Separation of Sequences from Host-Pathogen Interface Using Triplet Nucleotide Frequencies

Jeppe Emmersen[1], Stephen Rudd[2], Hans-Werner Mewes[3] and Igor V. Tetko[3,4,*]

1 -- Institut for Miljø og Bioteknologi, Aalborg Universitet, Sohngaardsholmsvej 49, 9000 Aalborg Denmark

2 -- Centre for Biotechnology, Tykistökatu 6, FIN-20521 Turku, Finland

3 -- GSF - Institute for Bioinformatics (MIPS), 85764 Neuherberg, Germany, D-85764 Neuherberg, Germany

4 -- Institute of Bioorganic & Petrochemistry (IBPC), Ukrainian Academy of Sciences, Kyiv, UA-02660, Ukraine

Address for correspondence: Igor V. Tetko

GSF – National Research Centre for Environment and Health
Institute for Bioinformatics (MIPS)
Ingolstädter Landstraße 1,
D-85764 Neuherberg, Germany
Telephone: +49-89-3187-3575
Fax: +49-89-3187-3585
e-mail: i.tetko@gsf.de

Abstract

The identification of genes involved in host-pathogen interactions is important for the elucidation of mechanisms of disease resistance and host susceptibility. A traditional way to classify the origin of genes sampled from a pool of mixed cDNA is through sequence similarity to known genes from either the pathogen or host organism or other closely related species. This approach does not work when the identified sequence has no close homologues in the sequence databases. In our previous studies we classified genes using their codon frequencies. This method, however, explicitly required the prediction of CDS regions and thus could not be applied to sequences composed from the non-coding regions of genes.

In this study we show that the use of sliding-window triplet frequencies extends the application of the algorithm to both coding and non-coding sequences and also increases the prediction accuracy of a Support Vector Machine classifier from 95.6±0.3 to 96.5±0.2. Thus the use of the triplet frequencies increased the prediction accuracy of the new method by more than 20% compared to our previous approach. A functional analysis of sequences detected gene families having significantly higher or lower probability to be correctly classified compared to the average accuracy of the method is described. The server to perform classification of EST sequences using triplet frequencies is available at http://mips.gsf.de/proj/est3.

1. Introduction

The completion of the draft genome sequences for several important plant pathogenic fungi such as *Magnaporthe grisea*, the causal agent of rice blast, and *Ustilago maydis*, the causal agent of corn smut, has marked a new era for the study of plant pathogens (Dean, Talbot et al. 2005). Access to these large-scale genomic data have allowed researchers to compare pathogenic fungi to non-pathogenic fungi, such as *Aspergilus nidulans* and *Neurospora crassa* and thus identify new genes conferring pathogenic properties (Galagan, Calvo et al. 2003). The Fungal Genome Initiative (FGI) hosted at The Broad Institute currently has 24 fungal genomes in the pipeline, of which 19 are in the finishing phase with initial assembly details and open reading frame predictions available, (http://www.broad.mit.edu/annotation/fgi/, July 2006). Other fungal species from various other projects are in the sequencing pipeline (see e.g., http://mips.gsf.de/projects/fungi or http://www.jgi.doe.gov/) but remain in the finishing and annotation process (July 2006). For a review of other public domain efforts, see (Yarden, Ebbole et al. 2003).

Although the number of sequenced fungal genomes is steadily increasing, it is unlikely that complete genomes will be sequenced for all the different pathogenic fungi or the fullest range of host plant species. This is due to both the cost of sequencing and because the subsequent annotation of each genome represents a major challenge and investment. Many species will therefore remain uncharacterized at the genome level in the foreseeable future. Sequencing of Expressed Sequence Tags (EST) remains the best alternative for gene discovery in plants and plant pathogens (Adams, Soares et al. 1993; Panabieres, Amselem et al. 2005; Posada-Buitrago and Frederick 2005). With broad EST sequencing from cDNA libraries, unique representations of expressed genes can be constructed through the assembly of identical and overlapping ESTs into unigenes (Miller, Christoffels et al. 1999; Pertea, Huang et al. 2003; Rudd 2005). For plant pathogens, EST sequencing of cDNA derived from mixed plant/pathogen tissue represents a special problem, since the resulting cDNA library will contain a mixture of clones derived from the plant and the pathogen (in addition to laboratory DNA contaminants such as *E. coli*, Lambda phage fragments, unspliced mRNA and fragmented genomic DNA). This mixing of mRNA is increased with obligate biotrophic parasites when the mRNA must be sampled from infected plant tissue. This is further compounded when the pathogen is endoparasitic - the plant may be completely invaded, resulting in a spread of the microorganism into all tissues. An example of this may be shown

with the fusarium wilt of tomatoes, caused by *Fusarium oxysporum f. sp. Lycopersici* (Bishop and Cooper 1993). The pathogen invades the vascular system of the plant, resulting in complete loss of turgor. Since gene expression can be radically different between the cultured state and its infectious counterpart, it might be preferable to work with mixed tissues. For the pathogen *M. grisea*, it was shown that expression of the gene MPG1 increased ~60 fold from *in-vitro* growth to infectious growth (Talbot et al, 1993). However, little is known about the later stages of infection and regulation of growth within the host tissue although *M. grisea* has been shown to exhibit intra- and intercellular growth of the fungus in the host (Viaud, Balhadere et al. 2002; Sesma and Osbourn 2004). The general pattern of gene expression might well change during these stages of growth.

However, instead of regarding this mixing process as a problem, it can be thought of as a novel method to gain insight into simultaneous gene expression of both plant and pathogen. This requires a method to classify the sequences according to their origin with a reasonable degree of confidence, without needing to know an entire genome sequence beforehand. Previous attempts limited to a few plant/pathogen pairs showed that it is possible to classify sequences from mixed libraries. Such methods include a probabilistic approach (Maor, Kosman et al. 2003), the GC counting method (Huitema, Torto et al. 2003), likelihood methods (Hraber and Weller 2001) and Support Vector Machines (SVM) (Friedel, Jahn et al. 2005; Rudd and Tetko 2005). SVM showed significantly higher prediction accuracy compared to previously reported results. However, given the small number of organisms included in previous studies, the general level of classification accuracy could not be evaluated with confidence. Because of the enormous diversity of microbes, any attempt for a generalized prediction method should include as many species as possible (Bennet 1997).

The need to calculate codon frequencies was an important limitation of our previous attempt to categorize sequences using SVM classification (Friedel, Jahn et al. 2005; Rudd and Tetko 2005). This requirement required an additional layer of complexity to find the in-frame coding sequence and discarded information buried within the non-coding parts of the mRNA sequence. Moreover, some sequences may contain a significant percentage, or consist entirely, of non-coding or untranslated regions, especially in the case of ESTs sequenced from the 3' end of cDNA clones (Seki, Narusaka et al. 2002). These classification problems would be reduced, if a simpler statistics such as trinucleotide frequencies (triplets) or higher orders of nucleotide organization can be used as input for the classifier. A non-codon based statistics

also means that the entire mRNA sequence and not just the coding part can be used for classification.

Here we investigate the classification of species origin for complete mRNA sequences, including their untranslated regions (UTR) in a large dataset set of 30 organisms representing plants of Dicotyledonous (15) and Monocotyledonous (5) origin as well as ten different plant pathogens or plant saprophytes, table 1. Classification accuracy using sliding windows of dinucleotide, triplet, quadruplet and hexamer nucleotide frequencies, and in-frame codon frequencies (codon usage) are compared for all pairs of organisms.

In our study we have additionally tried to address a number of other important questions: Can we *a priori* estimate an accuracy of separation of plant-pathogen genomes according to their differences in GC-content? Is it more difficult to separate sequences from one or different kingdoms? Are there some families of genes, which are more easy/difficult to separate, thus creating a group bias in the classification? How many sequences are needed to calculate a generalized model?

2. MATERIALS AND METHODS

2. 1 Host-pathogen mRNA-derived data

The aim of our study was to investigate the classification accuracy of distinguishing mixed collections of plant and fungal pathogen EST sequences. However, for the initial analysis we choose to use full or partially characterized mRNA sequences for the evaluation of SVM performance. The sequence fidelity of mRNA was presumed to be higher for mRNA sequences than EST sequences, thus minimizing noisy data due to sequencing errors prevalent in EST sequences. Based on mRNA sequence availability, full or partial mRNA sequences were obtained from GenBank or the Pedant Fungi database at MIPS (see table 1; Benson, Karsch-Mizrachi et al. 2005; Guldener, Mannhaupt et al. 2006). Unlike EST sequences, most mRNA sequences deposited in GenBank contain at least annotation of the coding region and often annotation of the UTR regions as well. To minimize classification bias caused by sequence redundancy we removed sequences, which were 95% identical by clustering of nearly identical sequences using the Blastclust program. The level of 95% identity was chosen to reflect normal levels of EST clustering (Quackenbush, Cho et al., 2001). Blastclust, v. 2.2.10, was evoked with the parameters -S 95 –b F –p F, and the longest sequence was kept for each cluster (Altschul, Madden et al. 1997).

A minimum of 100 mRNA sequences was chosen as the limit for inclusion. The number of non-redundant sequences included for each species ranged from 103 (*P. infestans*) to 70916 (*A. thaliana*). The number of annotated sequence features such as UTR regions was usually less than the total mRNA sequences, as some mRNA records only contained coding sequence. The 5' UTR set was smallest of all sequence sets with a mean sequence length of 92 bp. mRNA sequences from genome based sequence predictions were included if present in GenBank. To compensate for missing UTR data in organisms with annotated genome sequences, we constructed 5' and 3' UTR sequences by including 100 bp upstream of the coding region for 5' UTRs and 300 bp downstream for 3' UTRs to match the average length of UTR regions (3' ~330 bp, 5' ~ 90 bp) detected for genomes with UTR regions.

Validated models of mRNA sequences from genome sequencing of *N. crassa, F. graminearum, A. nidulans, M. grisea, S. sclerotiorum*, and *S. nodorum* were obtained from the MIPS Fungal genome database (http://mips.gsf.de/projects/fungi/fungi_db.html). The Oomyceetes *P. sojae* and *P. ramorum* predicted mRNA sequence sets (version 1) were downloaded from the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/.

2.2 EST sequences from host-pathogen pairs

To validate the performance found with annotated mRNA sequences, we extracted EST sequences from 23 organisms, from which at least 1000 unigenes were available (see table 1). The openSputnik software was used to prepare unigene sets and extract CDS data as described previously (Rudd 2005). The openSputnik performs vector clipping, pre-clustering of EST sequences using the HPT2 algorithm and final assembly of unigenes, using the CAP3 algorithm. Coding regions are predicted using a combination of a BlastX search to known proteins and ESTScan based predictions. For EST-based validation of the SVM classifiers of each pair, we used both the entire unigene set (UNI) as well as two subsets; one where either the coding sequence (CDS+) could be extracted from each unigene set and one where no CDS regions were found (CDS-). The CDS+ set is presumed to contain less contaminating sequences, such as unspliced mRNA and genomic DNA.

2. 3 Calculation of dinucleotide bias

Dinucleotide bias and relative dinucleotide distances were calculated according to (Gentles and Karlin 2001). The dinucleotide bias r^* is defined as

$$r^*_{XY} = f^*_{XY}/f^*_X f^*_Y, \qquad (1)$$

where f^*_X and f^*_Y are the frequencies of nucleotide X and Y respectively and f^*_{XY} is the frequency of each dinucleotide XY, calculated for each sequence with its reverse complement concatenated. The overall dinucleotide bias was calculated for each species over the total sequence set.

The relative dinucleotide abundance distance for two sequences p,q is defined as

$$d^*(p,q) = 1000/16 \, \Sigma_{XY} | r(p) - r(q) | , \qquad (2)$$

where the summation is over all dinucleotides. The relative dinucleotide abundance distance for two species was calculated using the first 100 kb of each sequence set.

2.4 Model development and validation

Input data for the SVM classifier was either the relative frequencies (probabilities) of di-, tri-, quadruplets and hexanucleotides generated by sliding windows over the nucleotide sequence or discrete windows of three nucleotides (in-frame codon frequencies) for coding (CDS) regions. Thus, for example for triplets a 64-dimensional input vector corresponding to frequencies of 4^3=64 possible combinations of four codon letters (A,T,C,G), i.e. AAA, AAT, … GGG was used. If the frequency of a particular *n*-mer was zero in the training set, this *n*-mer was excluded from the analysis. This was particular the case for quadruplets and hexamers, which projected the input sequences to much higher dimensional space (256 and 4096 for quadruplets and hexamers, respectively). The output values {1,-1} were used to indicate if the analyzed EST sequence had plant or pathogen origin. Unless stated otherwise, performance results are stated for triplet frequencies calculated from sliding windows. The input data for the SVM algorithm were normalized to a (0, 1) interval before the data analysis.

The sequence sets from host and pathogen species varied in size up to 700 fold (*A. thaliana* vs. *P. infestans*). In order to avoid any bias from imbalanced data sets, we always selected an equal number of sequences from each genome (by random from genomes with large amount of data) for the model development. In addition, for all reported studies, unless otherwise mentioned, the maximum number of sequences per genome included in the training set was 1,000. The remaining sequences were used as independent test sets for further validation of the method. The test sets were not used at any stage of model development.

We assumed equal costs for both type I and II classification errors and used the accuracy of predictions (defined as percentage of correct predictions of sequences) as a measure of the performance of the SVM (Chawla, Bowyer et al. 2002). In the case of imbalanced datasets, for example assuming higher classification error cost for pathogen ESTs, different criteria such as sensitivity, positive prediction value, Receiver Operator Curve (ROC), Area Under the ROC Curve (AUC) could be applied. For each analysis the mean accuracy and standard error of the mean are reported. For example, for plant-plant classification, there were k=20 genomes. Thus $N = k!/((k-2)!2!)$ = 19*20/2 = 190 genome pairs were considered to calculate mean accuracy and standard error of the mean.

We used the open source LibSVM package, version 2.8 (Chang and Lin 2005). In SVM learning, the input variables are first mapped into a higher dimensional feature space by the use of a kernel function, and then a linear model is constructed in this feature space. For the purposes of the current study we restricted our analysis to the RBF (Radial Basis Function) kernel, which was also used in previous studies and demonstrated superior prediction performance (Friedel, Jahn et al. 2005). The RBF kernel $k(\mathbf{x},\mathbf{x}')=exp(-\gamma|\mathbf{x}-\mathbf{x}'||)$ may behave like the linear and sigmoid kernels for certain values of kernel width parameter, γ. Another important parameter of the SVM, C, is used to account for mislabeled data (noise) in the training set and additionally provides soft-margin classification. Using the RBF kernel, only these two parameters need to be optimized: thus a simple grid-search on C and γ was performed for each dataset using the internal cross-validation procedure of the LibSVM program. The range of the cost parameter C was 2^{-5}, 2^{-3}, … , 2^{15} and the range of the kernel width parameter γ was 2^{-15}, 2^{-13}, …, 2^{3} as recommended in the LibSVM manual.

We employed a double cross-validation (see fig. 1 and also (Tetko, Solov'ev et al. 2006)). All compounds in each training data set were split five times on subsets containing 4/5 of all molecules (“cross-validation training sets”) and their complements containing 1/5 of all molecules (“cross-validation test sets”). The cross-validation training sets were employed to optimize internal parameters of SVM and the selected parameters were applied to predict corresponding cross-validation test sets as well as independent test set. No significant differences in model performances estimated using double cross-validation procedure and prediction of the independent test set molecules were found (see Results section).

A typical analysis of all data for one condition (e.g., estimation accuracy of EST predictions using triplets for sequence length of 50 base pairs, fig. 5) generated 435 tasks and required about one day of CPU time on a cluster of 14 64-bits Athlon computers.

3. RESULTS

3.1 Classification accuracy of n-mers vs codon usage

By including untranslated regions in the training process, it was not known which kind of nucleotide representation was best suited for classification. Thus, we developed SVM classifiers using different *n*-mers representations of complete mRNA and compared their performance to the performance of SVM classifiers developed using in-frame codon frequencies in the CDS region of mRNA.

The use of triplet frequencies derived from complete mRNA sequences (96.5±0.2) significantly improved the average prediction performance of models compared to those calculated with in-frame codon (95.2±0.3) or triplet frequencies (95.8±0.3) for the CDS component (fig. 2). Codon frequencies can also be formally used for the complete mRNA data. Such application of the codon frequency method, calculated a prediction accuracy of 95.6±0.3, which is significantly lower than the use of sliding triplet windows (96.5±0.2). The classifiers developed with codon frequencies provided lower prediction accuracy compared to the triplet-based classifiers for datasets with shorter average sequence length. A negative linear correlation was observed (r = -0.4, N = 200) between the average sequence length of genes in the dataset and the decrease in performance of codon vs triplet-based methods for N = 20*10 = 200 plant-pathogen genome pairs. For short sequences the sliding triplet frequencies extracted more information compared to the codon frequencies.

The use of dinucleotides or quadruplets decreased prediction performance of the SVM methods for mRNA compared to triplet usage. The use of hexanucleotides demonstrated the lowest prediction ability (data not shown). Since triplet frequencies provide the highest prediction ability for the method, we decided to use this representation of sequences for all subsequent studies.

3.2 Classification accuracy of UTR regions

To estimate the classification accuracy of non-protein coding sequence, we extracted DNA sequence corresponding to the 3' UTR and 5' UTR regions and compared its performance with classifiers developed using the entire mRNA sequence.

Classification performance using either 3' UTR or 5' UTR regions was not as accurate as using a complete mRNA sequence. The UTR regions (3' ~330 bp, 5' ~ 90 bp) are much shorter than complete mRNA sequences (>1000 bp), table 2. However, these regions still contribute to the sequence classification, since accuracy using whole mRNA sequence was significantly better than using just the CDS (fig. 2).

The classification accuracy was in excess of 90% for the separation of plant/fungi and fungi/fungi, table 2. The average accuracy between plant/plant pairs was in general more than 10% less compared to plant/fungi and fungi/fungi. This is perhaps a reflection of reduced evolutionary diversity among the plants.

3.3 Correlation of SVM performance with different measures of nucleotide composition

Organisms exhibit unique dinucleotide signatures that can be shown by comparing global dinucleotide biases from completely sequenced organisms (Karlin 1998). The GC content of a given sequence can be used to determine the origin of the sequence, if the difference in GC content between the species is large enough. Thus, we wanted to know how different measures of nucleotide composition, GC content and dinucleotide bias would correlate with the discriminating power of the SVM.

For each pair of organisms, we calculated the difference in their GC content and compared this to the cross-validated performance of the SVM classifier. We used the Pearson linear correlation coefficient to correlate nucleotide composition with classification performance of 435 pairs of organisms. For the complete mRNA sequence sets, we found a high correlation coefficient ($r = 0.74$, $N = 435$) between GC content and SVM performance. The G+C content is a one-dimensional variable, which provides little information in itself. This was demonstrated by the higher correlation ($r = 0.82$, $N = 435$) between the dinucleotide bias distance and SVM performance for mRNA sequences (fig. 3). Dinucleotide distances do not explain all the performance, as evolutionary distances are also important for accurate classification. Plant-fungi pairs showed higher classification accuracy compared to plant-plant

pairs (fig. 3). Firstly, plant-fungi pairs have larger dinucleotide distances, which contribute to higher classification accuracy of this data. Secondly, plant-fungi data pairs demonstrate significantly higher classification accuracy compared to plant-plant pairs with exactly the same dinucleotide distances. For example, there are 58 and 109 plant-plant and plant-fungi pairs with dinucleotide distances in ranging from 150 to 200 units. The average classification accuracy of plant-plant pairs, 84.6%, is significantly lower compared to that of plant-fungi pairs, 96.5%. When comparing separation of plant pairs, we found a value of 90.7% for dicot/monocot separation, whereas for dicot/dicot and monocot/monocot the average accuracies were 81.6% and 79.0%. For monocot/pathogen pairs the separation was 94.3% and 98.0% for dicot/pathogen pairs.

Although performance is generally poor when dinucleotide distances are below 50 for plant-plant pairs, plant-pathogen species pairs in this range are well separated with an average accuracy above 90%. At a dinucleotide distance of >100, no pair of organisms fell below 95% accuracy. It is interesting to note, that the spread of classification of fungi-fungi pairs is similar to that of plant-plant pairs (fig. 3). Thus, SVM accuracy between kingdoms is less correlated with dinucleotide distance than within kingdoms. This shows that classification of sequences from plant and fungi can be performed with high accuracy even though nucleotide compositions are similar.

3.4 SVM accuracy as function of the number of sequences in the training set

The number of unigenes or mRNA sequences may be limited for some organisms. Thus the accuracy of the SVM classifier was tested using different number of sequences for training, ranging from 50 to 2000 sequences per organism as input (fig. 4). Some organisms were limited with respect to the number of sequences available; these were omitted from this analysis (table 1). This exclusion increased standard errors of the mean for datasets with larger number of sequences (fig. 4). For plant-plant pairs having smaller dinucleotide distance between genomes, the training set size was crucial to achieve a better separation. For plant-fungi pairs the number of sequences was less important and reliable predictions could be obtained from the independent test sets with as few as 100-200 sequences used for model generation. Fig. 4 demonstrates that accuracies estimated by cross-validation and observed for the test sets are not significantly different. Some test sets were more than 1,000 times larger than the training sets for classifiers developed using only 50 sequences per organism.

3.5 SVM accuracy as function of sequence length

A typical EST sequence varies between 100 bp to 700 bp in length, depending on the sequence quality. To analyze the effect of sequence length on classification accuracy, the SVM classifier was tested using different lengths of mRNA sequences ranging from 50 bp to 1,000 bp, using 1,000 sequences per organism (fig. 5). If A sequence was longer than the specified length, we randomly picked a subsequence of the specified length. As expected, the accuracy of prediction increased with the length of the sequence (fig. 5). In agreement with previous results, the classification of sequences within kingdoms was less accurate than between the kingdoms.

3.6 Classification of Unigene data derived from the openSputnik repository

The previous sections relied on mRNA sequences to evaluate the performance of the SVM approach. These were either manually validated sequences or data from sequencing projects. Public ESTs collected in public repositories, e.g. openSputnik, represent less ideal datasets. These data are usually produced as result of automatic sequencing projects and contain significant contamination, sequence length variation and frame shift errors. Some errors could be removed during processing and assembly of ESTs into unigenes but genomic contamination and unspliced mRNA remain. It is essential to compare the performance of clean "gold-standard" mRNA dataset with unigene dataset automatically derived from openSputnik.

As described in section 2.2 only organisms with at least 1000 unigenes were used for analysis. In addition to the whole set of unigenes (UNI) we also subdivided it on sets of genes for which the coding sequences could be (CDS+) or could not be (CDS-) extracted from the openSputnik pipe-line (Rudd 2005).

Only the coding part of the CDS+ set sequences was used in the first analysis. The average performance of the SVM model trained using triplet frequencies (93.9±0.3) provided higher accuracy compared to the codon frequencies (93.3±0.3). This result is in agreement with our finding for the mRNA analysis: the use of triplet frequencies improves the prediction performance of the SVM compared to the codon usage.

The prediction performance of the triplet-based classifiers decreased to 92.1±0.3% when the complete UNI set was analyzed. This can be explained by higher noise in the whole EST-

derived unigene set compared to the CDS+ data. Indeed, a presence of coding part in the unigene identified by BlastX to known genes guarantees a higher quality of the data. Contrary to the CDS+ set, some of the CDS- unigenes could contain significant sequence contamination. Indeed, the prediction accuracy for the CDS- subset was significantly lower, 83.3±0.5% compared to the average result for the UNI set. At the same time the accuracy for the CDS+ subset was 93.8±0.3%, i.e. corresponded to the 93.9±0.3% calculated using only the CDS+ data. Thus SVM accuracy for the CDS+ set did not significantly decrease due to presence of possible contaminations in the UNI set.

The EST-trained classifier provided an average accuracy of 95.6±0.04% for the mRNA set, i.e. demonstrated a comparable performance to that of the mRNA classifiers. A large number of unigenes without CDS regions for a specific organism does not always result in low prediction accuracy. For example, the number of unigenes without CDS data ranged from less than 10% in *P. vulgaris* to about 40% in *H. vulgare*. Both these species demonstrated a high average accuracy of classification, 96.7% (*P. vulgaris*) and 94.1% (*H. vulgare*), when paired with *U. maydis*.

3.7 Functional annotation of classified sequences

In order to analyze whether there was a systematic bias of classification with respect to genes with similar function, we investigated the general functional connection between sequences with true or false classifications of plant-pathogen pairs. The sequence annotation was performed according to the FunCat functional annotation scheme (Ruepp, Zollner et al. 2004). Manually classified sequences from *A. thaliana, N. Crassa, F. graminearum* and *S. cerevisiae* were used to annotate classified sequences by automatic transfer of function based on BlastX sequence similarity (only matches with e-values $< 10^{-5}$ were considered). The sequences annotated to FunCat category 38 – Transposable elements, viral and plasmid proteins and category 99 – Unclassified proteins had significantly lower accuracy of classification compared to other proteins ($p<0.01$, Bonferoni corrected). Contrary to that, proteins annotated with categories 01 – Metabolism, 02.01 – glycolysis and gluconeogenesis and 12 – Protein synthesis had significantly higher accuracy of classification compared to other proteins.

3. DISCUSSION

Using mRNA and EST sequence data from 30 different plant and fungal organisms, we were able to show the general applicability of a binary SVM classifier to determine the origin of

sequences in mixed sequence sets. Furthermore, we showed that the entire mRNA sequence should be used for classification and not just the coding region, greatly simplifying the classification procedure and expanding the applicability to non-coding mRNA regions. The models developed in this study have been made available to the public by a web server (http://mips.gsf.de/proj/est3). Following classification of new data, the users are prompted to a page with statistical report of the used classifier. The web server also allows new host/pathogen pairs to be explored, as new classification models can be constructed by uploading sequence data to the server.

Our results showed that using sliding windows to calculate triplet frequencies provided the best overall performance for mixed mRNA data. The use of triplet frequencies also provided higher prediction performance of SVM compared to the use of in-frame codon frequencies for EST unigenes. Using sliding windows to calculate triplet frequencies probably provides some limited information on immediate sequence neighbors, expanding the input space for the SVM classifier compared to the in-frame codon usage frequencies. The use of higher order n-mer frequencies, e.g. quadruplets and hexamers, projected the input sequences to much higher dimensional space (256 and 4096 for quadruplets and hexamers, respectively compared to 64-dimensional space for triplets). This probably produced too sparse representation of sequence for SVM classification and lowered its prediction ability.

Contrary to mRNA data, the EST-derived unigene sets may contain a degree of contamination including unspliced mRNA and genomic contaminants. This explains a lower classification performance from complete unigene sets compared to the unigene subsets of predicted CDS regions (CDS+). We can assume that contaminants are classified by chance. If this is the case, the observed 10% decrease in the average performance of the unigene subsets using predicted CDS sequences (93.8%) and entire unigene sets (83.3%) can be explained by the existence of contaminants. Supposing that most contaminated sequences are present in the CDS- subset, which on average constituted ~20% of the entire unigene sets, the level of EST contamination for the complete unigene sets can be estimated at ~4%. Indeed, a 10% of decrease in performance for prediction of unigenes without CDS (CDS-) is explained by a 20% level of contamination of this subset, which equals 4% contamination of complete unigene sets. This percentage agrees with similar estimations of large unigene collections (Rudd 2005). The presence of the sequence contamination within unigene sets also explains overall decrease in prediction accuracy of classifiers developed using EST data compared to those developed using mRNA collections. Still the average classification accuracy calculated for separation of

EST data from fungi and plant was above 92%, which should be sufficient for most practical purposes. Future improvements might include a calculation of the confidence score for each prediction could better differentiate reliable vs non-reliable prediction and thus be used to achieve higher prediction accuracy of classifiers within their applicability domains (Tetko, Bruneau et al., 2006).

When scanning complete genomes, dinucleotide bias is mostly invariable in windows of 50kbp across the genome, except for repetitive sequence (Gentles and Karlin 2001). Assuming a mean length of 1500 for mRNA sequences, this corresponds to 33 complete mRNA sequences. Using 50 sequences as input for training, we find an average accuracy of 91.5±0.4%, which increases to 96.5±0.2% using a limit of 200 sequences for plant/fungi classification. The increase in accuracy indicates that the input space for the SVM classifier is not sufficiently exhausted with 50 sequences (fig. 4). However, no increase in accuracy for plant-fungi classification was observed using an input space of more than 200-500 sequences (depending on the host/pathogen pair). The same result was observed for separation of fungi-fungi species. On the contrary, separation of plant-plant EST sequences was gradually increased with the size of the training set.

The organisms included in the study were selected based on mRNA abundance in GenBank and real world interaction. These interactions are listed in the "interactions" column of table 1. Some combinations of plant-fungi organisms used for sequence classification are not found in nature but interaction may exist for species evolutionary close to the chosen organism. The fungi *Fusarium graminearum*, causal agent of fusarium head blight, attacks a broad range of crop plants including the graminous species barley, maize and wheat (McMullen, Jones et al. 1997). Another destructive Fusarium species is *F. oxysporum*, for which Fusarium wilt has described in ~80 plant species (Recorbet, Steinberg et al. 2003). Although not immediate relatives, *F. graminearum* and *F. oxysporum* both belong to the order *Hypocreales* of the fungi.

Although the simple measure of difference in GC content between genomes was highly correlated with the accuracy of the SVM classifier, dinucleotide bias showed even higher correlation. Both fungi and monocots have a tendency for higher GC content, particular at the third codon position, whereas dicots have a tendency towards higher AT content at the third codon position (Maor, Kosman et al. 2003). This was evident from the average accuracy of dicot/monocot and dicot/monocot pairs, which on average was higher than dicot/dicot pairs.

Classification accuracy of monocots/fungi pairs was also less accurate than for dicot/fungi pairs, but still above an average of 95% for most pairs when using triplet frequencies. The dinucleotide distance provides a good way to make an initial estimate of classification performance for a given pair - at distances above 100, all species pairs were classified with at least 95% average accuracy.

With an accurate classifier publicly available to discriminate between plant and microbe, the combination of EST sequencing and sequence classification provides a new tool for investigating plant-microbe interactions, especially for the study of obligate biotrophic pathogens such as *Blumeria graminis f. sp. hordei*, which infects barley (Both, Csukai et al. 2005). Contrary to normal mRNA preparation from plant or pathogen tissue, where contamination of foreign tissue is unwanted, it would be better to equalize the amount of mRNA from each organism. Standard EST sequencing of such mixed libraries may then offer new insight of genes involved in plant-pathogen interaction beyond the point of initial surface contact and into later stages of disease with deeper penetration of host tissues.

The application of the developed method is not restricted to EST data. The next generation of sequencing technologies, such as 454 Life Sciences (Margulies, Egholm et al. 2005) (see also review of other technologies in (Meyers, Souret et al. 2006)) are likely to produce many short sequence tags, and EST sequencing of pathogen-infected host tissue is likely to shift in the direction of these lower-cost, higher throughput methods. ESTs are far more expensive than microarrays, for example, so the new technologies offer a cheap alternative for arrays in a way that traditional ESTs cannot compete. The 454 technology currently creates reads with median value of 110 base pairs and in preliminary experiments the authors achieved sequencing of 200 base long reads where the proposed methodology can classify sequences with average accuracy of about 87-92% (fig. 5). Thus it may serve as a preliminary data filtering step before or after the assembly step.

Another challenge comes from applications of the new technologies in the context of metagenomics of uncultured microorganisms, in which the read will be sequenced from very complex populations like sea/lake/river water, soil samples (Handelsman 2004). Such data can contain thousand of genomes. This is a rapidly growing field and the ability to distinguish different organisms, before the assembly of sequence reads, may improve genome finishing. While in the current study we only considered binary situations, the current methodology can also be used for separation of mixtures using multi-class classification SVMs.

ACKNOWLEDGEMENTS

The authors are grateful to Ulrich Güldener and Martin Münsterkötter for providing us curated genomic data of *F. gramenarium*. This work was supported by grants 031U212C BFAM (BMFB) to HWM, TE/308/1-1 (DFG) to IVT/HWM and Danish technical research council (grant 26-00-0141 “Exploring the biosynthetic potential of potato”) to JE.

References

Adams, M. D., M. B. Soares, et al. (1993). "Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library." Nat Genet **4**(4): 373-80.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Bennet, J. W. (1997). "Fungal genome initiative. Genomics for filamentous fungi." Genet Biol **21**: 3-7.

Benson, D. A., I. Karsch-Mizrachi, et al. (2005). "GenBank." Nucleic Acids Res **33**(Database issue): D34-8.

Bishop, D. and R. M. Cooper (1993). "An ultrastructural-study of root invasion in three vascular wilt diseases." Physiol. Plant Pathol **22**(15-27).

Both, M., M. Csukai, et al. (2005). "Gene expression profiles of Blumeria graminis indicate dynamic changes to primary metabolism during development of an obligate biotrophic pathogen." Plant Cell **17**(7): 2107-22.

Chang, C. C. and C. J. Lin (2005). LIBSVM: a Library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chawla, N. V., K. W. Bowyer, et al. (2002). "SMOTE: Synthetic minority over-sampling technique." Journal of Artificial Intelligence Research **16**: 321-357.

Dean, R. A., N. J. Talbot, et al. (2005). "The genome sequence of the rice blast fungus Magnaporthe grisea." Nature **434**(7036): 980-6.

Friedel, C. C., K. H. Jahn, et al. (2005). "Support vector machines for separation of mixed plant-pathogen EST collections based on codon usage." Bioinformatics **21**(8): 1383-8.

Galagan, J. E., S. E. Calvo, et al. (2003). "The genome sequence of the filamentous fungus Neurospora crassa." Nature **422**(6934): 859-68.

Gentles, A. J. and S. Karlin (2001). "Genome-scale compositional comparisons in eukaryotes." Genome Res **11**(4): 540-6.

Guldener, U., G. Mannhaupt, et al. (2006). "FGDB: a comprehensive fungal genome resource on the plant pathogen Fusarium graminearum." Nucleic Acids Res **34**(Database issue): D456-8.

Handelsman, J. (2004). "Metagenomics: application of genomics to uncultured microorganisms." Microbiol Mol Biol Rev **68**(4): 669-85.

Hraber, P. T. and J. W. Weller (2001). "On the species of origin: diagnosing the source of symbiotic transcripts." Genome Biol **2**(9): RESEARCH0037.

Huitema, E., T. A. Torto, et al. (2003). "Combined ESTs from plant-microbe interactions: using GC counting to determine the species of origin." Methods Mol Biol **236**: 79-84.

Karlin, S. (1998). "Global dinucleotide signatures and analysis of genomic heterogeneity." Curr Opin Microbiol **1**(5): 598-610.

Maor, R., E. Kosman, et al. (2003). "PF-IND: probability algorithm and software for separation of plant and fungal sequences." Curr Genet **43**(4): 296-302.

Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-80.

McMullen, M., R. Jones, et al. (1997). "Scab of Wheat and Barley: A Re-emerging Disease of Devastating Impact." Plant Disease **81**(12): 1340-1348.

Meyers, B. C., F. F. Souret, et al. (2006). "Sweating the small stuff: microRNA discovery in plants." Curr Opin Biotechnol **17**(2): 1013-1019.

Miller, R. T., A. G. Christoffels, et al. (1999). "A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base." Genome Res **9**(11): 1143-55.

Panabieres, F., J. Amselem, et al. (2005). "Gene identification in the oomycete pathogen Phytophthora parasitica during in vitro vegetative growth through expressed sequence tags." Fungal Genet Biol **42**(7): 611-23.

Pertea, G., X. Huang, et al. (2003). "TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets." Bioinformatics **19**(5): 651-2.

Posada-Buitrago, M. L. and R. D. Frederick (2005). "Expressed sequence tag analysis of the soybean rust pathogen Phakopsora pachyrhizi." Fungal Genet Biol **42**(12): 949-62.

Quackenbush. J., J. Cho, et al (2001) "The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species." Nucleic Acids Res **29**(1): 159-164.

Recorbet, G., C. Steinberg, et al. (2003). "Wanted: pathogenesis-related marker molecules for Fusarium oxysporum." New Phytologist **159**: 73-92.

Rudd, S. (2005). "openSputnik--a database to ESTablish comparative plant genomics using unsaturated sequence collections." Nucleic Acids Res **33**(Database issue): D622-7.

Rudd, S. and I. V. Tetko (2005). "Eclair--a web service for unravelling species origin of sequences sampled from mixed host interfaces." Nucleic Acids Res **33**(Web Server issue): W724-7.

Ruepp, A., A. Zollner, et al. (2004). "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes." Nucleic Acids Res **32**(18): 5539-45.

Seki, M., M. Narusaka, et al. (2002). "Functional annotation of a full-length Arabidopsis cDNA collection." Science **296**(5565): 141-5.

Sesma, A. and A. E. Osbourn (2004). "The rice leaf blast pathogen undergoes developmental processes typical of root-infecting fungi." Nature **431**(7008): 582-6.

Tetko, I. V., P. Bruneau, et al. (2006). "Can we estimate the accuracy of ADME-Tox predictions?" Drug Discov Today **11**(15/16): 700-707.

Tetko, I. V., V. P. Solov'ev, et al. (2006). "Benchmarking of linear and nonlinear approaches for quantitative structure-property relationship studies of metal complexation with ionophores." J Chem Inf Model **46**(2): 808-19.

Viaud, M. C., P. V. Balhadere, et al. (2002). "A Magnaporthe grisea cyclophilin acts as a virulence determinant during plant infection." Plant Cell **14**(4): 917-30.

Yarden, O., D. J. Ebbole, et al. (2003). "Fungal biology and agriculture: revisiting the field." Mol Plant Microbe Interact **16**(10): 859-66.

Figures legends

Fig. 1. Dataflow for SVM analysis of data using triplets frequencies. On the first step we calculated triplet frequencies for all sequences from both compared genomes (e.g., plant and pathogen). An equal numbers of sequences (e.g., N_i=1000) were sampled without replication from each genome, *i=1,2*, (each sequence was selected only once) to form the training set. The remaining sequences formed the independent test set. The five-fold cross-validation was performed to optimize the SVM parameters and to predict the sequences from the cross-validation and the independent test sets.

Fig. 2. SVM performance for separation of plant-fungi pairs using codon and n-mers frequencies. The x-axis indicates fraction of host-pathogen pairs classified with given accuracy, i.e. for mRNA-4 (tetranucleotides frequencies derived from mRNA) 40% and 80% of host-pathogen pairs had classification accuracy < 80% and <90%, respectively. The use of mRNA-3 (triplet nucleotide frequencies) calculated < 96% and < 98% correct predictions for the same 40% and 80% of host-pathogen pairs.

Fig 3. SVM classification accuracy using triplet frequencies as a function of dinucleotide d* bias distance.

Fig. 4. SVM performance as function of the number of EST sequences in the training set for different genome pairs. The performance for the training (black lines) and test (grey lines) sets are shown. Circles, squares and rhombs correspond to separation of plant-pathogen, pathogen-pathogen and plant-plant pairs. If there are more than 100 sequences per genome, the increase in the number of EST sequences increases classification accuracy within each kingdom (plants, fungi) but practically does not change the SVM performance for plant-pathogen separation.

Fig. 5. SVM performance as function of the length of sequences in the training set for different genome pairs. Circles, squares and rhombs correspond to separation of plant-pathogen, pathogen-pathogen and plant-plant pairs.

Table 1. Description of organisms and mRNA sequences used

#	name	TaxId	mRNA	3' utr	5' utr	Unigenes (UNI)	CDS-[a]	interactions
Dicots								
d1	*S. oleracea*	3562	263	249	218	--[b]	--[b]	1, 4
d2	*G. hirsutum*	3635	358	294	240	10250	1183	4, 5
d3	*C. sativus*	3659	229	173	135	2445	352	1, 4, 5
d4	*A. thaliana*	3702	70916	52716	44649	48088	6543	4
d5	*B. napus*	3708	507	401	333	15787	1620	1 4
d6	*G. max*	3847	952	798	642	53890	7845	1 4 5
d7	*M. sativa*	3879	283	235	194	3476	692	4
d8	*M. truncatula*	3880	235	199	235	37738	9821	4
d9	*P. vulgaris*	3885	231	188	130	3584	328	1, 4
d10	*P. sativum*	3888	802	680	570	1175	1175	4
d11	*D. carrota*	4039	238	195	156	--[b]	--[b]	4, 5
d12	*L. esculentum*	4081	1146	958	785	31288	3152	1, 4
d13	*N. tabacum*	4097	1548	1271	1067	19103	4445	1, 4
d14	*S. tuberosum*	4113	724	585	458	35592	10584	1, 4
d15	*H. annuus*	4232	210	158	122	18984	5048	4
Monocots								
m1	*H. vulgare*	4513	275	228	170	49581	20864	3, 4, 6
m2	*O. sativa*	4530	38115	7730	7208	45084	13197	4, 6
m3	*T. aestivum*	4565	1353	1051	818	106313	33292	4, 6
m4	*Z. mays*	4577	1725	1458	1198	56003	21707	3, 4, 6
m5	*M. acuminata*	4641	141	89	54	--[b]	--[b]	4
Pathogens								
1	*P. infestans*	4787	103	86	83	25241	5121	--
2	*N. crassa*	5141	10220	110	100	5279	683	--
3	*U. maydis*	5270	23846	6512	6522	4338	450	--
4	*F.graminearum*	5518	14098	14036	13996	4483	884	--
5	*A. nidulans*	227321	11295	9541	9541	12968	1881	--
6	*M. grisea*	148305	1000	11109	11109	11364	3620	--
7	*P. sojae*	67593	18906	--	--	7579	2957	--
8	*P. ramorum*	164328	15572	--	--	--[b]	--[b]	--
9	*S. sclerotiorum*	5180	14459	14522	14522	--[b]	--[b]	--
10	*S. nodorum*	13684	16597	16597	16597	--[b]	--[b]	--

[a]Number of unigenes without CDS regions (see 2.2).

[b]These organisms had less than 1000 EST sequences in unigenes and thus were not analyzed. The last column lists pathogens genomes for which there are literature evidences of their interactions with the plant.

Table 2. Prediction accuracy of classifiers developed with data from non-coding regions and entire mRNA

sequence region	plant-pathogen	pathogen-pathogen	plant-plant	all pairs of species
mRNA	96.5±0.2	92.7±1	84.6±0.6	90.3±0.4
3'UTR	94.5±0.4	87.6±1	77.2±0.7	85.3±0.6
5'UTR	83.0±1	77.6±3	73.5±1	77.8±0.7

The mean and standard mean errors are indicated.

Figure 1

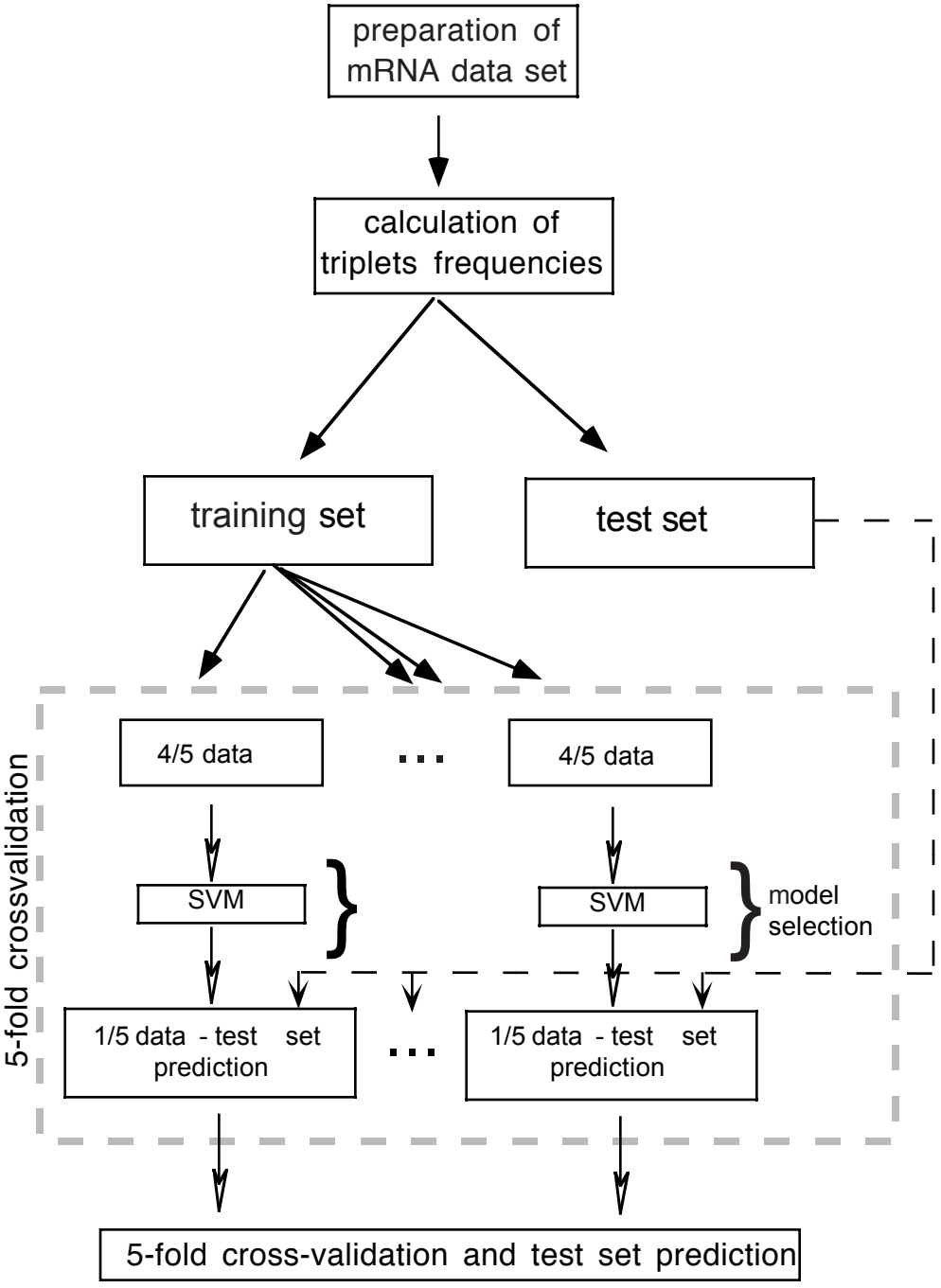

preparation of mRNA data set
calculation of triplets frequencies
training set
test set
4/5 data
4/5 data
SVM
SVM
model selection
1/5 data - test set prediction
1/5 data - test set prediction
5-fold crossvalidation
5-fold cross-validation and test set prediction

Figure 2

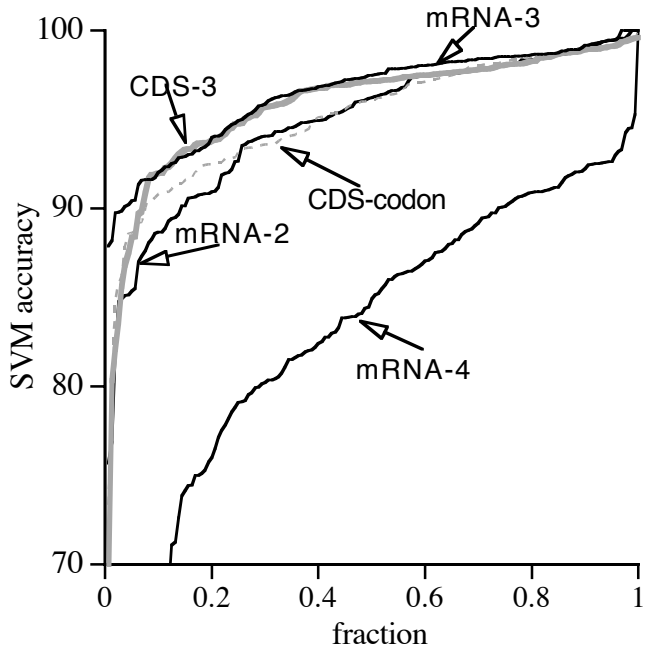

mRNA-3
CDS-3
CDS-codon
mRNA-2
mRNA-4
SVM accuracy
100
90
80
70
0
0.2
0.4
0.6
0.8
1
fraction

Figure 3

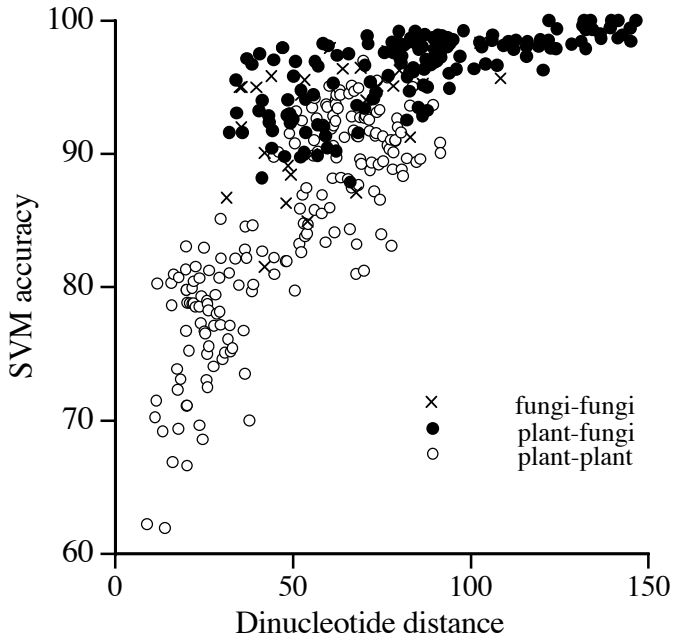

100
90
80
70
60
SVM accuracy
0
50
100
150
Dinucleotide distance
fungi-fungi
plant-fungi
plant-plant

Figure 4

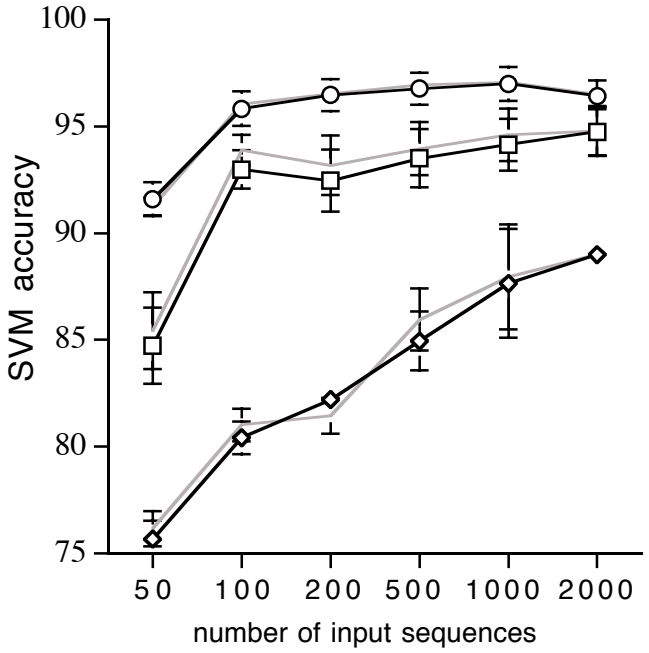
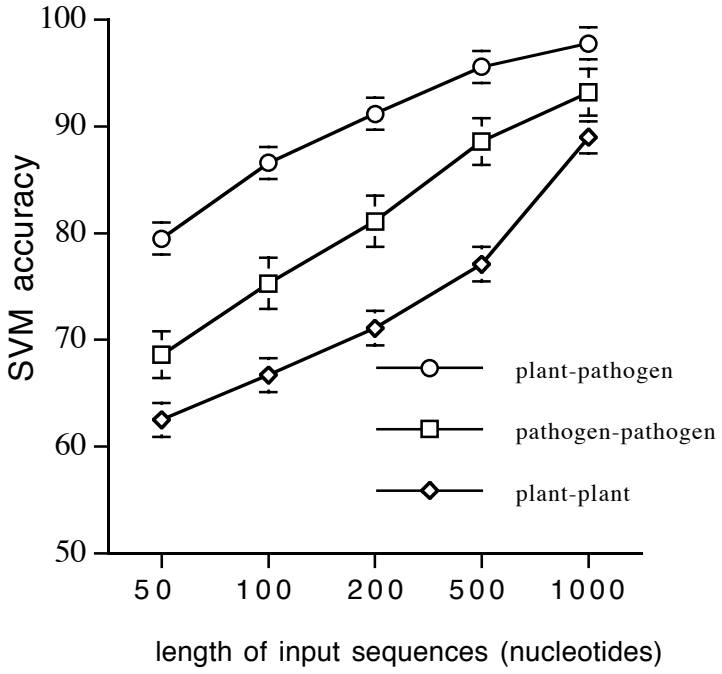

100
95
90
85
80
75
SVM accuracy
50
100
200
500
1000
2000
number of input sequences

Figure 5


100
90
80
70
60
50
SVM accuracy
50
100
200
500
1000
length of input sequences (nucleotides)
plant-pathogen
pathogen-pathogen
plant-plant