Supplementary Material

Supplementary materials and methods

Algorithm

Input data variants: Raw oligonucleotide frequencies and z-scores

All input DNA sequences were extended by their reverse complements in order to compensate for asymmetries between leading and lagging strand oligonucleotide compositions. For training TaxSOM, either raw oligonucleotide frequencies or Markov model-derived z-scores were used as input data. In the case of raw frequencies, all possible oligonucleotides of a given length were counted and the counts subsequently normalized on sequence length. When z-scores were used as input data, the normalized raw counts were z-transformed (Teeling et al., 2004). A z-transformation is the normalization of the differences between the observed and expected counts on the standard deviation (Eqn: 1). For example, if the observed frequency *N* of an oligonucleotide of length four (tetranucleotide) within a sequence is denoted as *N(n1n2n3n4)*, the corresponding expected frequency is denoted as *E(n1n2n3n4)* and the variance as *var(N(n1n2n3n4))*, then the corresponding z-score *Z(n1n2n3n4)* is calculated as follows:

$$Z(n1n2n3n4) = \frac{N(n1n2n3n4) - E(n1n2n3n4)}{\sqrt{\mathrm{var}(N(n1n2n3n4))}} \quad \text{(Eqn. 1)}$$

The expected frequency within Eqn. 1 can be computed via the maximal-order Markov model given in Eqn. 2, and the variance via the Markov-model approach in Eqn. 3 (Schbath et al., 1995).

$$E(n1n2n3n4) = \frac{N(n1n2n3) * N(n2n3n4)}{N(n2n3)} \quad \text{(Eqn. 2)}$$

$$\mathrm{var}(N(n1n2n3n4)) = E(n1n2n3n4) * \frac{\left[N(n2n3) - N(n1n2n3)\right] * \left[N(n2n3) - N(n2n3n4)\right]}{N(n2n3)^2} \quad \text{(Eqn. 3)}$$

Irrespective of whether raw counts or z-scores were used as inputs, in both cases input DNA sequences were transformed into numerical vectors (one for each DNA sequence), with a size determined by the number of possible oligonucleotide permutations. For a given oligonucleotide length *n*, the number of possible permutations is 4^n (3 = 64; 4 = 256; 5 = 1024; etc.).

SOM algorithm variants: Growing and batch-learning Self-Organizing Maps

TaxSOM implements two variants of the SOM algorithm, the batch-learning and the growing SOM. TaxSOM's batch-learning implementation is a variation of the one described by Abe (Abe et al., 2003), while the growing SOM implementation follows the approach described by Chan (Chan et al., 2007).

Batch-learning SOMs (BLSOMs)

In the batch-learning approach, all input vectors are presented to the SOM at once. Hence, the batch-learning SOM is independent of the temporal order of the input vectors. This has the advantage that SOMs can be initialized with pre-ordered input data instead of random input data, which causes the algorithm's iterative phases to need fewer cycles for convergence (Kohonen et al., 2001). In our case, a Principal Components Analysis (PCA) was applied to the input vectors prior to SOM initialization. PCA transforms the data to a new coordinate system in a way that the highest variance is represented by the first coordinate (first principal component, PC1), the second highest variance on the second coordinate (second principal component, PC2), etc.. This is achieved by rotating the vector space with the eigenvectors (the principal components) of the covariance matrix as a new basis. The subsequent batch-learning SOM algorithm works as follows:

Step 1: Initialization

The first step in the initialization of a batch-learning SOM is to determine its lattice size, i.e. the number of nodes in both dimensions that are required to represent the variation of the input data. These numbers can be estimated from the standard deviations of the first two principal components of the input data (Abe et al., 2003). The number of nodes in the first dimension (i) was set to five times the spread of the standard deviation of PC1 ($5\sigma_1$), and for the second dimension (j) it was approximated as the nearest integer exceeding $\sigma_2/\sigma_1 \times I$. After determining the SOM's lattice dimensions, all nodes were initialized with numerical data. These initial weight vectors were computed as follows:

$$w_{ij} = x_{av} + \left[b_1\left(i - \frac{I}{2} \right) + b_2\left(j - \frac{J}{2} \right) \right] \quad \text{(Eqn. 4)}$$

In Eqn. 4, w_{ij} denotes the weights of the node at position *ij,* x_{av} denotes the average of all vectors of the input data, and b_1 and b_2 denote the eigenvectors of the first and second principal components.

Step 2: Iterative association

For each input vector, the node with the minimum distance was determined, i.e. the node with the weight vector $w_{i'j'}$ most similar to the input vector x_k. In TaxSOM, the Euclidean distance is used for this purpose, but in principle other distance measures can be used as well.

Step 3: Iterative learning

After all input vectors have been associated with nodes, the learning phase starts, where the node's weight vectors $w_{i'j'}$ are iteratively adjusted to better reflect the input data. These adjustments were made in the following way:

$$w_{ij(new)} = w_{ij} + \alpha(r)\left(\frac{\sum_{x_k \in S_{ij}} x_k}{N_{ij}} - w_{ij} \right) \quad \text{(Eqn: 5)}$$

The factor α denotes the learning rate. It decreases during the algorithm's iterations r according to $\alpha(r) = \max\{0.01, \alpha(1)(1 - r/T)\}$, where T is the start value of α. Elements of S_{ij} are all input vectors x_k associated with the node $w_{i'j'}$ plus those in its neighborhood. The neighborhood is defined as the area around $w_{i'j'}$ satisfying the conditions $i - \beta(r) \le i' \le i + \beta(r)$ and $j - \beta(r) \le j' \le j + \beta(r)$. The parameter β determines the size of the neighborhood and deceases over iterations according to $\beta(r) = \max\{0, \beta(1) - r\}$. N_{ij} represents the number of elements in S_{ij}.

Growing SOMs (GSOMs)

In the batch-learning SOM algorithm the lattice size has to be estimated from the level of variation within the input data. This can lead to non-optimal map sizes, which affects the

separation quality of the data. The growing Self-Organizing Map algorithm addresses this issue. It starts out with only a small number of nodes but more nodes are iteratively added when this is needed to better reflect the variation of the input data (Alahakoon et al., 2000). The growth of the map is controlled by the Growth Threshold (GT) parameter that is defined as $GT = -D \times \ln(SF)$. D is the dimensionality of the input data and SF is a user-defined spread factor with a value ranging between 0 and 1. A SF of zero causes minimal growth, while a SF of one causes maximal growth. The GSOM algorithm consists of three phases, the initialization phase, the growing phase and the smoothing phase. During growth, for each input vector the node with the minimal Euclidean distance is found (winning node). Since this distance represents the precision with which an input data vector is represented by the map, it can be interpreted as an error value. Within the iterations of a GSOM, this value is accumulated every time the same node wins according to:

$$E_{winner}(t+1) = E_i(t) + \sqrt{\sum_{k=1}^{Dim} (v_k - w_{winner,k})^2} \qquad \text{(Eqn\_ 6)}$$

If the winning node's position is at the boundary of the lattice and its error value exceeds GT, than this node grows new nodes at every possible free position around it. The new nodes are initialized with weight vectors that are similar to the weights of their neighboring nodes, so that they integrate well in the map (Alahakoon et al., 2000). If the winning node's position is not at the boundary of the lattice and its error exceeds GT, this node's error is distributed to the surrounding nodes. This provides non-boundary nodes with the ability to indirectly initiate node growth. The GSOM algorithm works as follows:

Step 1: Initialization

The weight vectors of the starting nodes are initialized with random values. Depending on whether a rectangular or a hexagonal topology is used, the initial lattice consists of four or seven nodes, respectively. The final size of the GSOM is controlled by the growth factor that is a function of the spread factor and the dimensionality of the input data.

Step 2: Growing

An element of the dataset is presented to the network. By calculating the Euclidean distances between the presented input vector and all nodes' weight vectors, the node containing a weight vector with minimal distance is determined; this node is considered

as the winner. Subsequently, the weight vectors of the winning node and those inside its neighborhood are adapted, and the error value of the winning node is increased. When the error of a node exceeds the growth threshold (GT) and it is a boundary node, new nodes are grown at every free position around it. If a non-boundary node reaches GT, the error is distributed to neighboring nodes. In case of growth, the new nodes' weight vectors are initialized to match the neighboring nodes weights. Finally, the learning rate (LR) is reset to its initial value. All steps described above are repeated until all elements of the dataset have been presented to the network and node growth is only minimal.

Step 3: Smoothing

In the smoothing phase, the learning rate (LR) is lowered and the starting neighborhood is set to a small size. Again, input data is presented to the network and winning nodes are updated the same way as in growing phase.

Supplementary figure legends

Suppl. Figure 1

Supplement to Figure 2: F-measure values for the GSOM-based classification of the simulated metagenome datasets mimicking habitats of varying complexities that were assembled with the two programs PHRAP and Arachne, respectively. (a) F-values for contigs of 8 kb or larger from the low (simLC) and medium (simMC) complexity datasets, and (b) F-values for all datasets including the high complexity dataset (simHC) without constraints in contig length. Different taxonomic levels are shown in different colors.

Suppl. Figure 2

BLSOM-based classification statistics of simulated datasets

Taxonomic classification accuracy of TaxSOM for the simulated metagenome datasets mimicking habitats of varying complexities that were assembled with the two programs PHRAP and Arachne, respectively. (a) Values for contigs of 8 kb or larger from the low (simLC) and medium (simMC) complexity datasets, and (b) values for all datasets including the high complexity dataset (simHC) without constraints in contig length. Different taxonomic levels are shown in different colors. All classifications were perfomed on a BLSOM trained with z-transformed tetranucleotide counts.

Suppl. Figure 3

Cross-evaluation of TaxSOM with the protein-based taxonomic classification tools (CARMA and Darkhorse) for the MIMAS metagenome dataset of April 14th. Both tools were applied to only those contigs ≥ 2.5 kb where both tools provided classification results (1 896 contigs in total).

Supplementary tables

Suppl. Table 1a

Classification statistics for known organisms with raw count-based GSOMs (classification specificities [%] - *sp.*, sensitivities [%] - *sn.* and F-measure values [%] - *fm.*) of leave-out-datasets with randomly selected parts of fully sequenced microbial species). Statistics for fragments of 0.5 kb to 50 kb sizes and motif lengths from di- to tetranucleotides are shown. Fragments were classified on GSOMs trained with oligonucleotide raw counts. Dinucleotide GSOMs were trained using DNA-sequences split into 10 kb fragments, while tri-, and tetranucleotide GSOMs were trained with 50 kb fragments.

	Superkingdom			Phylum			Class			Order			Family			Genus			Species		
Size	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*
Tetranucleotide raw counts																					
0.5 kb	95	98	97	61	85	71	44	74	55	32	58	41	25	48	33	19	31	23	6	5	5
1.0 kb	97	98	97	72	86	78	59	77	67	49	65	56	42	56	48	33	39	36	12	8	10
2.5 kb	98	98	98	86	89	87	79	83	81	74	74	74	68	67	68	60	51	55	30	14	19
5.0 kb	99	99	99	93	92	92	89	87	88	86	80	83	83	75	79	77	59	67	48	20	28
10 kb	99	99	99	96	94	95	95	90	92	93	85	89	92	81	86	88	66	76	65	26	37
25 kb	100	99	100	98	96	97	97	93	95	97	89	93	96	86	91	95	73	83	78	33	46
30 kb	100	99	100	98	96	97	98	94	96	97	90	93	96	87	91	95	74	83	79	33	47
50 kb	100	99	100	99	97	98	98	94	96	97	91	94	97	88	92	96	77	85	83	38	52
Trinucleotide raw counts																					
0.5 kb	94	99	96	60	87	71	43	77	55	30	60	40	24	50	32	17	32	23	6	6	6
1.0 kb	96	99	97	70	88	78	57	80	66	45	66	54	38	58	46	30	40	34	12	10	11
2.5 kb	98	99	98	83	90	87	75	85	80	68	75	71	62	69	65	53	52	53	27	17	21
5.0 kb	99	99	99	91	93	92	86	88	87	82	81	82	78	76	77	71	61	66	44	23	30
10 kb	99	99	99	95	94	95	93	91	92	91	86	88	88	81	85	84	68	75	60	30	40
25 kb	100	99	100	98	96	97	97	93	95	96	90	93	95	86	90	93	74	83	73	36	48
30 kb	100	100	100	98	96	97	97	94	95	96	90	93	95	87	91	93	76	83	76	38	50
50 kb	100	100	100	98	96	97	97	94	96	97	91	94	96	88	92	94	77	85	79	41	54
Dinucleotide raw counts																					
0.5 kb	93	99	96	55	84	66	37	71	49	25	53	34	19	41	26	14	28	19	5	7	6
1.0 kb	95	99	97	63	85	72	48	74	58	36	59	45	29	49	36	23	35	28	9	11	10
2.5 kb	97	99	98	75	87	80	64	79	71	55	67	60	47	59	52	40	45	42	19	18	18
5.0 kb	98	99	98	83	89	86	76	83	79	69	73	71	62	66	64	56	53	54	31	24	27
10 kb	99	99	99	90	91	90	85	86	85	80	78	79	75	72	74	70	60	65	46	31	37
25 kb	99	99	99	94	93	94	92	89	90	89	83	86	87	78	82	84	68	75	62	39	48
30 kb	99	99	99	95	93	94	92	89	91	91	83	86	88	78	83	86	68	76	65	40	50
50 kb	100	99	99	96	93	94	94	90	92	92	84	88	90	80	85	88	70	78	68	43	53

Suppl. Table 1b

Classification statistics for known organisms with z-score-based GSOMs - classification accuracy of leave-out-datasets fragmented as described for Suppl. Table 1a, showing classification specificities [%] - *sp.*, sensitivities [%] - *sn.*, and F-measure values [%] - *fm.* of a growing SOMs trained with tri- and tetranucleotide normalized z-scores.

	Superkingdom			Phylum			Class			Order			Family			Genus			Species		
Size	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*
Tetranucleotide z-scores																					
0.5 kb	89	100	94	43	98	59	24	95	39	14	91	25	11	84	19	7	51	12	2	8	4
1.0 kb	92	100	96	51	98	67	35	96	52	24	93	38	19	87	31	13	63	21	5	14	7
2.5 kb	96	100	98	71	99	82	59	97	74	48	94	64	42	90	57	33	76	46	16	27	20
5.0 kb	98	100	99	86	99	92	80	97	88	73	95	82	67	92	78	59	82	68	36	38	37
10 kb	99	100	100	95	99	97	92	98	95	89	96	92	86	93	90	81	85	83	61	49	54
25 kb	100	100	100	98	100	99	97	99	98	97	98	97	96	96	96	93	89	91	81	58	67
30 kb	100	100	100	98	100	99	98	99	98	97	98	97	96	96	96	94	90	92	83	58	68
50 kb	100	100	100	99	100	99	98	99	99	97	98	98	97	97	97	96	91	93	86	61	71
Trinucleotide z-scores																					
0.5 kb	94	99	97	13	41	20	8	37	13	2	8	4	1	3	2	1	3	1	1	2	1
1.0 kb	93	100	96	56	95	71	40	91	55	28	83	42	23	76	35	16	55	25	7	16	10
2.5 kb	97	100	98	74	95	84	64	92	75	54	87	67	49	83	61	40	67	50	22	28	25
5.0 kb	99	100	99	88	96	92	82	94	88	76	91	83	72	87	79	66	75	70	44	39	41
10 kb	99	100	100	95	98	96	93	96	94	90	93	92	88	91	90	85	81	83	66	49	56
25 kb	100	100	100	98	98	98	97	97	97	96	96	96	95	94	95	94	86	90	82	56	67
30 kb	100	100	100	98	99	98	97	98	98	97	96	96	96	95	95	95	87	90	83	57	68
50 kb	100	100	100	98	99	99	98	98	98	97	97	97	97	96	96	95	89	92	85	61	71

Suppl. Table 1c

Classification statistics for known organisms with raw count-based BLSOMs - classification accuracy of leave-out-datasets fragmented as described for Suppl. Table 1a, showing classification specificities [%] - *sp.*, sensitivities [%] - *sn.*, and F-measure values [%] - *fm.* of batch-learning SOMs trained with di-, tri-, and tetranucleotide raw counts.

	Superkingdom			Phylum			Class			Order			Family			Genus			Species		
Size	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*
Tetranucleotide raw counts																					
0.5 kb	94	99	97	55	89	68	36	78	49	23	59	33	16	45	23	10	25	14	3	5	4
1.0 kb	96	99	97	64	89	75	49	80	61	36	65	46	27	52	36	19	32	24	7	7	7
2.5 kb	97	99	98	79	91	84	69	83	76	60	73	66	52	64	57	41	44	42	19	14	16
5.0 kb	98	99	99	88	93	90	82	87	85	76	79	78	71	73	72	61	54	57	36	20	25
10 kb	99	99	99	94	95	94	91	91	91	88	84	86	84	79	82	79	61	69	55	26	35
25 kb	100	100	100	97	96	97	96	94	95	95	89	92	93	86	89	91	69	79	73	32	45
30 kb	100	100	100	98	96	97	96	94	95	95	90	92	94	86	90	92	70	80	75	33	46
50 kb	100	100	100	98	97	97	97	95	96	96	91	94	95	88	92	94	74	83	80	36	50
Trinucleotide raw counts																					
0.5 kb	95	99	97	57	85	68	39	72	51	27	54	36	21	44	28	14	27	19	4	5	4
1.0 kb	96	99	97	65	86	74	51	75	61	39	60	48	32	51	39	24	33	28	8	7	7
2.5 kb	97	99	98	78	88	83	69	80	74	60	69	64	54	61	57	43	43	43	18	12	14
5.0 kb	98	99	99	87	90	88	81	84	82	76	75	75	70	68	69	62	51	56	32	17	22
10 kb	99	99	99	93	92	92	90	87	88	86	80	83	83	75	79	78	59	67	49	22	31
25 kb	100	100	100	97	94	95	95	90	93	94	85	89	93	81	86	91	66	76	67	28	40
30 kb	100	100	100	97	94	96	96	91	93	94	86	90	93	82	87	91	67	77	70	29	41
50 kb	100	100	100	98	94	96	96	91	94	96	86	91	95	83	88	94	69	80	74	32	44
Dinucleotide raw counts																					
0.5 kb	94	96	95	62	60	61	43	40	41	31	25	28	23	18	20	18	12	14	6	3	4
1.0 kb	96	96	96	72	62	67	57	45	50	47	31	38	39	23	29	33	16	21	14	5	7
2.5 kb	98	96	97	85	67	75	76	53	63	71	40	51	64	32	43	59	23	33	32	8	13
5.0 kb	99	97	98	92	71	80	86	60	71	84	47	60	79	39	52	77	30	43	52	11	19
10 kb	99	97	98	96	75	84	92	64	76	92	53	67	89	45	60	88	35	51	68	15	24
25 kb	100	98	99	98	79	87	95	69	80	95	58	72	93	50	65	93	40	56	77	18	29
30 kb	100	98	99	98	79	87	95	69	80	95	58	72	93	51	66	94	41	57	78	18	29
50 kb	100	98	99	98	80	88	96	70	81	96	59	73	94	53	67	94	43	59	80	19	31

Suppl. Table 1d

Classification statistics of leave-out-datasets fragmented as described for Suppl. Table 1a, showing classification specificities [%] - *sp.*, sensitivities [%] - *sn.,* and F-measure values [%] - *fm.* of batch-learning SOMs trained with batch-learning SOMs trained with tri-, and tetranucleotide z-scores.

	Superkingdom			Phylum			Class			Order			Family			Genus			Species		
Size	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*	*sp.*	*sn.*	*fm.*
Tetranucleotide z-scores																					
0.5 kb	90	99	94	44	96	61	24	91	38	14	81	23	10	71	18	6	43	10	2	7	3
1.0 kb	92	99	96	52	96	67	34	93	50	22	85	35	17	76	28	10	51	17	4	10	6
2.5 kb	96	100	98	69	97	80	55	94	70	44	88	58	37	80	50	27	62	37	12	18	14
5.0 kb	98	100	99	84	98	90	76	95	84	68	90	77	61	83	71	51	70	59	29	26	27
10 kb	99	100	100	94	99	96	90	96	93	87	92	89	83	87	85	77	76	77	54	35	42
25 kb	100	100	100	98	99	98	97	97	97	96	95	95	95	92	93	92	81	86	77	42	54
30 kb	100	100	100	98	99	98	97	98	97	96	95	96	95	92	94	93	82	87	79	42	55
50 kb	100	100	100	98	99	99	98	98	98	97	96	96	96	93	95	95	84	89	84	44	58
Trinucleotide z-scores																					
0.5 kb	89	100	94	44	95	60	23	90	37	13	78	22	9	68	15	5	40	10	2	7	3
1.0 kb	92	100	96	52	95	67	34	91	49	21	81	34	16	73	26	11	49	18	4	10	5
2.5 kb	96	100	98	67	95	79	54	93	68	42	85	56	35	79	49	27	61	37	11	18	14
5.0 kb	98	100	99	81	96	88	73	94	82	64	88	75	59	84	69	50	69	58	26	26	26
10 kb	99	100	99	92	97	94	88	95	91	84	91	87	80	88	84	74	75	75	50	34	40
25 kb	100	100	100	97	98	98	96	97	96	94	95	94	93	92	93	91	81	86	76	42	54
30 kb	100	100	100	97	98	98	96	97	97	95	95	95	94	93	94	93	82	87	78	43	56
50 kb	100	100	100	99	99	99	98	98	98	98	96	96	97	93	95	96	84	89	84	44	58

Suppl. Table 2

Informations on the separation quality of the SOMs used for the taxonomic classification of simulated metagenomes that were constructed from all bacterial and archaeal DNA sequences exceeding 485 kb (roughly the size of *Nanoarchaeum equitans*) in the NCBI GenBank database as of October 2008 (release no. 167). Given are the total number of nodes in the SOM, number of pure nodes (i.e. nodes containing DNA-fragments of a single taxon on the respective taxonomic level) and number of ambiguous nodes (nodes representing DNA-fragments of more than one taxon on the respective taxonomic level). The last column reflects the number of nodes containing DNA fragments from an organism with no taxonomic information on that taxonomic level (e.g. the Epsilonproteobacteria *Sulforovum sp.* NBC 37-1 has no defined taxonomic classification for the taxonomic levels class, order and genus).

Suppl. Table 2a

GSOM - 50 kb training sequence length - tetranucleotide z-scores

taxon	total nodes	pure nodes	ambiguous nodes	without taxonomy
Superkingdom	10 412 (100%)	10 407 (100%)	5 (0%)	0 (0%)
Phylum	10 412 (100%)	10 355 (99%)	51 (0%)	6 (0%)
Class	10 412 (100%)	9 702 (93%)	100 (1%)	610 (6%)
Order	10 412 (100%)	10 042 (96%)	183 (2%)	187 (2%)
Family	10 412 (100%)	9 468 (91%)	278 (3%)	666 (6%)
Genus	10 412 (100%)	9 543 (92%)	731 (7%)	138 (1%)
Species	10 412 (100%)	6 988 (67%)	3 424 (33%)	0 (0%)

Suppl. Table 2b

BLSOM - 50 kb training sequence length - tetranucleotide z-scores

taxon	total nodes	pure nodes	ambiguous nodes	without taxonomy
Superkingdom	9 834 (100%)	9 827 (100%)	7 (0%)	0 (0%)
Phylum	9 834 (100%)	9 765 (99%)	63 (1%)	6 (0%)
Class	9 834 (100%)	9 393 (96%)	117 (1%)	324 (3%)
Order	9 834 (100%)	9 474 (96%)	209 (2%)	151 (2%)
Family	9 834 (100%)	9 049 (92%)	324 (3%)	461 (5%)
Genus	9 834 (100%)	8 937 (91%)	786 (8%)	111 (1%)
Species	9 834 (100%)	6 587 (67%)	3 247 (33%)	0 (0%)

Suppl. Table 3

Informations on the separation quality of the SOMs that were used for the classification of the leave-out-datasets. Information shown as described for Suppl. Table 2.

Suppl. Table 3a

GSOMs as described in Suppl. Table 1a.

taxon	total nodes	pure nodes	ambiguous nodes	without taxonomy
Tetranucleotide rawcounts				
Superkingdom	10 801 (100%)	10 722 (99%)	78 (1%)	1 (0%)
Phylum	10 801 (100%)	10 325 (96%)	475 (4%)	1 (0%)
Class	10 801 (100%)	9587 (89%)	689 (6%)	525 (5%)
Order	10 801 (100%)	9595 (89%)	1 045 (10%)	161 (1%)
Family	10 801 (100%)	9032 (84%)	1 315 (12%)	454 (4%)
Genus	10 801 (100%)	8502 (79%)	2 254 (21%)	45 (0%)
Species	10 801 (100%)	5234 (48%)	4 989 (46%)	578 (5%)
Trinucleotide rawcounts				
Superkingdom	15 472 (100%)	15 364 (99%)	104 (1%)	4 (0%)
Phylum	15 472 (100%)	14 783 (96%)	685 (4%)	4 (0%)
Class	15 472 (100%)	13 920 (90%)	964 (6%)	588 (4%)
Order	15 472 (100%)	13 715 (89%)	1 526 (10%)	231 (1%)
Family	15 472 (100%)	13 017 (84%)	1 877 (12%)	578 (4%)
Genus	15 472 (100%)	12 166 (79%)	3 253 (21%)	53 (0%)
Species	15 472 (100%)	8 069 (52%)	6 585 (43%)	818 (5%)
Dinucleotide rawcounts				
Superkingdom	20 793 (100%)	20 621 (99%)	167 (1%)	5 (0%)
Phylum	20 793 (100%)	19 316 (93%)	1 472 (7%)	5 (0%)
Class	20 793 (100%)	17 980 (86%)	2 012 (10%)	801 (4%)
Order	20 793 (100%)	17 378 (84%)	3 081 (15%)	334 (2%)
Family	20 793 (100%)	16 239 (78%)	3 678 (18%)	876 (4%)
Genus	20 793 (100%)	15 510 (75%)	5 159 (25%)	124 (1%)
Species	20 793 (100%)	11 674 (56%)	7 991 (38%)	1 128 (5%)

Suppl. Table 3b

GSOMs as described in Suppl. Table 1b.

taxon	total nodes	pure nodes	ambiguous nodes	without taxonomy
Tetranucleotide zscores				
Superkingdom	9 063 (100%)	9 055 (100%)	5 (0%)	3 (0%)
Phylum	9 063 (100%)	9 017 (99%)	43 (0%)	3 (0%)
Class	9 063 (100%)	8 488 (94%)	87 (1%)	488 (5%)
Order	9 063 (100%)	8 794 (97%)	159 (2%)	110 (1%)
Family	9 063 (100%)	8 456 (93%)	250 (3%)	357 (4%)
Genus	9 063 (100%)	8 359 (92%)	670 (7%)	34 (0%)
Species	9 063 (100%)	6 159 (68%)	2 256 (25%)	648 (7%)
Trinucleotide zscores				
Superkingdom	22 145 (100%)	22 092 (100%)	52 (0%)	1 (0%)
Phylum	22 145 (100%)	21 698 (98%)	446 (2%)	1 (0%)
Class	22 145 (100%)	20 513 (93%)	664 (3%)	968 (4%)
Order	22 145 (100%)	20 800 (94%)	1 075 (5%)	270 (1%)
Family	22 145 (100%)	19 741 (89%)	1 402 (6%)	1 002 (5%)
Genus	22 145 (100%)	19 489 (88%)	2 565 (12%)	91 (0%)
Species	22145 (100%)	14418 (65%)	6014 (27%)	1713 (8%)

Suppl. Table 3c

BLSOMs as described in Suppl. Table 1c.

taxon	total nodes	pure nodes	ambiguous nodes	without taxonomy
Tetranucleotide rawcounts				
Superkingdom	4 918 (100%)	4 893 (99%)	22 (0%)	3 (0%)
Phylum	4 918 (100%)	4 740 (96%)	175 (4%)	3 (0%)
Class	4 918 (100%)	4 407 (90%)	247 (5%)	264 (5%)
Order	4 918 (100%)	4 426 (90%)	397 (8%)	95 (2%)
Family	4 918 (100%)	4 129 (84%)	493 (10%)	296 (6%)
Genus	4 918 (100%)	4 047 (82%)	846 (17%)	25 (1%)
Species	4 918 (100%)	2 904 (59%)	1 622 (33%)	392 (8%)
Trinucleotide rawcounts				
Superkingdom	5 765 (100%)	5 719 (99%)	35 (1%)	11 (0%)
Phylum	5 765 (100%)	5 465 (95%)	289 (5%)	11 (0%)
Class	5 765 (100%)	5 206 (90%)	392 (7%)	167 (3%)
Order	5 765 (100%)	5 092 (88%)	584 (10%)	89 (2%)
Family	5 765 (100%)	4 878 (85%)	714 (12%)	173 (3%)
Genus	5 765 (100%)	4 584 (80%)	1 146 (20%)	35 (1%)
Species	5 765 (100%)	3 178 (55%)	2 309 (40%)	278 (5%)
Dinucleotide rawcounts				
Superkingdom	33 850 (100%)	32 702 (97%)	1 140 (3%)	8 (0%)
Phylum	33 850 (100%)	25 806 (76%)	8 036 (24%)	8 (0%)
Class	33 850 (100%)	22 601 (67%)	10 009 (30%)	1 240 (4%)
Order	33 850 (100%)	20 276 (60%)	13 205 (39%)	369 (1%)
Family	33 850 (100%)	17 832 (53%)	14 653 (43%)	1 365 (4%)
Genus	33 850 (100%)	16 391 (48%)	17 383 (51%)	76 (0%)
Species	33 850 (100%)	11 961 (35%)	20 495 (61%)	1 394 (4%)

Suppl. Table 3d

BLSOMs as described in Suppl. Table 1d.

Tetranucleotide z-scores:

taxon	total nodes	pure nodes	ambiguous nodes	without taxonomy
Tetranucleotide zscores				
Superkingdom	5 932 (100%)	5 916 (100%)	14 (0%)	2 (0%)
Phylum	5 932 (100%)	5 863 (99%)	67 (1%)	2 (0%)
Class	5 932 (100%)	5 654 (95%)	120 (2%)	158 (3%)
Order	5 932 (100%)	5 672 (96%)	209 (4%)	51 (1%)
Family	5 932 (100%)	5 431 (92%)	314 (5%)	187 (3%)
Genus	5 932 (100%)	5 268 (89%)	650 (11%)	14 (0%)
Species	5 932 (100%)	3 811 (64%)	1 677 (28%)	444 (7%)
Trinucleotide zscores				
Superkingdom	6 299 (100%)	6 287 (100%)	10 (0%)	2 (0%)
Phylum	6 299 (100%)	6 184 (98%)	113 (2%)	2 (0%)
Class	6 299 (100%)	5 956 (95%)	165 (3%)	178 (3%)
Order	6 299 (100%)	5 940 (94%)	280 (4%)	79 (1%)
Family	6 299 (100%)	5 657 (90%)	363 (6%)	279 (4%)
Genus	6 299 (100%)	5 643 (90%)	630 (10%)	26 (0%)
Species	6 299 (100%)	4 079 (65%)	1 791 (28%)	429 (7%)

Suppl. Table 4a

Computation speed of the GSOM-based classifications of the simulated metagenome datasets mimicking habitats of varying complexities (see Figure 1):

contigs of 8 kb or larger (simLC, simMC)			all contigs (simLC, simMC, simHC)		
dataset	no. of elements	time [s]	dataset	no. of elements	time [s]
simLCPhrap	229	16	simLCPhrap	12 665	288
simLCPArachne	202	15	simLCArachne	2 362	62
simMCPhrap	401	20	simMCPhrap	15 197	336
simMCArachne	301	18	simMCArachne	7 307	173
			simHCPhrap	23 398	530
			simHCArachne	578	26

Suppl. Table 4b

Computation speed of the BLSOM-based classifications of simulated metagenome datasets mimicking habitats of varying complexities (see Figure 2):

contigs of 8 kb or larger (simLC, simMC)			all contigs (simLC, simMC, simHC)		
dataset	no. of elements	time [s]	dataset	no. of elements	time [s]
simLCPhrap	229	12	simLCPhrap	12 665	271
simLCArachne	202	11	simLCArachne	2 362	56
simMCPhrap	401	15	simMCPhrap	15 197	329
simMCArachne	301	15	simMCArachne	7 307	178
			simHCPhrap	23 398	478
			simHCArachne	578	22

Suppl. Table 5a

Computation speed for the classification of known organisms with raw count-based GSOMs, z-score-based GSOMs, raw-count based BLSOMs, and z-score-based BLSOMs. Time values for the classifications of fragments of 0.5 kb to 50 kb sizes along with the total number of elements in the particular dataset and motif lengths from di- to tetranucleotides are shown. Fragments were on SOMs trained with oligonucleotide raw counts and z-scores, as described in Suppl. Table 1a to d.

			raw count GSOM	z-score GSOM	raw count BLSOM	z-score BLSOM
	dataset	no. of elements	time	time	time	time
Tetranucleotides						
	0.5 kb	1 085 582	405 min 13 s	355 min 20 s	184 min 23 s	223 min 30 s
	1.0 kb	543 138	199 min 23 s	169 min 41 s	92 min 7 s	111 min 30 s
	2.5 kb	217 688	79 min 57 s	68 min 4 s	37 min 12 s	45 min 12 s
	5 kb	109 213	42 min 36 s	34 min 13 s	18 min 41 s	23 min 1 s
	10 kb	54 978	20 min 31 s	17 min 18 s	9 min 23 s	11 min 51 s
	25 kb	22 485	8 min 30 s	7 min 9 s	3 min 53 s	4 min 47 s
	30 kb	18 885	7 min 7 s	6 min 2 s	3 min 18 s	3 min 56 s
	50 kb	11 673	4 min 28 s	3 min 49 s	2 min 4 s	2 min 27 s
Trinucleotides						
	0.5 kb	1 085 582	170 min 30 s	241 min 30 s	55 min 36 s	63 min 9 s
	1.0 kb	543 138	85 min 58 s	119 min 52 s	29 min 24 s	34 min 50 s
	2.5 kb	217 688	34 min 35 s	48 min 8 s	11 min 28 s	12 min 39 s
	5 kb	109 213	17 min 26 s	24 min 23 s	5 min 54 s	6 min 19 s
	10 kb	54 978	8 min 49 s	12 min 21 s	2 min 57 s	3 min 15 s
	25 kb	22 485	3 min 40 s	5 min 7 s	1 min 14 s	1 min 21 s
	30 kb	18 885	2 min 36s	4 min 20 s	1 min 5 s	1 min 8 s
	50 kb	11 673	1 min 57 s	2 min 44 s	0 min 39 s	0 min 43 s
Dinucleotides						
	0.5 kb	1 085 582	81 min 3 s		158 min 45 s	
	1.0 kb	543 138	39 min 49 s		88 min 25 s	
	2.5 kb	217 688	16 min 0 s		36 min 47 s	
	5 kb	109 213	8 min 5 s		14 min 58 s	
	10 kb	54 978	4 min 8 s		7 min 9 s	
	25 kb	22 485	1 min 44 s		2 min 59 s	
	30 kb	18 885	1 min 28 s		3 min 12 s	
	50 kb	11 673	0 min 56 s		1 min 58 s	

Suppl. Table 6

Time statistics for biodiversity assessments of the North Sea metagenomes over time using Taxonomic classification of assemblies exceeding 2.5 kb with TaxSOM, as described in Material and Methods (Real-world dataset) and Figure 3.

dataset	no. of elements	time
11th February 2009	227	7 s
31th March 2009	2 321	29 s
7th April 2009	3 229	38 s
14th April 2009	2 999	36 s
16th June 2009	1 137	16 s

References

Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T. (2003). Informatics for unveiling hidden genome signatures. *Genome Res* **13**: 693-702.

Alahakoon D, Halgamuge SK, Srinivasan B. (2000). Dynamic Self-Organizing Maps with Controlled Growth for Knowledge Discovery. *IEEE Transactions on Neural Networks* **11**: 601-614.

Chan C-KK, Hsu AL, Tang S-L, Halgamuge SK. (2007). Using Growing Self-Organising Maps to Improve the Binning Process in Environmental Whole-Genome Shotgun Sequencing. *Journal of Biomedicine and Biotechnology* **2008**:

Kohonen T, Kohonen T, Schroeder MR, Huang TS, Maps SO. (2001). Springer-Verlag New York. *Inc, Secaucus, NJ*

Schbath S, Prum B, de Turckheim E. (1995). Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J Comput Biol* **2**: 417-437.

Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO. (2004). Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol* **6**: 938-947.