Gene Functional Classification from Heterogeneous Data (2001, P.Pavlidis, et.al...)

Presenter: James J. Winkle

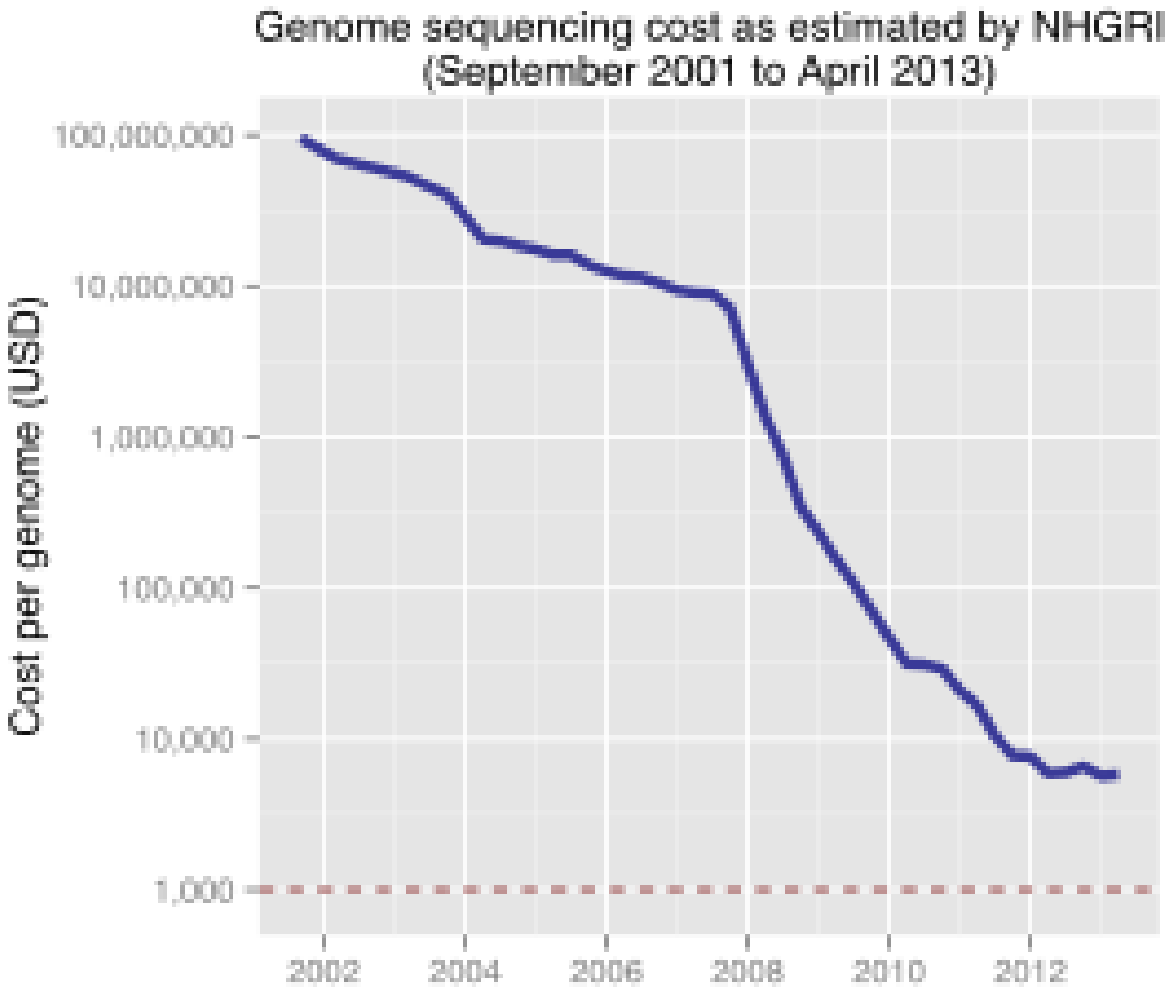
8 May, 2014

Outline

Overview of Paper

- Biology seeks to understand the "molecular machinery of the cell."
- A data-centric complementary view of this machinery is provided by the following types of ("heterogeneous") data:
  - DNA μ -array hybridization experiments
  - Genomic Sequences: Phylogenetic Profiles
- This paper hopes to advance computational techniques toward a long-term goal of learning about gene-function from many different types of genomic data.
- Various Kernel combinations are tested to address how best to combine heterogeneous data for genomic classification

Genomic Sequencing Cost



Illumina Stock Price (ILMN)

Heterogeneous Data Sources

The paper cites previous work from:

- Brown *et al.*(2000): applied SVM techniques to yeast expression data with "excellent classification performance"
- Combining heterogeneous data sets is mentioned (Marcotte, Pellegrini, . . . 1999) but with data sets considered separately rather than at once.

This paper asserts that: "the performance of SVM's when data types are combined and a single hypothesis is formed is superior to combining two independent hypotheses."

Data Type Definitions

DNA μ-expression data

- "The first data set derives from a collection of **DNA μ-array hybridization** experiments. Each data point represents the logarithm of the ratio of expression levels of a particular gene under two different experimental conditions."

$$X_i = \frac{log(E_i/R_i)}{[\Sigma_{j=1}^{79} log^2(E_j/R_j)]^{\frac{1}{2}}}$$

- A snapshot of the messenger RNA expression levels during various time points of "cell events" (diauxic shift, cell division, sporulation, "shocks")
- If two genes have a functional link, they should be expressed together during the functional event

Data Type Definitions

Phylogenetic Profiles

- Two genes with similar phylogenetic profiles are *likely* to have similar functions, under the assumption that their similar pattern of inheritance across species is the result of a functional link.
- "In its simplest form, a **phylogenetic profile** is a bit string, in which the Boolean value of each bit indicates a close homolog in the genome." In this paper, each genome position in the data vector is $-logE_{val}$ from BLAST in a search against the complete genome (negative values truncated to 0).

Assigning protein functions by comparative genome analysis protein phylogenetic profiles

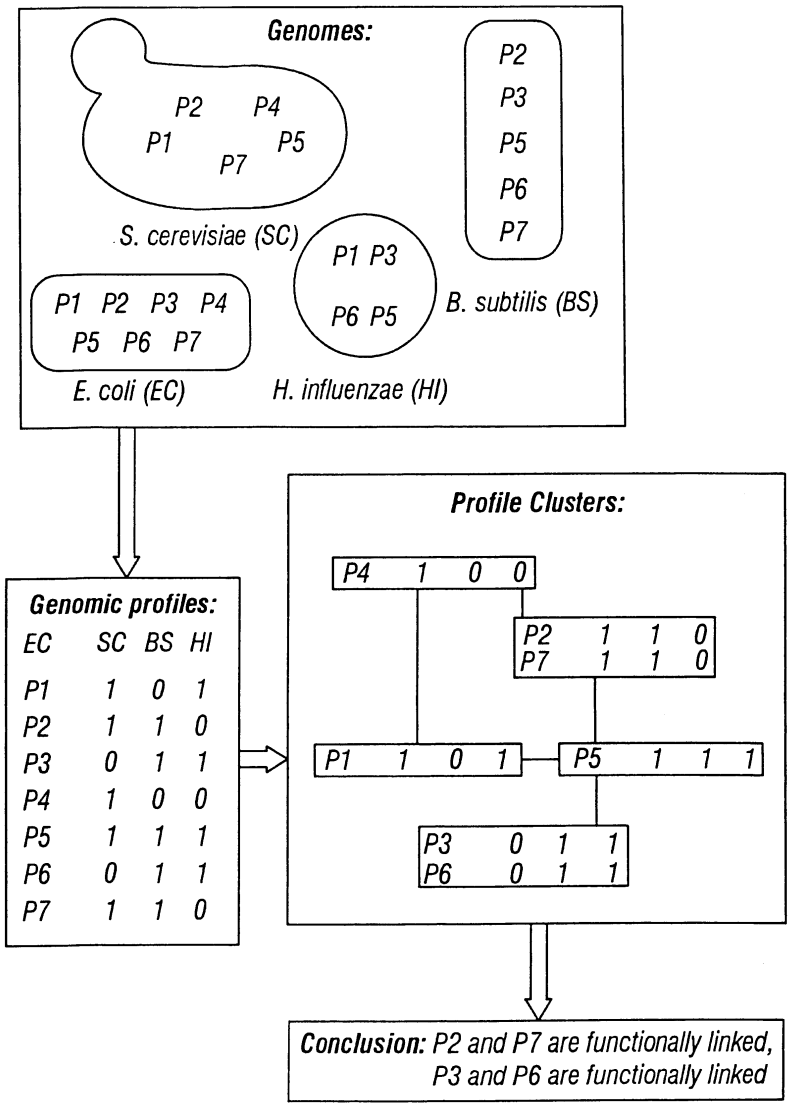


FIG. 3

US 6564151 B1

Inventors: M. Pellegrini, . . .

Heterogeneous Data Sources

- Classification comes from the CYGD (MIPS Comprehensive Yeast Genome Database), which contains "several hundred functional classes"
- Classes containing 10 or more genes are selected
- 108 Classes are initially selected (but later narrowed to 27 "learnable classes")
- The two genomic data vectors are of length:

$$\mathbf{x_g} = [\mu\text{-array expression data}] \qquad n = 79$$
$$\mathbf{x_p} = [\text{phlyo}] \qquad n = 24$$

- There are N = 2465 yeast genes used as the data set (selected for "accurate functional annotations")

Polynomial Kernel

- The Kernel function selected is polynomial degree 3 (data vector is projected on the unit sphere)

$$K(\mathbf{x}, \mathbf{y}) = \left(\frac{< \mathbf{x}, \mathbf{y} >}{\|\mathbf{x}\| \|\mathbf{y}\|} + 1 \right)^3$$

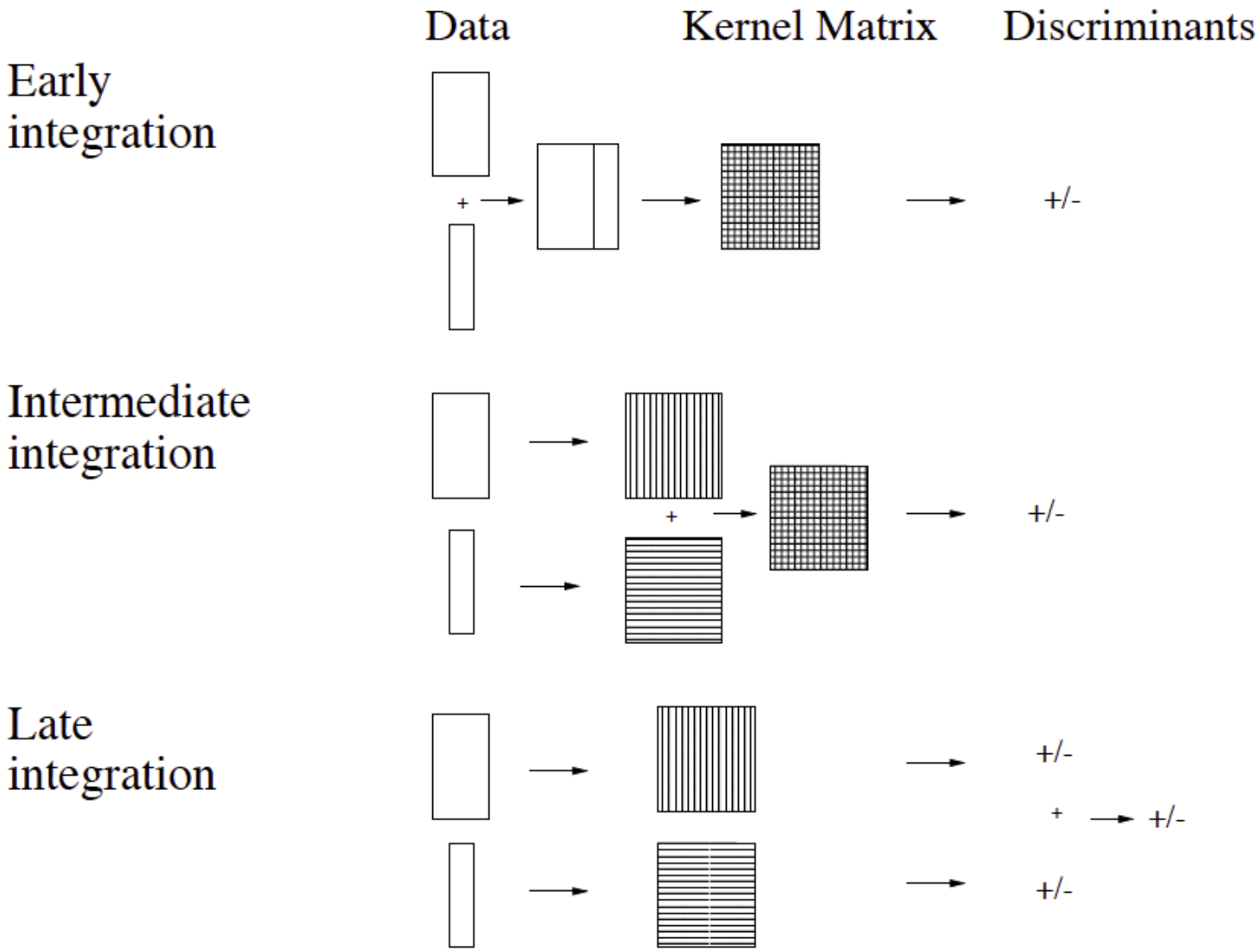
- The polynomial "takes into account pairwise and tertiary correlations..."
- For a fixed vector $\mathbf{x}$, the *level sets* of K are radial in $\mathbf{y}$
- This would clearly present difficulty for radially symmetric (about $\mathbf{0}$) classes

SVM

- Each class is trained with one-against-others (binary) SVMs
- The two types of data are integrated in 3 different ways:
  1. Early: concatenate the vectors
  2. **Intermediate**: add kernel values for each separately
  3. Late: one SVM for each type of data
- The Intermediate integration can be expressed as a new ("Heterogeneous") Kernel:

$$K(\cdot,\cdot) = K(\mathbf{x}_g, \mathbf{y}_g) + K(\mathbf{x}_p, \mathbf{y}_p)$$

Integration



Restricting Correlations

Argument for the use of the Intermediate Integration:

- The heterogeneous kernel creates *local* features of polynomial relationships of one type of data only
- The local features are combined linearly
- Thus, polynomial relationships between different *types* of data are ignored
- Removal of these correlations reduces overfitting

Validation

- For most of the experiments, 3-fold cross-validation is used
- A method-cost is used to evaluate the performance of a method M (early, intermediate, late)

$$C(M) = (f_p(M) + 2 \cdot f_n(M))/n$$

- False negatives f_n are given more weight than false positives f_p (n = number in class)
- Failing to recognize a limited class member is worse than recognizing a non-member
- The method-cost is normalized to [0,1], with 1 being a perfect classifier, as follows: $S(M) = (C(N) - C(M))/2$

Example Normalization Calculation

- For a method that classifies perfectly, we have $C(M) = 0$ ($f_p = f_n = 0$)
- The cost of classifying all data as negative is:

$$C(N) = (0 + 2 \cdot n)/n = 2$$

- Thus, a perfect classifier is normalized to $(2 - 0)/2 = 1$
- ...and a null classifier is normalized to $(2 - 2)/2 = 0$

**the formula shown in the paper does not work as written and is fixed here (it varies from the previously published Brown paper also).*

Results

Class	Exp	Phylo	Early	Intermediate	Late
amino acid transporters	0.05 ± 0.04	$\mathbf{0.77 \pm 0.10}$	0.50 ± 0.04	$\mathbf{0.71 \pm 0.08}$	0.49 ± 0.07
ribosomal proteins	0.71 ± 0.02	0.09 ± 0.03	$\mathbf{0.76 \pm 0.01}$	0.71 ± 0.01	0.69 ± 0.01
sugar and carbohydrate transporters	0.33 ± 0.07	$\mathbf{0.67 \pm 0.02}$	$\mathbf{0.68 \pm 0.06}$	$\mathbf{0.70 \pm 0.01}$	0.63 ± 0.03
glycolysis and gluconeogenesis	0.21 ± 0.03	$\mathbf{0.43 \pm 0.05}$	0.28 ± 0.02	$\mathbf{0.39 \pm 0.05}$	$\mathbf{0.39 \pm 0.04}$
mitochondrial organization	$\mathbf{0.40 \pm 0.03}$	0.15 ± 0.01	$\mathbf{0.43 \pm 0.03}$	$\mathbf{0.42 \pm 0.02}$	0.35 ± 0.02
tricarboxylic acid pathway	0.21 ± 0.11	0.15 ± 0.07	$\mathbf{0.32 \pm 0.08}$	$\mathbf{0.42 \pm 0.07}$	$\mathbf{0.25 \pm 0.13}$
deoxyribonucleotide metabolism	0.07 ± 0.05	$\mathbf{0.31 \pm 0.11}$	$\mathbf{0.24 \pm 0.15}$	$\mathbf{0.39 \pm 0.11}$	$\mathbf{0.31 \pm 0.12}$
organization of cytoplasm	0.35 ± 0.01	0.18 ± 0.01	$\mathbf{0.38 \pm 0.01}$	0.34 ± 0.02	$\mathbf{0.35 \pm 0.02}$
transport ATPases	0.13 ± 0.04	$\mathbf{0.37 \pm 0.05}$	0.23 ± 0.05	$\mathbf{0.32 \pm 0.04}$	0.22 ± 0.03
amino acid biosynthesis	0.18 ± 0.02	0.28 ± 0.02	0.29 ± 0.03	$\mathbf{0.36 \pm 0.04}$	0.27 ± 0.02
purine ribonucleotide metabolism	0.17 ± 0.03	$\mathbf{0.26 \pm 0.05}$	0.20 ± 0.04	$\mathbf{0.33 \pm 0.04}$	0.19 ± 0.03
pyrimidine ribonucleotide metabolism	0.03 ± 0.02	$\mathbf{0.33 \pm 0.06}$	0.11 ± 0.04	$\mathbf{0.28 \pm 0.03}$	0.17 ± 0.03
cytoplasmic degradation	$\mathbf{0.32 \pm 0.01}$		$\mathbf{0.32 \pm 0.06}$	$\mathbf{0.30 \pm 0.03}$	0.17 ± 0.02
respiration	$\mathbf{0.32 \pm 0.02}$		$\mathbf{0.30 \pm 0.04}$	0.23 ± 0.04	0.17 ± 0.03
organization of chromosome structure	$\mathbf{0.31 \pm 0.01}$		$\mathbf{0.30 \pm 0.01}$	$\mathbf{0.29 \pm 0.02}$	0.13 ± 0.03
phosphate utilization	$\mathbf{0.22 \pm 0.04}$	0.08 ± 0.05	$\mathbf{0.26 \pm 0.05}$	$\mathbf{0.21 \pm 0.04}$	$\mathbf{0.22 \pm 0.04}$
organization of plasma membrane	0.07 ± 0.02	$\mathbf{0.25 \pm 0.01}$	$\mathbf{0.24 \pm 0.03}$	$\mathbf{0.26 \pm 0.03}$	$\mathbf{0.26 \pm 0.02}$
pentose phosphate pathway		$\mathbf{0.20 \pm 0.15}$		$\mathbf{0.26 \pm 0.07}$	$\mathbf{0.15 \pm 0.10}$
cellular import	0.04 ± 0.02	$\mathbf{0.25 \pm 0.04}$	$\mathbf{0.18 \pm 0.05}$	0.17 ± 0.03	$\mathbf{0.21 \pm 0.04}$
protein folding and stabilization		$\mathbf{0.24 \pm 0.04}$	$\mathbf{0.20 \pm 0.04}$	$\mathbf{0.23 \pm 0.05}$	0.14 ± 0.04
proteolysis	$\mathbf{0.23 \pm 0.02}$		$\mathbf{0.24 \pm 0.02}$	$\mathbf{0.18 \pm 0.06}$	0.17 ± 0.01
pheromone response generation	$\mathbf{0.24 \pm 0.05}$		0.15 ± 0.03	$\mathbf{0.14 \pm 0.08}$	
nuclear organization	$\mathbf{0.21 \pm 0.01}$	0.07 ± 0.01	$\mathbf{0.24 \pm 0.03}$	$\mathbf{0.24 \pm 0.02}$	0.17 ± 0.02
drug transporters		$\mathbf{0.23 \pm 0.09}$			
organization of endoplasmatic reticulum	$\mathbf{0.20 \pm 0.02}$		$\mathbf{0.22 \pm 0.03}$	$\mathbf{0.19 \pm 0.05}$	0.13 ± 0.03
organization of cell wall	$\mathbf{0.12 \pm 0.04}$	$\mathbf{0.19 \pm 0.06}$	$\mathbf{0.14 \pm 0.08}$	$\mathbf{0.16 \pm 0.07}$	$\mathbf{0.21 \pm 0.08}$
anion transporters		$\mathbf{0.21 \pm 0.02}$			
Mean cost savings	0.19 ± 0.02	0.21 ± 0.04	0.27 ± 0.03	0.31 ± 0.03	0.24 ± 0.03
Number of best-performing	10	12	17	21	8
Number of non-learnable	4	6	3	2	3

Results

Exp	Phylo	Early	Intermediate	Late
0.05 ± 0.04	$\mathbf{0.77 \pm 0.10}$	0.50 ± 0.04	$\mathbf{0.71 \pm 0.08}$	0.49 ± 0.07
0.71 ± 0.02	0.09 ± 0.03	$\mathbf{0.76 \pm 0.01}$	0.71 ± 0.01	0.69 ± 0.01
0.33 ± 0.07	$\mathbf{0.67 \pm 0.02}$	$\mathbf{0.68 \pm 0.06}$	$\mathbf{0.70 \pm 0.01}$	0.63 ± 0.03
0.21 ± 0.03	$\mathbf{0.43 \pm 0.05}$	0.28 ± 0.02	$\mathbf{0.39 \pm 0.05}$	$\mathbf{0.39 \pm 0.04}$
$\mathbf{0.40 \pm 0.03}$	0.15 ± 0.01	$\mathbf{0.43 \pm 0.03}$	$\mathbf{0.42 \pm 0.02}$	0.35 ± 0.02
0.21 ± 0.11	0.15 ± 0.07	$\mathbf{0.32 \pm 0.08}$	$\mathbf{0.42 \pm 0.07}$	$\mathbf{0.25 \pm 0.13}$
0.07 ± 0.05	$\mathbf{0.31 \pm 0.11}$	$\mathbf{0.24 \pm 0.15}$	$\mathbf{0.39 \pm 0.11}$	$\mathbf{0.31 \pm 0.12}$
0.35 ± 0.01	0.18 ± 0.01	$\mathbf{0.38 \pm 0.01}$	0.34 ± 0.02	$\mathbf{0.35 \pm 0.02}$
0.13 ± 0.04	$\mathbf{0.37 \pm 0.05}$	0.23 ± 0.05	$\mathbf{0.32 \pm 0.04}$	0.22 ± 0.03
0.18 ± 0.02	0.28 ± 0.02	0.29 ± 0.03	$\mathbf{0.36 \pm 0.04}$	0.27 ± 0.02
0.17 ± 0.03	$\mathbf{0.26 \pm 0.05}$	0.20 ± 0.04	$\mathbf{0.33 \pm 0.04}$	0.19 ± 0.03
0.03 ± 0.02	$\mathbf{0.33 \pm 0.06}$	0.11 ± 0.04	$\mathbf{0.28 \pm 0.03}$	0.17 ± 0.03
$\mathbf{0.32 \pm 0.01}$		$\mathbf{0.32 \pm 0.06}$	$\mathbf{0.30 \pm 0.03}$	0.17 ± 0.02
$\mathbf{0.32 \pm 0.02}$		$\mathbf{0.30 \pm 0.04}$	0.23 ± 0.04	0.17 ± 0.03
$\mathbf{0.31 \pm 0.01}$		$\mathbf{0.30 \pm 0.01}$	$\mathbf{0.29 \pm 0.02}$	0.13 ± 0.03

Results

0.22 ± 0.04	0.08 ± 0.05	**0.26 ± 0.05**	**0.21 ± 0.04**	**0.22 ± 0.04**
0.07 ± 0.02	**0.25 ± 0.01**	**0.24 ± 0.03**	**0.26 ± 0.03**	**0.26 ± 0.02**
	0.20 ± 0.15		**0.26 ± 0.07**	**0.15 ± 0.10**
0.04 ± 0.02	**0.25 ± 0.04**	**0.18 ± 0.05**	0.17 ± 0.03	**0.21 ± 0.04**
	0.24 ± 0.04	**0.20 ± 0.04**	**0.23 ± 0.05**	0.14 ± 0.04
0.23 ± 0.02		**0.24 ± 0.02**	**0.18 ± 0.06**	0.17 ± 0.01
0.24 ± 0.05		0.15 ± 0.03	**0.14 ± 0.08**	
0.21 ± 0.01	0.07 ± 0.01	**0.24 ± 0.03**	**0.24 ± 0.02**	0.17 ± 0.02
	0.23 ± 0.09			
0.20 ± 0.02		**0.22 ± 0.03**	**0.19 ± 0.05**	0.13 ± 0.03
0.12 ± 0.04	**0.19 ± 0.06**	**0.14 ± 0.08**	**0.16 ± 0.07**	**0.21 ± 0.08**
	0.21 ± 0.02			
0.19 ± 0.02	0.21 ± 0.04	0.27 ± 0.03	0.31 ± 0.03	0.24 ± 0.03
10	12	17	21	8
4	6	3	2	3

5 Most Learnable Classes

Class	Size	FP	FN
amino acid transporters	22	2.0 ± 0.4	5.6 ± 0.2
ribosomal proteins	173	26.6 ± 1.2	34.2 ± 1.1
sugar and carbohydrate transporters	32	2.4 ± 0.7	9.0 ± 0.0
deoxyribonucleotide metabolism	9	0.2 ± 0.2	4.6 ± 0.7
mitochondrial organization	296	84.8 ± 1.8	128.4 ± 1.7

Conclusions

- SVM's have extended to other data types in this domain (phylogenetic profiles)
- The results of intermediate integration do not show *radical* improvement
- But *some* improvement can be worth a lot
- No analysis of other kernels was made (but claimed no expectation of helping one method over another)
- There is no claim that gene functional ID wants to be perfect (cf. digit recognition); the domain here is to be better (via SVMs).