Supplementary Table 2

**Supplementary Table 2:** Classification accuracy of the SVM with a gaussian versus a linear kernel. Values for the gaussian kernel are shown in columns 3-4, and for the linear kernel in columns 5-6. Performance was evaluated for some parameter combinations at different taxonomic levels. The composition space is defined by the word length *w*, the number of literal characters *l,* and the step size *s*. For the tests, the all-vs.-all multi-class SVM classifier was trained on coding sequences in 3-fold crossvalidation on 2/3 of the sequenced organisms. Accuracy was evaluated on coding sequences from organisms that were excluded from training (ie unknown organisms to the classifier). The overall classification success is measured by the sensitivity (or micro-accuracy) in the different tests; the normalized specificity denotes the average proportion of correct assignments for every clade. Note that the specificity here is low as post-processing with the OVA classifier is not performed. The last column shows the difference in micro-accuracy for the gaussian versus the linear kernel, which for feature-spaces with more than $w^2$ dimensions is generally positive.

| Level | | Sn.$_{gaussian}$ (%) | Sp.$_{gaussian}$ (%) | Sn.$_{linear}$ (%) | Sp.linear (%) | Δ(Sn.$_{gaussian}$ - Sn.$_{linear}$) |
|---|---|---|---|---|---|---|
| Genus | w2,l2,s1 | 38.1 | 28.7 | 41.1 | 23.2 | -3 |
| Genus | w4,l4,s3 | 75.3 | 42.4 | 63.4 | 39.9 | 11.9 |
| Genus | w6,l4,s3 | 80.9 | 47.1 | 67.6 | 48.5 | 13.3 |
| Class | w2,l2,s1 | 36.2 | 35.7 | 33.7 | 29.2 | 2.5 |
| Class | w2,l2,s3 | 27.5 | 26.2 | 23.7 | 23 | 3.8 |
| Class | w3,l2,s3 | 50 | 40.6 | 51.2 | 42 | -1.2 |
| Class | w4,l4,s3 | 67.9 | 54.5 | 53 | 53 | 14.9 |
| Class | w6,l4,s3 | 72.7 | 59.2 | 59.5 | 58 | 13.2 |
| Class | w6,l5,s3 | 71.9 | 58.5 | 62.4 | 56.1 | 9.5 |
| Class | w6,l6,s1 | 67.7 | 54.2 | 58.5 | 52.5 | 9.2 |
| Class | w6,l6,s3 | 70.2 | 57.5 | 61 | 54.9 | 9.2 |
| Phylum | w3,l3,s3 | 58.4 | 50.8 | 50.5 | 38.1 | 7.9 |