Algorithms for Molecular Biology

Research

P-value based visualization of codon usage data

Peter Meinicke*[1], Thomas Brodag[2], Wolfgang Florian Fricke[3] and Stephan Waack[2]

Address: [1]Abteilung Bioinformatik, Institut für Mikrobiologie und Genetik, Georg-August-Universität Göttingen, Goldschmidtstr. 1, 37077 Göttingen, Germany, [2]Institut für Numerische und Angewandte Mathematik, Universität Göttingen, Lotzestr. 16, 37083 Göttingen, Germany and [3]Göttingen Genomics Laboratory, Universität Göttingen, Grisebachstr. 8, 37077 Göttingen, Germany

Email: Peter Meinicke* - pmeinic@gwdg.de; Thomas Brodag - Thomas.Brodag@T-Online.de; Wolfgang Florian Fricke - wfricke@gwdg.de; Stephan Waack - waack@cs.uni-goettingen.de

* Corresponding author

Abstract

Two important and not yet solved problems in bacterial genome research are the identification of horizontally transferred genes and the prediction of gene expression levels. Both problems can be addressed by multivariate analysis of codon usage data. In particular dimensionality reduction methods for visualization of multivariate data have shown to be effective tools for codon usage analysis. We here propose a multidimensional scaling approach using a novel similarity measure for codon usage tables. Our probabilistic similarity measure is based on P-values derived from the well-known chi-square test for comparison of two distributions. Experimental results on four microbial genomes indicate that the new method is well-suited for the analysis of horizontal gene transfer and translational selection. As compared with the widely-used correspondence analysis, our method did not suffer from outlier sensitivity and showed a better clustering of putative alien genes in most cases.

Background

The standard genetic code of protein coding DNA sequences shows a redundancy, since different triplet codons may be used to code for the same amino acid. In general, codon usages show organism-specific patterns. However, codon usage variation within a single genome can be an important source of information about gene expression levels and events of horizontal gene transfer. In particular, dimensionality reduction methods have widely been used for the analysis of codon usage patterns in microbial genomes. These methods provide a low-dimensional point representation of genes, where the proximity of gene-specific points indicates a similar codon usage of the associated genes. Hence, the resulting two-dimensional scatter plots enable a total view on the genome which may reveal a clustering of genes according to groups of nearby points. These clusters can for instance provide evidence for horizontal gene transfer according to groups of putative alien genes [1,2] or for translational selection according to groups of highly expressed genes [3,4].

As a standard method for scatter plot visualization of codon usage data, researchers mostly resort to the so-called correspondence analysis (CA) which has originally been developed for the analysis of contingency tables [5]. From the original formulation it is not completely clear how CA applies to codon counts. Because different pre-processing and normalization schemes have been proposed, the use of CA in codon usage studies has not been

without controversy [6]. Nevertheless, CA has been applied for the analysis of many bacterial genomes, including those of *Escherichia coli* [1,3], *Bacillus subtilis* [4,7,8], *Borrelia burgdorferi* [9,10], *Chlamydia trachomatis* [11], *Mycoplasma genitalium* [12], *Helicobacter pylori* [13] and *Pseudomonas aeruginosa* [14].

Recently, self-organizing maps [15] have been proposed as an alternative visualization method for codon usage data [2,16,17]. Although this method provides a simultaneous clustering of the data which may be useful in certain contexts, it requires to choose the size of a discrete grid on which the genes are mapped in a non-linear way. The grid-size is a critical parameter of the method and directly controls the final clustering in the visualization. Unfortunately, the grid-size of self-organizing maps is a so-called *hyperparameter* which usually cannot be inferred from the data in an unsupervised manner. Therefore the resulting visualizations bare the risk of being highly subjective.

Here we present a visualization method, which has been tailored to the analysis of codon usage data while not depending on difficult to tune hyperparameters. Our visualization method is based on multidimensional scaling and a new similarity measure for codon usage data. In the following we first introduce our probabilistic similarity measure for codon usage tables and outline the corresponding algorithm for multidimensional scaling based on P-values. Then we provide some visualizations for the analysis of four microbial genomes and discuss our results in comparison with the results obtained from the classical correspondence analysis method.

P-values for multidimensional scaling

For the analysis of codon usage tables we developed a special similarity measure which has been derived from the well-known chi-square test for the comparison of two distributions. Unlike the classical chi-square test we do not decide whether two distributions are equal or not, but instead we only use the corresponding P-values to compute a similarity measure for the underlying codon usage tables. For each pair of genes we compare the corresponding codon distributions on the basis of the codon frequencies in the two genes. For a suitable similarity score we average the P-values of the amino acid specific chi-square tests. We start with the counts N_{ij}^l for codon c_i^l of amino acid a_i in the j-th gene. These counts sum up to $n_{ij} = \sum_{l=1}^{L_i} N_{ij}^l$ over the number L_i of different codons for amino acid a_i. Note that n_{ij} corresponds to the number of occurrences of amino acid a_i in gene j. With these counts

we compute the chi-square statistic for each pair (j, k) of genes:

$$\chi_{ijk}^2 = \sum_{l=1}^{L_i} \frac{\left(\sqrt{n_{ik}/n_{ij}} N_{ij}^l - \sqrt{n_{ij}/n_{ik}} N_{ik}^l \right)^2}{N_{ij}^l + N_{ik}^l} \qquad (1)$$

The classical chi-square test for comparison of two distributions is based on the following proposition: under the null hypothesis that the corresponding samples were drawn from the same probability distribution, the variable χ_{ijk}^2 is asymptotically chi-square distributed with L_i degrees of freedom. Here we do not perform a chi-square test, but rather calculate the P-value P_{ijk} associated with the chi-square statistic χ_{ijk}^2. The P-values are obtained from the chi-square probability function which is an incomplete gamma function [18]. A small value of P_{ijk} indicates a significant difference between the codon distributions of gene j and k with respect to amino acid a_i. For a number of M genes in a genome we then assemble the $M \times M$ matrix **S** of similarity scores with non-negative elements

$$S_{jk} = 1/n_a \sum_{i=1}^{n_a} P_{ijk} \qquad (2)$$

where n_a is the number of amino acids. Note that **S** has unit diagonal elements, i.e. $S_{jj} = 1$, because the P-value for tables with identical counts is one. Consequently all off-diagonal elements are in the range [0, 1].

In order to derive a suitable low-dimensional point representation of genes we apply classical multidimensional scaling (see e.g. [19]) to the above similarities. The objective is to find a two-dimensional point configuration with interpoint distances reflecting the codon usage similarities of the corresponding genes. To perform classical scaling based on similarities we first transform the similarity matrix **S** into a positive semi-definite matrix **C** by subtracting the smallest eigenvalue $\lambda_{\min}$ of **S** from all of its diagonal elements:

$$\mathbf{C} = \mathbf{S} - \lambda_{\min}\mathbf{I} \qquad (3)$$

where **I** is the $M \times M$ identity matrix. Note that this transformation preserves the equality of diagonal elements. With the $M \times M$ centering matrix **H** with elements

$$H_{jk} = \begin{cases} 1 - 1/M & \text{if } j = k, \\ -1/M & \text{else.} \end{cases} \qquad (4)$$

we finally obtain the matrix

B = **HCH**.   (5)

It can be shown that for a positive semi-definite matrix **C** the distance matrix **D** with elements obtained by the standard transformation $D_{jk} = (C_{jj}^2 - 2C_{jk} + C_{kk}^2)^{-1/2}$ is Euclidean and **B** is a centered inner product matrix ([19], pp. 402). Therefore principal components can be obtained from (partial) eigenvalue decomposition of **B**. Thus, for 2D-visualization we compute the two leading eigenvectors x_1 and x_2 of **B** associated with the largest and second largest eigenvalue, respectively. The *M* components of x_1 and x_2 provide the x_1 and x_2 coordinates for the *M* genes, which are utilized for scatter plot visualization.

Experimental results
Data sets
To evaluate our multidimensional scaling (MDS) approach, we focused on visualizations of ribosomal protein genes and putative alien genes for different microbial genomes. Ribosomal protein genes belong to the class of highly expressed genes which tend to use codons associated with the prevalent tRNAs present in the organism. If translational selection is one of the main sources for codon preferences in a particular genome, then codon usage can in turn be used for the prediction of putative highly expressed genes [20]. Another source of codon usage variation in microbial genomes is provided by the insertion of foreign DNA by means of horizontal gene transfer. Thus, putative alien genes may also be predicted on the basis of codon usage analysis [2,21]. While ribosomal protein genes can be identified from the annotations of completely sequenced genomes, reliable information about putative alien genes is much more difficult to obtain. We combined predictions of the SIGI-HMM tool [22] with existing references from the literature in order to obtain suitable test sets for our evaluations. SIGI-HMM is based on a Hidden Markov Model for the detection of genomic islands and, in contrast to our MDS-based visualization method, it explicitly uses information about the locations of genes on the corresponding chromosomes. However, unlike MDS, SIGI-HMM does not consider codon usage correlations between different amino acids. Using the two complementary kinds of information exclusively, both methods provide completely different approaches to codon usage analysis [22].

For the evaluation of the MDS-based visualizations we analyzed the microbial genomes of *Escherichia coli K-12*, *Bacillus subtilis*, *Vibrio cholerae* and *Thermus thermophilus HB8*. We used annotated DNA sequence data in the EMBL format publicly available from EBI [23]. Ribosomal protein genes were extracted from the datasets of the completely annotated genomes. Putative alien genes were selected according to the following information: On chromosome 1 of *V. cholerae* two genomic islands were predicted by SIGI-HMM that comprise a gene cluster for a toxin-coregulated pilus and fragments of a temperate filamentous phage described in [24]. Both clusters are closely associated with the pathogenicity of *V. cholerae*. For *Bacillus subtilis* 10 integrated prophages have been described based on experimental evidence and theoretical considerations [25-28]. Nine of these prophages overlap with genomic islands as predicted by SIGI-HMM. For *Escherichia coli* K-12 the authors of [29] used different compositional variables and estimated that about 18% of the genome have been imported by horizontal gene transfer. In contrast, SIGI-HMM predicted 580 genes (13,6%) to be putatively alien. The largest genomic islands comprise the cryptic prophages CP4-6, DLP12, e14, Rac, Qin, CP4-44, CPS-53, Eut, CP4-57, and the phage-like element KpLE2 (reviewed in [30]). For the extremophilic bacterium *Thermus thermophilus HB8* no genomic islands have been described so far. SIGI-HMM predicted a contiguous gene cluster of 5 genes associated with functions in cell wall biosynthesis to be putative alien. The total number of putative alien genes and the number of ribosomal protein genes for all species considered here are summarized in table 1. Additional file 1 provides a detailed list of all putative alien genes used for the visualization.

Visualization
We compared our multidimensional scaling (MDS) approach with the correspondence analysis (CA) method as implemented in the *CodonW* program [31] of J. Peden. Computations were based on *relative synonymous codon usage* (*RSCU*) values which is the most common way to perform CA on codon usage data [6]. For both methods

Table 1: Number of genes used for the visualization for all species under consideration. Given are the number of putative alien genes, the number of ribosomal protein genes and the total number of genes on the respective chromosomes.

species	# genes (total)	# ribosomal protein genes	# putative alien genes
E. coli	4254	61	206
B. subtilis	4106	57	317
V. cholerae Chr1	2736	64	41
V. cholerae Chr2	1092	0	216
T. thermophilus	1973	60	5

the resulting coordinates were normalized according to a unit variance of the leading two factors and principal components, respectively.

The CA-based visualization for E. coli (Fig. 1) shows the typical "rabbit head" structure which has been described in [1]. The "ears" correspond to two branches of the distribution with low density. The "left ear" in the upper left corner shows a cluster of ribosomal protein genes while putative alien genes are mainly located around the other branch of the distribution. The MDS plot in Fig. 1 shows a similar picture with ribosomal protein genes and putative alien genes again concentrated in the two branches of the distribution which here appears rotated by 180 degrees. Comparing the visualizations, most of the ribosomal protein genes are well-clustered in both plots while putative alien genes are slightly more concentrated in the MDS plot. Note that the CA-based visualization shows an outlier at the lower boundary of the plot which is not among the putative alien genes.

For *B. subtilis* (Fig. 2) both visualization methods show a good clustering of putative alien genes and ribosomal protein genes in the branches of the distribution. Again the lower boundary of the CA plot is determined by an outlier which does not belong to the set of putative alien genes.

For the first chromosome of *V. cholerae* (Fig. 3) the comparison shows a similar situation as for *B. subtilis*: in both plots, most of the ribosomal protein and putative alien genes are well-clustered in the two branches of the distribution. In the lower left corner of the CA-based plot there is an outlier which is not in the set of putative alien genes. As chromosome II of *V. cholerae* does not contain any ribosomal protein genes, the visualization of this replicon is restricted to putative alien genes (Fig. 4). These genes are slightly more concentrated in the MDS-based plot. Again, the lower boundary of the CA-plot is determined by an outlier which is not among putative alien genes.

For *T. thermophilus* (Fig. 5) the outlier sensitivity of CA results in a highly distorted plot which makes it difficult to draw any conclusions from the visualization at all. While ribosomal protein genes are clumped together with the remaining genes in a small region of the plot, putative alien genes are widespread in a region of low density. In contrast, the MDS-based plot shows a specific proximity of putative alien genes in a tail at the right border and the ribosomal protein genes at least show some weak clustering in the upper right part of the core distribution.

Conclusion

We proposed an approach for the visualization of codon usage data, using multidimensional scaling (MDS). In that context we introduced a novel similarity measure for
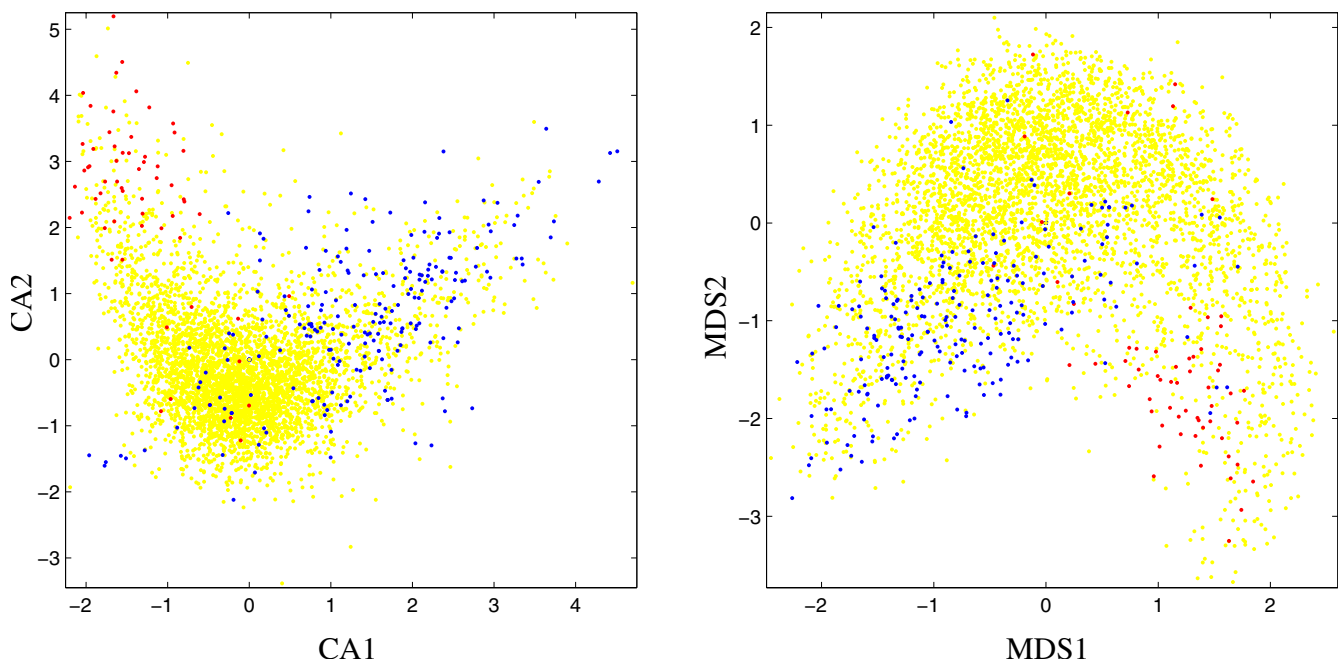

Figure 1
Scatter plots for *E. coli* based on first two components of correspondence analysis (left, CA) and P-value based multidimensional scaling (right, MDS), respectively. Red dots: ribosomal protein genes; blue dots: putative alien genes; yellow dots: all other genes.
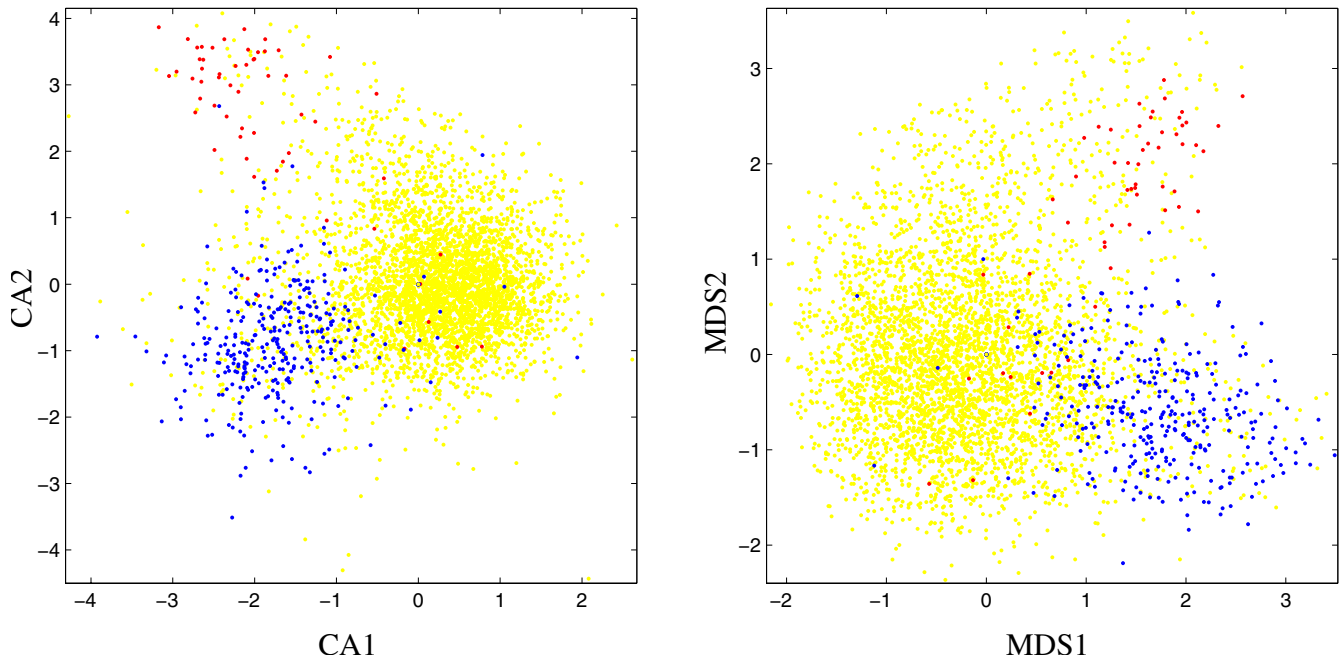
Figure 2
Scatter plots for *B. subtilis* based on first two components of correspondence analysis (left, CA) and P-value based multidimensional scaling (right, MDS), respectively. Red dots: ribosomal protein genes; blue dots: putative alien genes; yellow dots: all other genes.
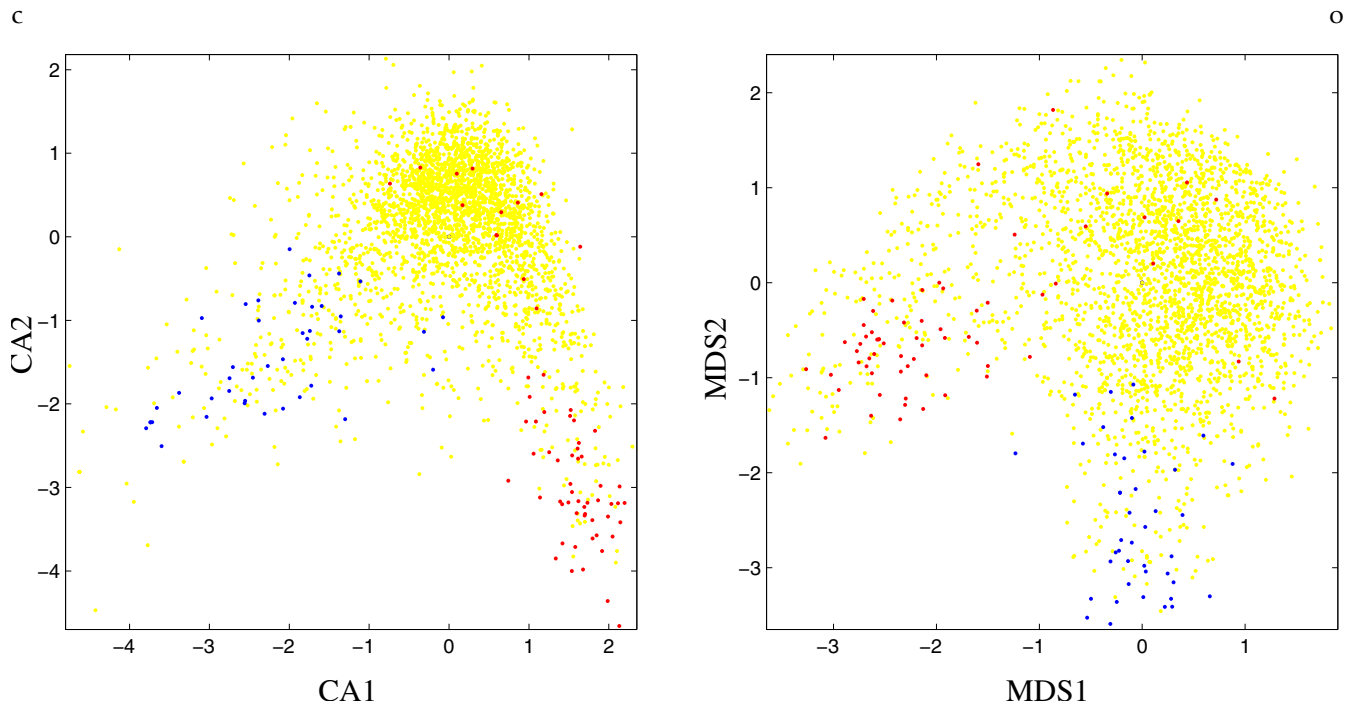

Figure 3
Scatter plots for *V. cholerae* (chromosome 1) based on first two components of correspondence analysis (left, CA) and P-value based multidimensional scaling (right, MDS), respectively. Red dots: ribosomal protein genes; blue dots: putative alien genes; yellow dots: all other genes.

Figure 4
Scatter plots for *V. cholerae* (chromosome 2) based on first two components of correspondence analysis (left, CA) and P-value based multidimensional scaling (right, MDS), respectively. Red dots: ribosomal protein genes; blue dots: putative alien genes; yellow dots: all other genes.
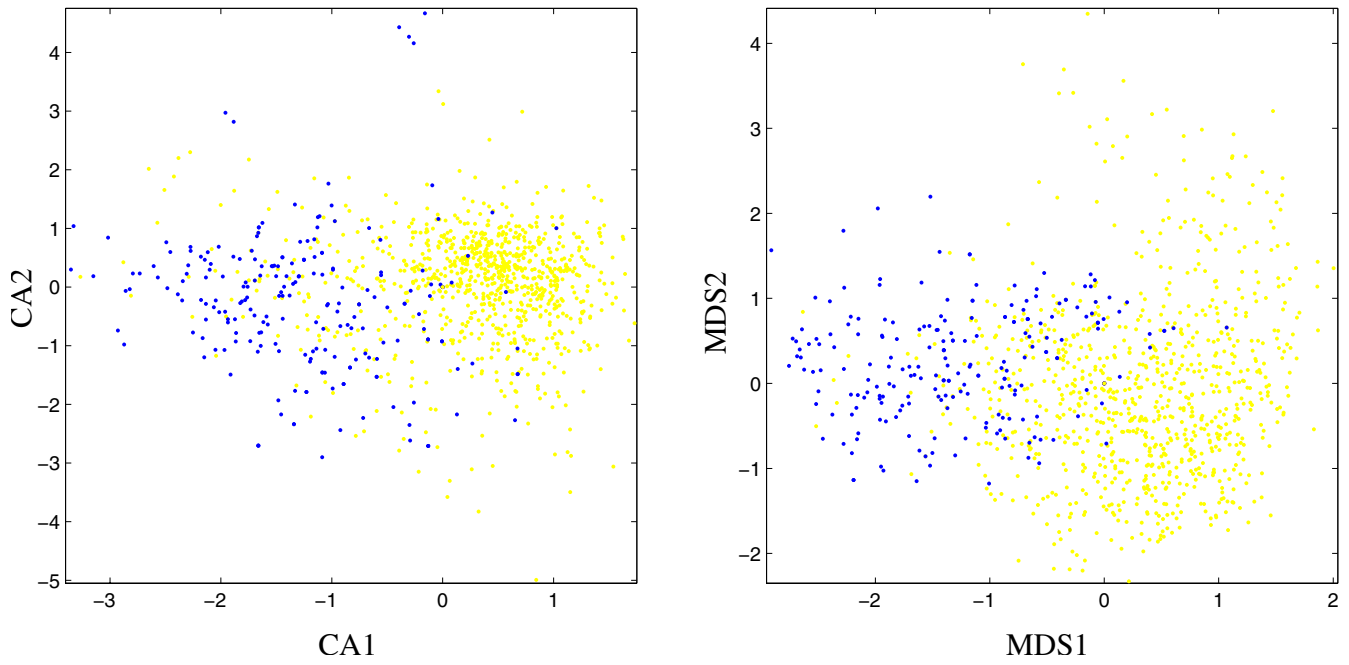

Figure 5
Scatter plots for *T. thermophilus* based on first two components of correspondence analysis (left, CA) and P-value based multidimensional scaling (right, MDS), respectively. Red dots: ribosomal protein genes; blue dots: putative alien genes; yellow dots: all other genes.
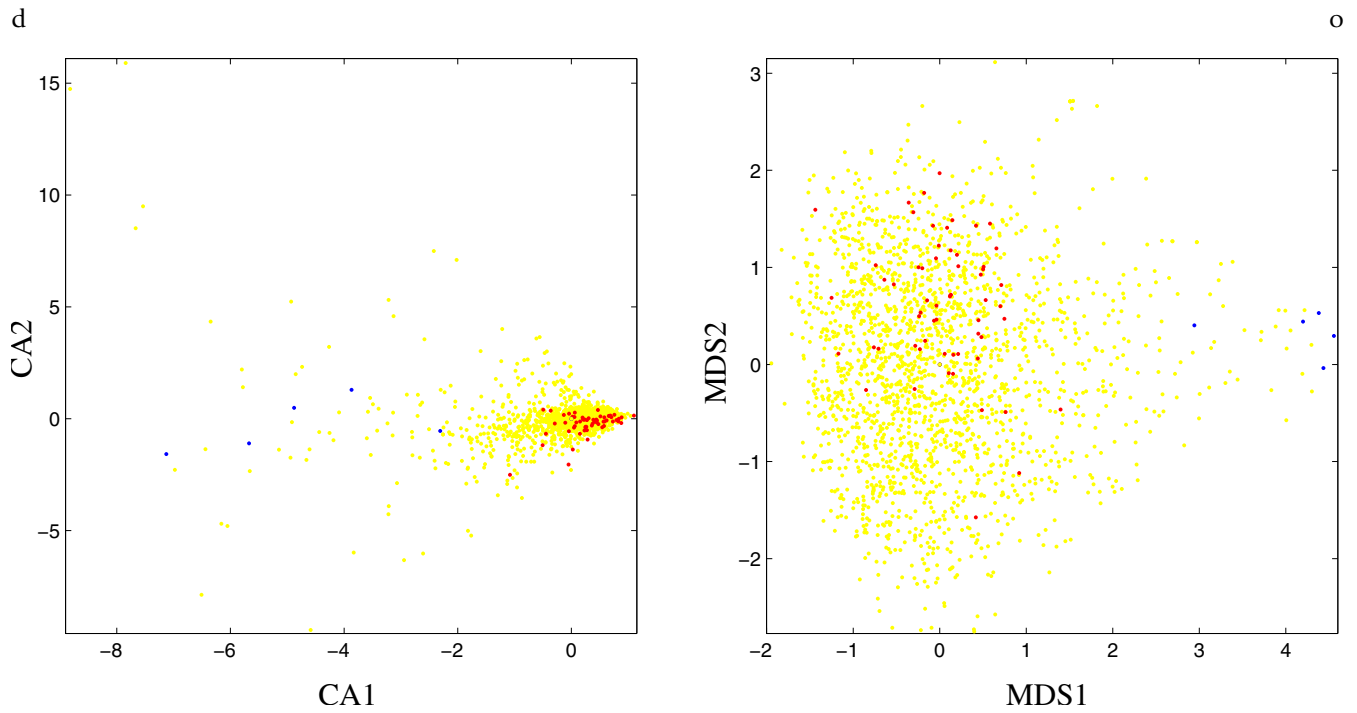
n usage tables, which has been derived from the classical chi-square test. An important feature of our P-value based similarity measure is that it does not involve any hyperparameters. Therefore a subjective "bias" on the visualization due to user-adjusted parameters is effectively avoided. Our comparisons with the widely-used correspondence analysis (CA) method in most cases showed a slightly better clustering of putative alien genes for our P-value based visualization. In particular the results indicate that our approach is more robust than the CA-based visualization method. The outlier-sensitivity of CA becomes apparent in the plots for all species considered here and has already been observed in previous studies [9]. While in most cases the CA-based visualizations are still useful in terms of a suitable clustering of ribosomal protein and putative alien genes, for *T. thermophilus* that sensitivity results in an inappropriate plot which complicates interpretation.

Additional material

Additional File 1
provides an Excel table (XLS) containing a detailed list of all putative alien genes used for the visualization.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1748-7188-1-10-S1.xls]

Acknowledgements

References

1.   Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A: **Evidence for horizontal gene transfer in Escherichia coli speciation.** *J Mol Biol* 1991, **222**:851-856.
2.   Wang HC, Badger J, Kearney P, Li M: **Analysis of codon usage patterns of bacterial genomes using the self-organizing map.** *Mol Biol Evol* 2001, **18**:792-792.
3.   Holm L: **Codon usage and gene expression.** *Nucleic Acids Res* 1986, **14**:3075-3087.
4.   Shields DC, Sharp PM: **Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases.** *Nucleic Acids Res* 1987, **15**:8023-8040.
5.   Hill MO: **Correspondence analysis: a neglected multivariate method.** *Appl Stat* 1974, **23**:340-354.
6.   Perrière G, Thioulouse J: **Use and misuse of correspondence analysis in codon usage studies.** *Nucleic Acids Res* 2002, **30**:4548-4555.
7.   Perrière G, Gouy M, Gojobori T: **NRSub: a non-redundant data base for the Bacillus subtilis genome.** *Nucleic Acids Res* 1994, **22**:5525-5529.
8.   Moszer I, Rocha EP, Danchin A: **Codon usage and lateral gene transfer in Bacillus subtilis.** *Curr Opin Microbiol* 1999, **2**:524-528.
9.   McInerney JO: **Replicational and transcriptional selection on codon usage in Borrelia burgdorferi.** *Proc Natl Acad Sci USA* 1998, **95**:10698-10703.
10. Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, Wolfe KH: **Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases.** *Nucleic Acids Res* 1999, **27**:1642-1649.
11. Romero H, Zavala A, Musto H: **Codon usage in Chlamydia trachomatis is the result of strand-specific mutational biases and a complex pattern of selective forces.** *Nucleic Acids Res* 2000, **28**:2084-2090.
12. McInerney JO: **Prokaryotic Genome Evolution as Assessed by Multivariate Analysis of Codon Usage Patterns.** *Microbial and Comparative Genomics* 1997, **2**:1-10.
13. Lafay B, Atherton JC, Sharp PM: **Absence of translationally selected synonymous codon usage bias in Helicobacter pylori.** *Microbiology* 2000, **146(Pt 4)**:851-860.
14. Gupta SK, Ghosh TC: **Gene expressivity is the main factor in dictating the codon usage variation among the genes in Pseudomonas aeruginosa.** *Gene* 2001, **273**:63-63.
15. Kohonen T: *Self-Organizing Maps* Springer, Berlin; 1995.
16. Kanaya S, Kinouchi M, Abe T, Kudo Y, Yamada Y, Nishi T, Mori H, Ikemura T: **Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the E. coli O157 genome.** *Gene* 2001, **276**:89-89.
17. Supek F, Vlahovicek K: **INCA: synonymous codon usage analysis and clustering by means of self-organizing map.** *Bioinformatics* 2004, **20**:2329-2330.
18. Press WH, Flannery BP, Teukolsky SA, Vetterling WT: *Numerical Recipes in C* 2nd edition. Cambridge University Press, Cambridge; 1992.
19. Mardia KV, Kent JT, Bibby JM: *Multivariate Analysis* Academic Press, London; 1979.
20. Karlin S, Mrazek J: **Predicted highly expressed genes of diverse prokaryotic genomes.** *J Bacteriol* 2000, **182(18)**:5238-5250.
21. Merkl R: **SIGI: score-based identification of genomic islands.** *BMC Bioinformatics* 2004, **5**:22.
22. Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, Surovcik K, Meinicke P, Merkl R: **Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models.** *BMC Bioinformatics* 2006, **7**:142.
23. **European Bioinformatics Institute** [http://www.ebi.ac.uk/genomes/]
24. Waldor MK, Mekalanos JJ: **Lysogenic conversion by a filamentous phage encoding cholera toxin.** *Science* 1996, **272(5270)**:1910-1914. Comment.
25. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S, Borriss R, Boursier L, Brans A, Braun M, Brignell SC, Bron S, Brouillet S, Bruschi CV, Caldwell B, Capuano V, Carter NM, Choi SK, Codani JJ, Connerton IF, Danchin A: **The complete genome sequence of the gram-positive bacterium Bacillus subtilis.** *Nature* 1997, **390(6657)**:249-256.
26. Takemaru K, Mizuno M, Sato T, Takeuchi M, Kobayashi Y: **Complete nucleotide sequence of a skin element excised by DNA rearrangement during sporulation in Bacillus subtilis.** *Microbiology* 1995, **141(Pt 2)**:323-327.
27. Wood HE, Dawson MT, Devine KM, McConnell DJ: **Characterization of PBSX, a defective prophage of Bacillus subtilis.** *J Bacteriol* 1990, **172(5)**:2667-2674.
28. Zahler SA, Korman RZ, Rosenthal R, Hemphill HE: **Bacillus subtilis bacteriophage SPbeta: localization of the prophage attachment site, and specialized transduction.** *J Bacteriol* 1977, **129(1)**:556-558.
29. Lawrence JG, Ochman H: **Molecular archaeology of the Escherichia coli genome.** *Proc Natl Acad Sci USA* 1998, **95(16)**:9413-9417.
30. Casjens S: **Prophages and bacterial genomics: what have we learned so far?** *Mol Microbiol* 2003, **49(2)**:277-300.
31. **CodonW** [http://codonw.sourceforge.net/]