (CAI) (5–11). Compositional methods based on higher-order oligonucleotides have also been proposed (12). In addition to these main two categories, we should also mention the existence of (i) surrogate methods (13), which attempt to answer the question of horizontal gene transfer without any need for orthologous sequences; and (ii) hybrid methods that combine one or more compositional methods into a single scheme (14). Finally, and independent of the used method, highly expressed genes and ribosomal proteins are typically filtered out (14,15).

Examination of all of the previously published methods readily reveals that they have focused entirely on the analysis of archaeal and bacterial genomes. Moreover, and precisely as one would intuitively expect, the performance of essentially all composition-based computational techniques generally improves if more data are available that can be used to define the average genome 'signature'. A better-defined genomic signature will in turn result in an improved signal-to-noise ratio when determining the provenance of a given gene. Naturally, for a given composition-based method, the uncertainty in deciding whether a gene is atypical with respect to the rest of the genome will increase in inverse proportion to the size of the processed genome: the smaller the genome, the harder it becomes to assess the atypicality of its genes. Analogous observations can be made for phylogenetic approaches: the larger the collection of orthologous genes that are available, the greater the confidence in the conclusions.

In view of the above observations, we set out to create a method that would work equally well with large and small genomes. In particular, we sought a method that would be able to sensitively detect horizontal transfers even when the set of genes used for training was small. Such a method would result in improved sensitivity when analyzing archaeal and bacterial genomes while at the same time permitting the study of the phenomenon of horizontal transfer in viral genomes.

Viral transduction has been known for quite some time (16) as a mechanism by which viruses, during their replication, incorporate in their genome genetic material from their host, which they then transfer to a new host. In fact, there is no reason as to why the final recipient host should even be related to the host in which the genetic material originated. Clearly, transduction holds substantial potential for virus-mediated genetic engineering. However, it remains unclear whether transduction is as an important enabler of evolutionary change, if at all. Because of the latter uncertainty and the lack of methods that could sensitively detect horizontal transfers in small genomes, viruses have remained until now uncharted territory from the standpoint of horizontal transfer analysis.

In what follows, we present a new method that builds on a representation scheme that we introduced recently (3). This new method can learn very effectively from small training sets and works well with archaeal and bacterial as well as viral genomes. Notably, the new method even improves upon the Wn method, which we introduced in (3), by an average of 10% across a very large collection of tested genomes.

MATERIALS AND METHODS

Brief overview of our generalized compositional framework

In (3), we introduced a generalized, composition-based framework for HGT detection. Summarily, our framework extends and generalizes composition-based methods in three distinct ways:

(i) it uses higher-order nucleotide sequences (templates); this leads to improved discrimination power and an improved ability to classify genes when compared with the previously proposed di- and tri-nucleotide models;
(ii) it extends composition-based schemes through the ability to 'ignore' certain nucleotide positions; this was achieved with the use of generating templates that include 'wildcards' and thus comprise non-consecutive nucleotides; and
(iii) it permits the optional consideration of the periodicity of the DNA code; in particular, when collecting the instances of a template, we can optionally align the template with codon boundaries.

Given a genome sequence, our ultimate objective is to characterize coding and non-coding regions of the genome in terms of how 'atypical' they are compared to the 'average' composition. Let $\phi(s) = (\alpha_1, \alpha_2, \ldots, \alpha_q)$ denote the compositional feature vector for any given DNA sequence s over a set of templates $\pi = \{\pi_1, \pi_2, \ldots, \pi_q\}$; α_i is the frequency of template π_i in sequence s. Similarly, we compute the compositional feature vector $\phi(G)$ for the whole genome G as the average of the compositional feature vectors of the given sequences. With the help of a standard similarity measure (e.g. correlation, covariance, χ^2 test, Mahalanobis distance, relative entropy, etc.), we assign a typicality score $S_G(g)$ to each gene g of genome G: the higher the score the more typical the gene is for the genome. Genes with low scores are thus candidates to be the result of horizontal gene transfer events. We called the resulting method Wn, where n is an integer greater than two and equal to the size of the template.

Our extensive analysis in (3) demonstrated that for template sizes greater than two, the optimal performance is obtained when the codon boundaries are ignored (i.e. all the templates are counted, including those that begin at the second and third codon positions), the templates include no wildcards, and covariance is used as the similarity measure for computing typicality scores. Moreover, the performance of the method increased with the size of the template, reaching a maximum for $n = 8$; performance deteriorated with further increases in the size of the template.

A new similarity measure: one-class support-vector machines (SVM)

Given a set of training data points in a high-dimensional input space, the objective of the one-class SVM method (17) is to learn a function that will take the value +1 in the region where the majority of the data points are concentrated, and the value -1 everywhere else. The function to be learned is modeled as a hyperplane in a transformed space (= feature space), and hyperplane parameters are estimated so that its margin with respect to the training data is maximized, as dictated by the data-driven distribution-free paradigm.

More formally, let us consider the training objects $x_1, \ldots, x_1 \in X$ and a feature map ϕ: $X \rightarrow \mathbb{R}^m$, which maps objects from the input space X to points in the feature space $\mathbb{R}^m$, where m is the number of features associated with each object. The maximum margin solution of the one-class SVM problem, i.e. the problem of finding the maximum-margin hyperplane in the feature space that separates the data from the origin, is obtained by solving the following quadratic optimization problem:

$$\min_{u \in \mathbb{R}^m,\, \xi_i \geqslant 0,\, \rho \in \mathbb{R}} \frac{1}{2}\|u\|^2 + \frac{1}{\nu l}\sum_{i=1}^{l} \xi_i - \rho$$

$$\text{s.t.} \quad \begin{array}{l} \langle u \cdot \phi(x) \rangle \geqslant \rho - \xi_i, \\ \xi_i \geqslant 0, 1 \leqslant i \leqslant l, \end{array}$$

where $u \in \mathbb{R}^m$ is a vector describing the hyperplane in the feature space, $\rho \in \mathbb{R}$ is the margin of the hyperplane with respect to the data, ξ_i are non-zero slack variables allowing for a soft margin, and $\nu \in (0,1)$ is a parameter that represents an upper bound on the fraction of outliers in the data. Finally, the decision function inferred by the learned hyperplane is:

$$f(x) = \operatorname{sgn}(\langle u \cdot \phi(x) \rangle - \rho).$$

The optimization problem is solved by applying the Lagrange multipliers, thus converting it to the equivalent dual problem:

$$\min_{\alpha \in \mathbb{R}^l} \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j \langle \phi(x_i) \cdot \phi(x_j) \rangle$$

$$\text{s.t.} \quad \begin{array}{l} \sum_{i=1}^{l} \alpha_i = 1, \\ 0 \leqslant \alpha_i \leqslant \frac{1}{\nu l}, \end{array}$$

with the hyperplane parameters given by $u = \sum_{i=1}^{l} \alpha_i \phi(x_i)$.

In the context of the HGT detection problem, the input space X is the set of all possible nucleotide sequences, whereas the feature space $\mathbb{R}^m$ comprises the selected set of compositional features of the nucleotide sequences, i.e. the frequencies of all templates of size n. We can use the learned decision function to induce a scoring measure S_G of genes belonging to a fixed genome G, where more atypical genes will receive lower scores: $S_G\ (x) = \langle u \cdot \phi(x) \rangle$.

When $\nu = 1$, this last measure, S_G, is proportional to the covariance of the two vectors involved in the inner product. Furthermore, it is worth pointing out that for any feature map ϕ, the typicality measure obtained from the solution of the one-class SVM optimization problem and the covariance measure discussed previously (3) will result in the same relative ranking of genes with respect to typicality. This can be shown with the help of the Lemma contained in Appendix A.

However, for values of $\nu < 1$, the optimal hyperplane solution will have some coefficients α_i assume a value of zero; the genes for which $\alpha_i = 0$ will not contribute their compositional features $\phi(x_i)$ to the computation of u.

From the above, we can give a natural interpretation to the optimal hyperplane u as a generalized genome signature: when $\nu = 1$, the generalized signature is equivalent to the usual genome signature which is computed as the mean of the gene signatures in the genome; for $0 < \nu < 1$, the generalized signature will comprise only a subset of special 'signature genes.' This also constitutes a natural interpretation of the parameter ν for the problem that we try to solve here as an upper bound on the fraction of gene transfers in the genome. In the following section, we use this fact to estimate, via a series of experiments, the optimal parameter ν for any given genome so that the number of recovered horizontal gene transfers is maximized. The worst-case performance of the one-class SVM-based method can be as good as the covariance-based method that we introduced previously (3) (this is again a direct sequence of the Lemma in Appendix A). However, in practice, the Wn-SVM method achieves an average improvement of >10% across the 123 archaea and bacterial genomes that we have used as a reference.

RESULTS

Evaluation of Wn-SVM: archaeal and bacterial genomes

For each of the 123 host organisms in turn, we conducted $k = 20$ experiments of simulated transfers from a gene pool. As in (3), this simulation was carried out using a pool of more than 350 000 archaeal and bacterial genes: in fact, we permitted all our genomes to exchange genes with one another while making sure that a given genome did not become a gene donor for itself. The genes were randomly selected from the pool and 'inserted' in the i-th genome: the task at hand for each of the tested methods was to recover as many as possible of these artificially inserted genes. To the best of our knowledge, it is important to note here that this simulation as well as those mentioned in (3) are unique in that they are carried out using donor sets comprising actual genes. The methods we tested included CAI, Wn and Wn-SVM. Moreover, the set of donors was the same as the set of acceptors, in other words we allowed the tested genomes to exchange genes with one another in any conceivable combination. As such, this is a realistic simulation of what happens naturally (as it is currently understood). In each experiment, the number of added genes was chosen to be a fixed percentage of the number of genes in the host genome. The 'transferred' genes were selected from the donor pool at random and with replacement. The simulated-transfer experiments were carried out for transfer percentages which ranged between 1 and 8% of the genes in the host genome under consideration.

Given each genome and transfer percentage combination, each of the tested methods had to recover as many of the artificially transferred genes as possible, without using any a priori knowledge about the host genome or the donor genes. The ideal method should recover each and every one of the artificially added genes. However, our artificial insertions compete for the top, putative-transfer positions with the horizontal gene transfers that are already present in the genome under consideration. Consequently, not all of the artificially inserted genes will occupy the top, putative-transfer positions: we use the term 'hit ratio' to refer to the fraction of the artificially inserted genes that a tested method manages to recover. We should

Table 1. Gene scoring methods

Name	Width	Step	Measure	Description
CAI	3	3	N/A	Codon Adaptation Index
W8	8	1	covariance	8 nt composition (no wildcards)
W8-SVM	8	1	SVM	8 nt composition (no wildcards)

point out that this situation poses no problem for the purposes of simulation as it holds true for all of the tested methods, and thus no method is favored at the expense of another.

In (3), we showed that the best performance was achieved by Wn for templates of size $n = 8$ and that the second best method was the CAI. These are the three methods that we evaluated. Table 1 summarizes the characteristics of the three methods. Method m's overall performance across the N genomes under consideration is defined as:

$$\mathrm{Perf}^m = \frac{1}{N}\sum_G \left(\frac{1}{k}\sum_{i=1}^{k} r_i^m(G)\right),$$

where $r_i^m(G)$ is the hit ratio obtained by the method m for genome G at the i-th iteration of the experiment (with $1 \leqslant i \leqslant k$).

For the one-class SVM method, we have the additional task of estimating the parameter v, which controls the fraction of genes that contribute to the genome signature. For each genome and each given percentage of added genes, we estimate the optimal value of the parameter v so that the fraction of the artificially inserted genes recovered by the SVM method is maximized. This estimation is carried out by varying the value of v from 0 to 1 using a step of 0.1 and conducting $k = 20$ experiments for each value; performance was averaged over these 20 experiments and the value of v that maximized the performance was chosen as the optimal value for v. The highly optimized SVM package LibSVM by Chang and Lin was used to solve a total of 200 quadratic problems per organism. The code and reference manuals for LibSVM can be found online at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

In Table 2, we list the overall performance of all methods for different percentages of artificially added genes. Table 3 shows the improvement that our new W8-SVM method achieves when compared with the remaining methods: the improvement is shown both in absolute percentage points and in terms of relative values and represents the average across the 20 experiments that we carried out for each genome and artificial transfer percentage value. Table 3 is also depicted graphically in Figure 1. The amount of relative improvement that W8-SVM achieves relative to method m is computed using the following formula:

$$\mathrm{Rel}^m = \frac{1}{N}\sum_G \mathrm{Rel}_G^m = \frac{1}{N}\sum_G \frac{\mathrm{Perf}^{\mathrm{W8/SVM}}(G) - \mathrm{Perf}^m(G)}{\mathrm{Perf}^m(G)},$$

and is a measure of how many more horizontal transfers are detected by W8-SVM. For example, in the experiments with 2% added genes from the pool, the W8-SVM method achieved a relative improvement of 10.6% compared with W8 and 33.6% compared with CAI.

Table 2. The overall performance Perf^m for the methods under evaluation is shown: higher numbers for the overall performance are more preferable—see also text for a definition of Perf^m

% HGT	CAI (%)	W8 (%)	W8-SVM (%)
1	46.3	51.6	56.6
2	51.6	56.2	60.6
4	56.5	60.9	64.1
8	61.5	65.4	67.7

Table 3. Improvement of the new W8-SVM method over CAI and over W8

% HGT	W8-SVM versus CAI (%)	W8-SVM versus W8 (%)
% Improvement in overall performance		
1	10.3	5.0
2	9.0	4.4
4	7.6	3.2
8	6.2	2.3
% Average relative improvement		
1	52.0	15.0
2	33.6	10.6
4	23.5	6.3
8	15.4	3.8

In Figure 2, we show a comparison between W8-SVM and W8 for each of the 123 genomes and for those experiments where we added 2% donor genes. As predicted theoretically, W8-SVM improves upon W8 across all the genomes with which we experimented (but of course is in no case inferior to W8).

In Figure 3, we compare W8-SVM with the CAI method: green solid bars indicate the cases where W8-SVM outperforms CAI, whereas red bars are used when CAI outperforms W8-SVM. The height of each bar corresponds to the relative improvement Rel_G^m achieved by our method over CAI as an average over the 20 experiments and can be either positive (green bars) or negative (red bars).

Evaluation of Wn-SVM: analysis of the human cytomegalovirus genome

In addition to the simulation and analysis of archaea and bacteria, we present an analysis of the human cytomegalovirus genome from the standpoint of horizontal gene transfer and compare our results with existing knowledge from the literature about the genes of this virus. This experiment is of particular relevance given that we set out to create a method that would be suitable for the analysis of large and small genomes. As an example of a small genome to analyze with our Wn-SVM method, we selected the human cytomegalovirus (also known as human herpesvirus 5 or HHV5). The reason for this particular choice is due to our long standing interest in the cytomegalovirus in conjunction with the fact that this is a virus that transmits very easily, knows no age or geographic boundaries, has no seasonal dependencies and affects a very large percentage of the population in modern societies (18–20).

Figure 4 shows a map of the HHV5 genome marked by Wn-SVM. The strain with which we worked with was the laboratory strain AD169 (21). In the absence of detailed knowledge as to the extent of horizontal transfers into the

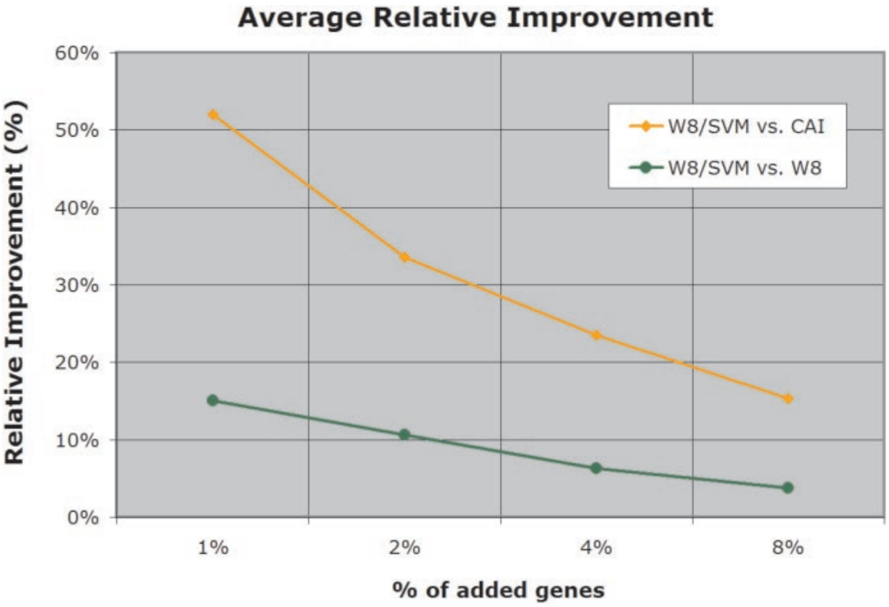

Figure 1. Achieved relative improvement of W8-SVM versus CAI and of W8-SVM versus W8. The results represent an average over all experiments and all genomes (see also text).

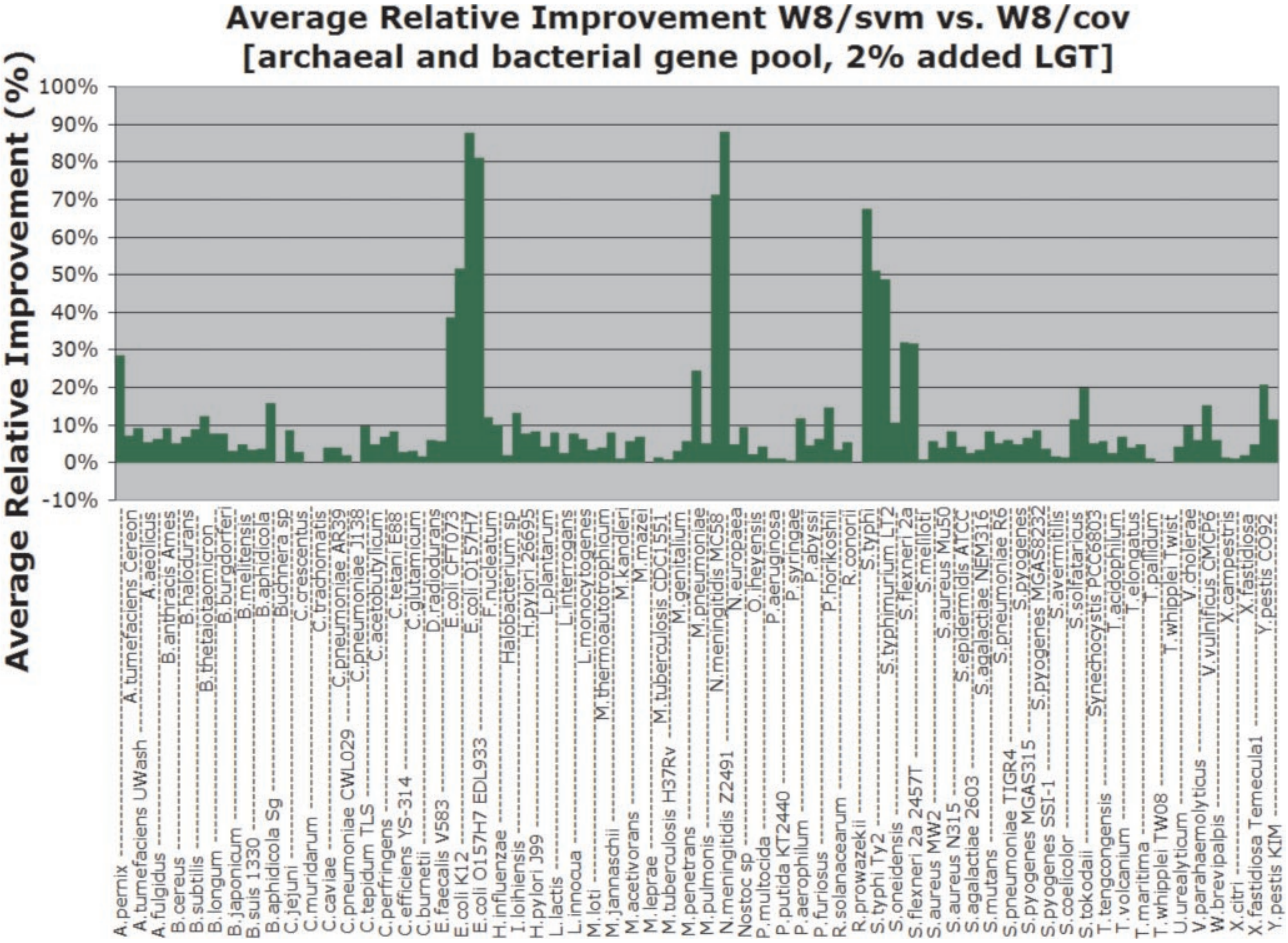

Figure 2. Average relative improvement $\mathrm{Rel}_G^{\mathrm{W8}}$ of W8-SVM over W8 for each one of 123 organisms. Each value is an average over 20 experiments with donor genes drawn from the archaeal and bacterial gene pool (see also text).

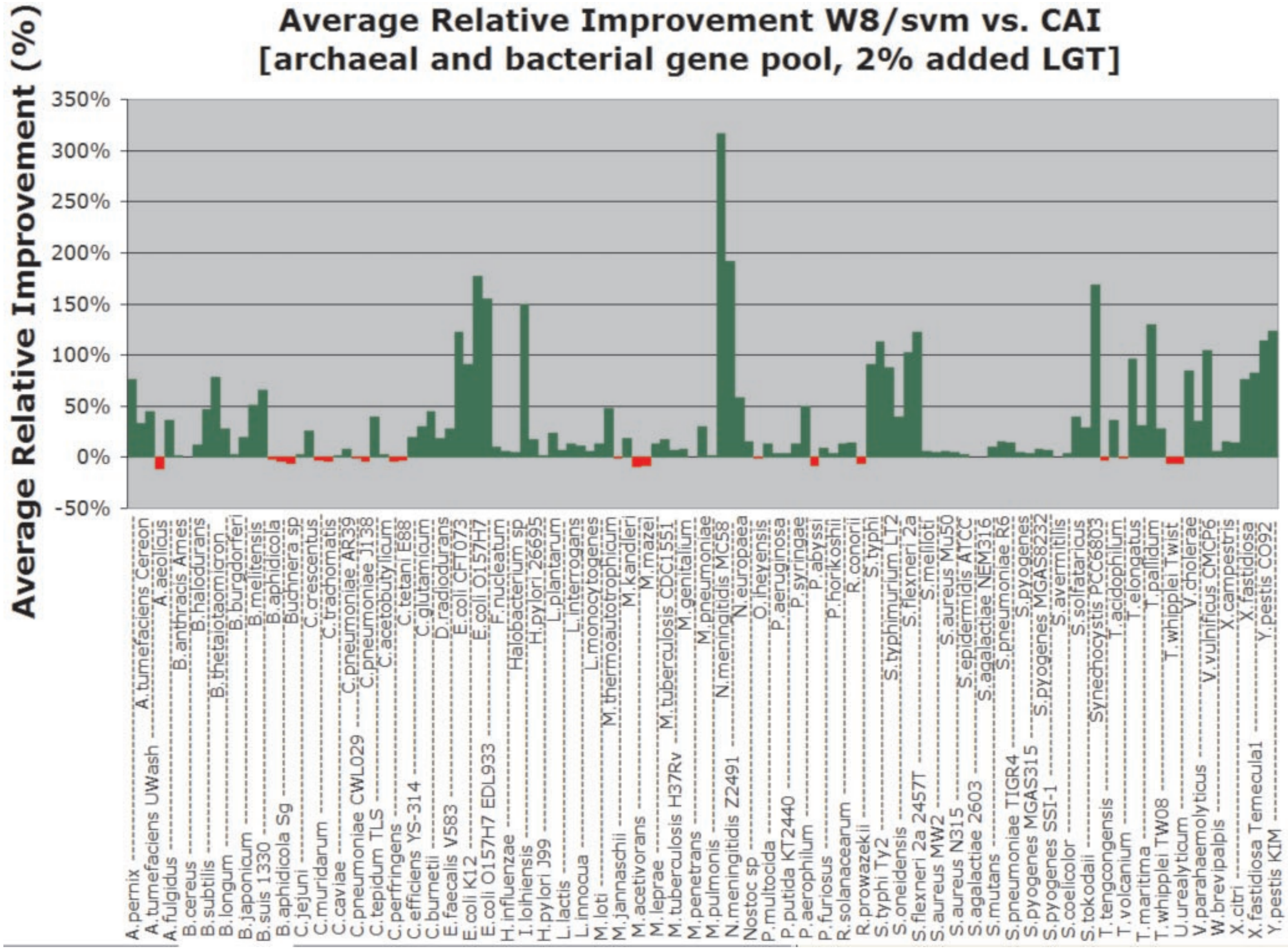

Figure 3. Average relative improvement $\mathrm{Rel}_G^{\mathrm{CAI}}$ of W8-SVM over CAI for each one of 123 organisms. Each value is an average over 20 experiments with donor genes drawn from the archaeal and bacterial gene pool (see also text).

cytomegalovirus genomes, we generated results for three values of v, namely 1.0, 0.9 and 0.8, and reported a region as a candidate for gene transfer if and only if it were marked as atypical by Wn-SVM at all three values of v. Supplementary Figure 1 shows the boundaries of the evaluated genomic regions, the genes that overlap with each region and the similarity score assigned by Wn-SVM to each region. The forward and reverse strands of the genome are treated separately and the genes are shown on their respective strand. The evaluated regions were 300 nt in length and consecutive regions had an overlap of 200 nt.

Several interesting results can be seen from Figure 4 and Supplementary Figure 1. With some very interesting exceptions that clearly demonstrate the capabilities of Wn-SVM and which we will discuss next, effectively every single one of the blocks of genes that are known to be conserved across the β-herpesvirinae is marked by Wn-SVM as typical (native) to the cytomegalovirus genome, precisely as described in (22). These blocks are genes UL22 through UL33, UL45 through UL53, UL69 through UL72, UL75 through UL80, UL85 through UL87, UL89 through UL105, UL112 through UL117, and the TRL/IRL and TRS/IRS regions.

Although the above mentioned blocks of genes are marked as herpesvirinae-specific, there are a few small regions within them with atypical composition. In particular, and as shown in Figure 4, genes UL33, UL78, US12 and US21 are all reported by Wn-SVM as atypical and thus as horizontal transfer candidates. This is in fact a correct result given that all four of these genes are G-protein coupled receptor homologs and thus eukaryotic in origin. Also marked, in a piece-meal fashion this time, was UL48, a gene coding for a virion protein that is known to comprise several distinct, non-contiguous domains (hence the piece-meal marking by Wn-SVM) with eukaryotic character. This was in fact described previously (19)—see relevant entry from Table 1.

A few additional observations are warranted here as they further demonstrate the new method's capabilities and increased sensitivity. First, we would like to point out that several areas of the genomic sequence, outside the gene blocks that are known to be conserved across herpesvirinae, show a typical composition and have been marked as horizontal transfer candidates. This is a very interesting result which does not contradict the current knowledge about the cytomegalovirus and which suggests several new avenues of investigation.

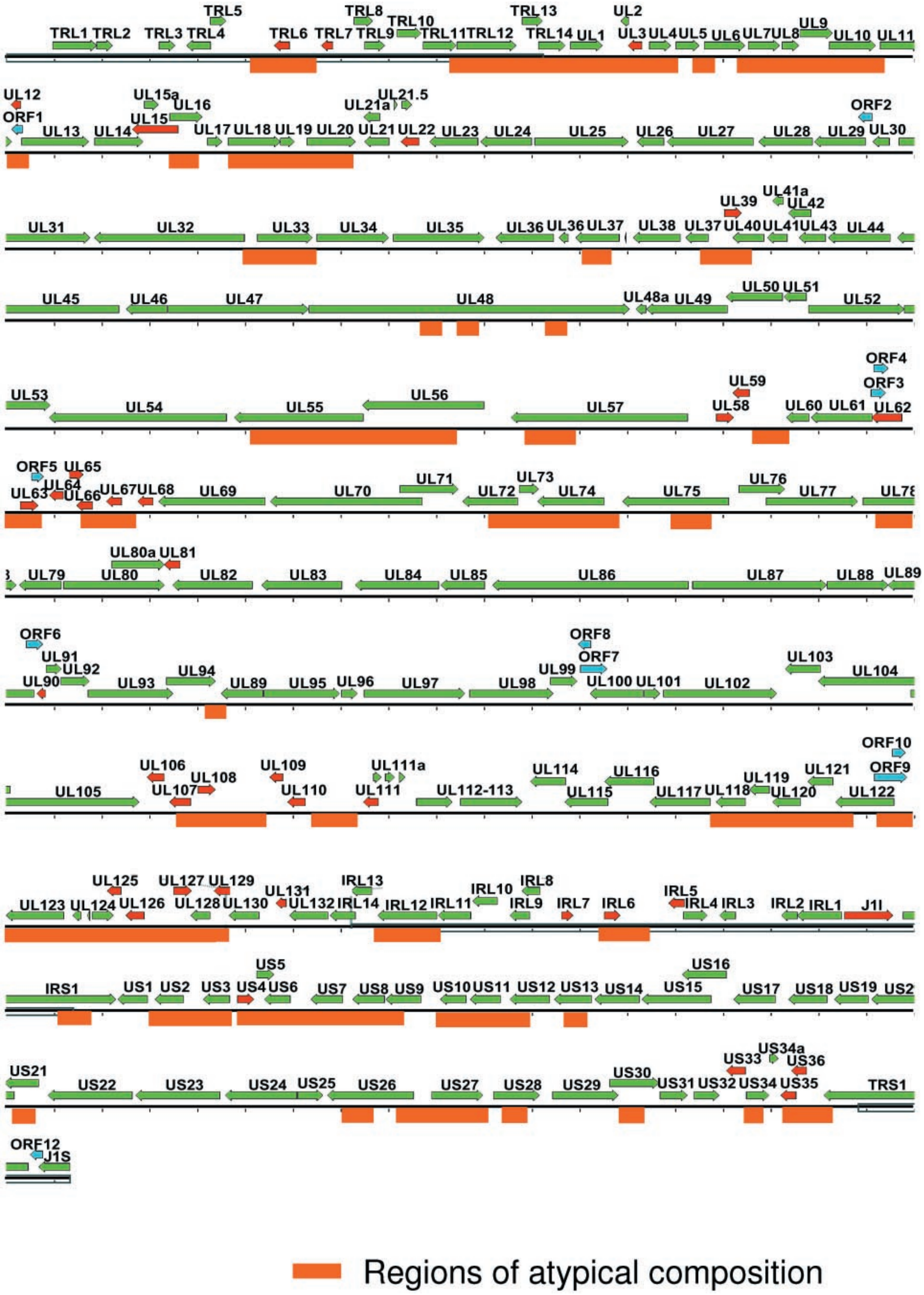
TRL1 TRL2 TRL3 TRL4 TRL5 TRL6 TRL7 TRL8 TRL9 TRL10 TRL11 TRL12 TRL13 TRL14 UL1 UL2 UL3 UL4 UL5 UL6 UL7 UL8 UL9 UL10 UL11
UL12 ORF1 UL13 UL14 UL15a UL15 UL16 UL17 UL18 UL19 UL20 UL21a UL21.5 UL21 UL22 UL23 UL24 UL25 UL26 UL27 UL28 ORF2 UL29 UL30
UL31 UL32 UL33 UL34 UL35 UL36 UL36 UL37 UL38 UL37 UL39 UL40 UL41a UL41 UL42 UL43 UL44
UL45 UL46 UL47 UL48 UL48a UL49 UL50 UL51 UL52
UL53 UL54 UL55 UL56 UL57 UL59 UL58 UL60 UL61 ORF4 ORF3 UL62
ORF5 UL65 UL63 UL64 UL66 UL67 UL68 UL69 UL70 UL71 UL72 UL73 UL74 UL75 UL76 UL77 UL78
UL79 UL80a UL81 UL80 UL82 UL83 UL84 UL85 UL86 UL87 UL88 UL89
ORF6 UL91 UL90 UL92 UL93 UL94 UL89 UL95 UL96 UL97 UL98 UL99 ORF8 ORF7 UL100 UL101 UL102 UL103 UL104
UL105 UL106 UL107 UL108 UL109 UL110 UL111a UL111 UL112-113 UL114 UL115 UL116 UL117 UL118 UL119 UL120 UL121 ORF10 ORF9 UL122
UL123 UL124 UL125 UL126 UL127 UL128 UL129 UL130 UL131 UL132 IRL13 IRL14 IRL12 IRL11 IRL10 IRL8 IRL9 IRL7 IRL6 IRL5 IRL4 IRL3 IRL2 IRL1 J1I
IRS1 US1 US2 US3 US4 US5 US6 US7 US8 US9 US10 US11 US12 US13 US14 US16 US15 US17 US18 US19 US20
US21 US22 US23 US24 US25 US26 US27 US28 US29 US30 US31 US32 US33 US34a US34 US36 US35 TRS1
ORF12 J1S
Regions of atypical composition

Figure 4. Atypical regions (candidate horizontal transfers) in the HHV5 genome, strain AD169.

Another interesting region is the one that genes UL107, UL108 and (in part) UL109 span. The region is marked as atypical and we believe that this is a correct result as well. Indeed, in our earlier analysis of this virus' genome that we described previously (20) concluded that UL106 through UL111 are unlikely to code for genes. This claim was verified by very recent work (23) that has shown that the corresponding 5 kb region does in fact code for a spliced intron. More importantly, this region is not conserved in the murine cytomegalovirus, a strain that is close to the analyzed AD169 strain. Taken together, these observations corroborate the Wn-SVM results regarding the region in question.

Finally, and as shown in Figure 4, the TRL6/7 and IRL6/7 regions are reported by Wn-SVM to be non-native to the human cytomegalovirus. Although this last statement may be in disagreement with the discussion presented previously (22), it bodes well with the more recent findings of (20) according to which these two blocks are unlike the rest of the TRL and IRL regions and may in fact be non-coding.

In closing, we would like to make some general comments on the findings of Figure 4: our analysis suggests that the cytomegalovirus genome comprises numerous regions with atypical composition, including ones that are known to have been the result of horizontal gene transfer, e.g. the G-protein-coupled receptors that we discussed above. If the rest of the atypical regions do indeed correspond to transfer events, then the cytomegalovirus must have incorporated these regions relatively recently [see for example the discussion in (6)]. It would then follow that such events may be happening much more frequently than we might have expected. This is a very important research topic in its own right, which we will be addressing in future work as it escapes the scope of this presentation.

CONCLUSION

In this paper, we continued our earlier work on horizontal gene transfer and introduced a new more sensitive method, Wn-SVM, for detecting atypical composition that is based on a one-class SVM. Wn-SVM utilizes the generalized compositional features that we proposed in our earlier work. Our current work represents a substantial point of departure in that Wn-SVM relies on a distribution-free, one-class SVM method in order to draw conclusions instead of defining an a priori model as in the case of the covariance measure. For each gene in turn, the new method computes a typicality score, which is then used as a proxy for the probability that the gene under consideration has been acquired through a horizontal transfer event.

Additional very important methodological differences involve the manner in which the genome's compositional signature ('reference signature') is now computed. In the earlier, covariance-based method, all genes of the genome at hand contributed equally to the genomic signature. However, in the Wn-SVM method weights are chosen optimally using the maximum margin criterion. As such, Wn-SVM extends the notion of a compositional genomic signature by enforcing genes to contribute their compositional features in a non-uniform fashion. In fact, due to the constraints of the optimization problem, some genes may end up not contributing at all to the genomic signature (they will be assigned a weight of zero). Interestingly enough, preliminary analysis shows that the informational genes are under-represented in this group of signature genes, exactly as anticipated: these genes tend to have atypical compositions and therefore should not be contributing to the genomic signature.

It is also worth pointing out that from a mathematical standpoint, our previous method, (3), can be viewed as a special case of the one-class SVM category of approaches. It in fact corresponds to a fixed parameter $\nu = 1$, which does not necessarily yield the optimal performance. Also, it should be pointed out that although the compositional features used in this paper were based on templates of size 8, further performance improvements may be possible through the application of Gaussian or polynomial kernels on the same features, or through the use of especially designed kernel functions that are applied directly on sequences without any need to first extract the compositional features [see chapter 8 of (24)].

We evaluated the performance of Wn-SVM by carrying out a comparative analysis of W8-SVM, W8 and CAI by inserting random, varying-size collections of genes in each of 123 host genomes (archaea and bacteria) and processing those artificially created genomes with each method in turn. Our findings clearly show that Wn-SVM offers significant sensitivity improvements over *Wn*. We further validated Wn-SVM by demonstrating its applicability to the analysis of smaller viral genomes, an area of research that has to date remained unexplored from the standpoint of horizontal gene transfer. As a case study, we analyzed the genome of the human cytomegalovirus (HHV 5) and showed that we can successfully mark genomic regions as atypical, in direct agreement with earlier independent studies. Finally, we have made available Wn-SVM's predictions for numerous, publicly available archaeal, bacterial and viral genomes on the worldwide web at http://cbcsrv.watson.ibm.com/HGT_SVM/.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Eain Murphy for kindly providing an electronic version of Figure 4, and the anonymous reviewers for their helpful and detailed comments and suggestions. Funding to pay the Open Access publication charges for this article was provided by IBM Research.

Conflict of interest statement. None declared.

REFERENCES

1. Avery,O.T., MacLeod,C.M. and McCarty,M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.*, **149**, 297–326.
2. Syvanen,M. and Kado,C. (2002) *Horizontal Gene Transfer. 2nd edn*. Academic Press, San Diego, CA, USA.
3. Tsirigos,A. and Rigoutsos,I. (2005) A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res.*, **33**, 922–933.

4. Syvanen,M. (1994) Horizontal gene transfer: evidence and possible consequences. *Annu. Rev. Genet.*, **28**, 237–261.
5. Ochman,H., Lawrence,J.G. and Groisman,E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.
6. Lawrence,J.G. and Ochman,H. (1998) Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl Acad. Sci. USA*, **95**, 9413–9417.
7. Ikemura,T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, **2**, 13–34.
8. Karlin,S., Mrázek,J. and Campbell,A.M. (1998) Codon usages in different gene classes of the *Escherichia coli* genome. *Mol. Microbiol.*, **29**, 1341–1355.
9. Karlin,S. and Mrázek,J. (1996) What drives codon usage in human genes? *J. Mol. Biol.*, **262**, 459–472.
10. Andersson,S.G.E. and Kurland,C.G. (1990) Codon preferences in free-living microorganisms. *Microbiol. Rev.*, **54**, 198–210.
11. Hooper,S. and Berg,O. (2002) Detection of genes with atypical nucleotide sequence in microbial genomes. *J. Mol. Evol.*, **54**, 365–375.
12. Pride,D.T. and Blaser,M.J. (2002) Identification of horizontally acquired genetic elements in *Helicobacter pylori* and other prokaryotes using oligonucleotide difference analysis. *Genome Lett.*, **1**, 2–15.
13. Ragan,M.A. (2001) On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol Lett.*, **201**, 187–191.
14. Garcia-Vallve,S., Romeu,A. and Palau,J. (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.*, **10**, 1719–1725.
15. Garcia-Vallve,S., Guzman,E., Montero,M.A. and Romeu,A. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.*, **31**, 187–189.
16. Waldor,M.K. and Mekalanos,J.J. (1996) Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science*, **272**, 1910–1914.
17. Scholkopf,B., Platt,J.C., Shawe-Taylor,J., Smola,A.J. and Williamson,R.C. (2001) Estimating the support of a high-dimensional distribution. *Neural Comput.*, **13**, 1443–1471.
18. Novotny,J., Rigoutsos,I., Coleman,D. and Shenk,T. (2001) *In silico* structural and functional analysis of the human cytomegalovirus (HHV5) genome. *J. Mol. Biol.*, **310**, 1151–1166.
19. Rigoutsos,I., Novotny,J., Huynh,T., Chin-Bow,S., Parida,L., Platt,D., Coleman,D. and Shenk,T. (2003) *In silico* pattern-based analysis of the human cytomegalovirus (HHV5) genome. *J. Virol.*, **77**, 4326–4344.
20. Murphy,E., Rigoutsos,I., Shibuya,T. and Shenk,T.E. (2003) Re-evaluation of human cytomegalovirus coding potential. *Proc. Natl Acad. Sci. USA*, **100**, 13585–13590.
21. Chee,M.S., Bankier,S., Beck,S., Bohni,R., Brown,C.R., Horsnell,T., Hutchisno,C.A.,III, Kouzarides,T., Martignetti,J.A., Preddie,E. *et al.* (1990) *Curr. Top. Microbiol. Immunol.*, **154**, 125–169.
22. Roizman,B. and Pellett,P.E. (2001) The family herpesviridae: a brief introduction. In Fields,B.N., Knipe,D.M., Howley,P.M., Chanock,R.M., Monath,T.P., Melnick,J.L., Roizman,B. and Straus,S.E. (eds), *Fields Virology*. Lippincott Williams & Wilkins.
23. Kulesza,C.A. and Shenk,T. (2004) Human cytomegalovirus 5-kilobase immediate-early RNA is a stable intron. *J. Virol.*, **78**, 13182–13189.
24. Cristianini,N. and Shawe-Taylor,J. (2000) *An introduction to Support Vector Machines*. Cambridge University Press, Cambridge, New York, ISBN: 0 521 78019 5.

APPENDIX

LEMMA. *For $v = 1$, the solution vector u of the optimization problem is equal to the feature vector of the entire genome $\phi(G)$, i.e. the genome compositional signature, defined as the average of the feature vectors of all the genes in the genome.*

PROOF. For $v = 1$, the constraints of the dual problem are simplified to $0 \leqslant \alpha_i \leqslant 1/l$ and $\sum_{i=1}^{l} \alpha_i = 1$. These constraints can only be satisfied if all α_i attain the maximum allowed value, i.e. if $\alpha_i = 1/l$. This is the only feasible point for the optimization problem, and therefore it must also be the optimal solution. This means that:

$$u = \sum_{i=1}^{l} \alpha_i \phi(x_i) = \frac{1}{l} \sum_{i=1}^{l} \phi(x_i) = \phi(G).$$

From this Lemma, we immediately conclude that because the two typicality measures are proportional to each other, they will induce identical rankings, and therefore the two methods will produce identical results with respect to identifying atypical genes in a genome.