Phylogenetic trees based on k-nucleotide frequencies

Markus Friberg*†, Gaston Gonnet†, Peter von Rohr† , and Kurt Tobler‡

January 6, 2004

Running heads: Phylogenetic trees based on k-nucleotide freq.

Keywords: Phylogenetic tree, dinucleotides, trinucleotides, tetranucleotides

*Corresponding author: friberg@inf.ethz.ch, phone: +41 1 632 3145, fax: +41 1 632 1374

†Institute of Computational Science, ETH, Hirschengraben 84, 8092 Zurich, Switzerland

‡Institute of Virology, University of Zurich, Winterthurerstr. 266a, 8057 Zurich, Switzerland

Abstract

We investigate an alternative approach to phylogenetic tree construction based on k-nucleotide frequencies. Compared to traditional phylogenetic analysis, where trees are produced from multiple sequence alignments, the k-nucleotide approach has the advantage that it uses all the available information. It is also almost insensitive to gene rearrangements, which are common in viruses and cause havoc to MSA methods.

There are several ways of comparing the k-nucleotide frequencies to compute the phylogenetic distance between two organisms. In previous work, different methods have been used producing different results. We study in detail the method of going from k-nucleotide counts to phylogenetic trees. The different methods are evaluated using a tree quality index. We analyze an extensive number of new and already reported methods to see which produce the most accurate phylogenetic trees. According to our analysis, the method that works best in general is the Euclidean distance of odds ratios.

This paper was motivated by the good results obtained for the classification of the SARS virus in the context of other coronaviruses and RNA viruses. As an additional example of how the method can be used, we study 11 vertebrate species for which the phylogenetic tree is known.

1 Introduction

Previous studies have shown that dinucleotide composition is similar in related organisms and different in unrelated organisms (Nussinov, 1984) (Karlin and Burge, 1995). Furthermore, the composition is relatively stable over the whole genome, in both coding and non-coding regions. Hence, it is possible to compute a genome signature for several organisms and use this measure to create a phylogenetic tree (Nakashima et al., 1998). We define the signature as the vector describing the k-nucleotide counts. Except for phylogenetic analysis, these signatures have been used in several applications, e.g. assessment of general relatedness of genomes (Karlin and Ladunga, 1994), genetic sequence classification (Chuzhanova et al., 1998), clustering of genes from different genomes (Nakashima et al., 1998), and recognition of species-specific sequence blocks (Abe et al., 2002).

The most popular way to build phylogenetic trees is from multiple sequence alignments (MSA). Either complete genomes are compared or the deduced amino acid sequences of coding regions. Both approaches suffer from disadvantages (Brocchieri, 2001). In the case of coding nucleic acids, the analysis has to be limited to the specific parts, which contain related proteins. Naturally, we have to know the coding regions, and we have to make an arbitrary choice of which regions to compare. Different regions can lead to different trees. If the sequences are of very different lengths or if they are only to a low degree similar, the resulting multiple sequence alignment will be of low quality, leading to unreliable trees. Gene rearrangements, common in viruses, will further complicate/destroy MSAs. Finally, finding a multiple sequence alignment for long regions of many organisms is so computationally intensive that it is often impossible to do in practice. In summary, MSAs may provide a good basis for phylogenetic analysis, but constructing the MSA themselves is plagued with problems.

In the case of k-nucleotide analysis, these problems do not exist. Although the relation between evolutionary distances and the signatures is not entirely clear, similar sequences will generate similar signatures. Another advantage is that k-nucleotide analysis does not, or only marginally, suffer from gene rearrangements that are quite common in viruses.

In the process of creating a phylogenetic tree from k-nucleotide data, several choices can be made (Figure 1). Many possible computational paths can take us from signatures to a tree. We use a distance matrix, containing the estimated pairwise distances between the organisms in our population, as a fixed intermediate step. From the distance matrix we compute the phylogenetic tree using the least squares method, as a second fixed step. However, different methods can be used to go from signatures to the distance matrix. In this paper, we evaluate several of these methods with respect to a quantifiable measures of quality: topological distance (Penny and Hendy, 1985) (Nei and Kumar, 2000) and the least squares error of the resulting phylogenetic tree.

Parsimony and maximum likelihood are other popular methods for creating phylogenetic trees. However, since they need a multiple sequence alignments, we have not included them in our comparison. We here focus on phylogenetic trees based on k-nucleotide frequencies, and show that they could be an interesting alternative when good multiple sequence alignments are not possible to compute.

In previous work, mostly dinucleotides have been used. Our analysis shows that trinucleotides contain more information, and may be more suitable for phylogenetic analysis.

2 Methods

[Figure 1 about here.]

Several methods can be used to go from the k-nucleotide counts to the distance matrix (Figure 1). For the purpose of evaluating the different methods, two databases have been compiled: one of 49 RNA viruses and one of the whole mitochondrial genome of 11 vertebrates (Russo et al., 1996). Each of the methods evaluated are used to create phylogenetic trees of the RNA viruses and the vertebrates. Quality measures are used to assess the quality of the trees produced. The construction of the tree from the distances is probably our strongest link. We have constructed each tree 1000 times and selected the one with the least error. We use Darwin (Gonnet et al., 2000) to do this computation. Source code is available online as a bio-recipe (Friberg, 2003).

2.1 k-nucleotide counts - signatures

The k-nucleotide counts a are counted over the whole genome, shifting the frame one base at a time. This allows partial overlap of the counts. E.g. in a genome consisting of n nucleotides we count $n - k + 1$ k-nucleotides. There are 4^k different combinations, so e.g. a trinucleotide count vector ($k = 3$) will have dimension 64.

2.2 Methods to build computed vectors

2.2.1 Expected and observed k-nucleotide frequencies

The expected frequency f_{exp} of a k-nucleotide is defined as the product of the corresponding nucleotide frequencies. E.g. the trinucleotide `ACG` has the expected frequency $f_{exp} = f_A f_C f_G$.

The observed frequency of a k-nucleotide is defined as the number of counts for that k-nucleotide divided by the total number of counts:

$$f_{obs} = \frac{a}{n - k + 1} \tag{1}$$

where n is the total number of nucleotides in the genome.

2.2.2 Odds ratios

The odds ratio is a measure of the bias between the observed and expected k-nucleotide frequencies:

$$r_{odds} = \frac{f_{obs}}{f_{exp}} \tag{2}$$

This method was used by (Campbell et al., 1999) (Gentles and Karlin, 2001) (Karlin and Burge, 1995).

2.2.3 Log-odds ratios

As an alternative to (2), we evaluate the logarithm of the odds ratio:

$$r_{logodds} = \log \frac{f_{obs}}{f_{exp}} \tag{3}$$

This method was used by (Nakashima et al., 1997).

2.2.4 Poisson deviates

Another measure of the bias between the observed and expected k-nucleotide frequencies is the normalized deviation from a Poisson distribution:

$$r_{poisson} = \frac{f_{obs} - f_{exp}}{\sqrt{n f_{exp}(1 - f_{exp})}} \tag{4}$$

2.2.5 Odds difference

This measure is similar to (4), but does not contain any normalization factor. It was used by (Nakashima et al., 1998).

$$r_{oddsdiff} = f_{obs} - f_{exp} \tag{5}$$

2.3 Computing distances from vectors

There are fairly well established ways of computing distances between vectors. Most of these boil down to computing the norm of the difference of the vectors. We use the norms:

$$||x|| = \sqrt[p]{\sum |x_i|^p} \tag{6}$$

where $p = 1$ gives the Manhattan distance, which was used by (Campbell et al., 1999) (Gentles and Karlin, 2001). $p = 2$ gives the Euclidean distance, and was used by (Nakashima et al., 1998). $p = \infty$ gives the max-norm (where only the largest absolute value contributes). The norms (6) give mathematically well defined distance functions for all positive values of p.

2.4 Relative entropy

In the following methods, the distance is computed directly from the count vectors. We use concepts from statistics, comparing the two hypothesis that i) two count vectors come from the same probability distribution or ii) they come from different probability distributions.

The relative entropy (or Kullback Leibler divergence) (MacKay, 2003) is defined as:

$$d(a, b) = \sum p_i \log \frac{p_i}{q_i} \tag{7}$$

where $p_i = a_i/n_a$ and $q_i = b_i/n_b$

It is sometimes called *KL distance*, even though it is not strictly a distance measure, since it is not symmetric (MacKay, 2003). However, we choose to include it for the completeness of the comparison. We also derived symmetric measures from the KL distance by computing both $d(a, b)$ and $d(b, a)$, and taking the max, min and sum of these, respectively.

2.5 Chi-square

Using the test statistic (χ^2) of the chi-square test of homogeneity (Rice, 1995) between a pair of k-nucleotide count vectors a and b, an element of the distance matrix is computed as:

$$\chi^2_{a,b} = \sum_{i=1}^{4^k} \left(\frac{(a_i - E\langle a_i \rangle)^2}{E\langle a_i \rangle} + \frac{(b_i - E\langle b_i \rangle)^2}{E\langle b_i \rangle} \right) \tag{8}$$

where 4^k is the length of the count vectors, and assuming a multinomial model for the counts a_i or b_i:

$$E\langle a_i \rangle = \frac{\left(\sum_{i=1}^{4^k} a_i\right)(a_i + b_i)}{\sum_{i=1}^{4^k} (a_i + b_i)} \tag{9}$$

A χ^2 value 0 corresponds to identical counts. The higher the value, the more likely it is that the distributions that generated a and b are different. χ^2 does satisfy all of the distance conditions except for the triangular rule. Since the value of $\chi^2_{a,b}$ is essentially a sum of squares of differences, we also include the square root of this measure.

2.6 Maximum likelihood

In the maximum likelihood method, we view the count vectors as generated by random independent events from a probability source. To get the distance between two count vectors a and b, we compare the likelihoods that they come from the same probability source as opposed to different ones. We compute these likelihoods by estimating the individual source probabilities by maximum likelihood. The maximum likelihood estimate of the probabilities is $p_i = a_i/n_a$ for a alone, $p_i = b_i/n_b$ for b alone and $p_i = (a_i + b_i)/(n_a + n_b)$ for a and b from the same distribution. The quotient of the likelihoods is then:

$$ML = \frac{ML_1(a) ML_1(b)}{ML_2(a,b)} \tag{10}$$

where

$$ML_1(a) = n_a! \prod \left(\frac{a_i}{n_a}\right)^{a_i} \frac{1}{a_i!} \tag{11}$$

and

$$ML_2(a,b) = n_a! n_b! \prod \left(\frac{a_i + b_i}{n_a + n_b}\right)^{a_i+b_i} \frac{1}{a_i! b_i!} \tag{12}$$

This quotient is ≥ 1 and it is equal to 1 when the two distributions are identical. It is customary to take the logarithm of such ratios as a score. We will use this logarithm as a distance function. It satisfies all the distance properties except for the triangular inequality, so strictly speaking it is not a distance function.

2.7 Least squares method for constructing phylogenetic trees

The least squares method computes a binary tree from a distance matrix. The tree is computed such that the weighted sum of the distances from node to node through the tree is minimal (weighted least squares matching of the given distance matrix).

The algorithm starts from a random tree (constructed by a randomized version of the neighbor joining algorithm) and iteratively improves the topology by 4-optim and

5-optim optimization. 4-optim is a node swapping heuristic that divides the trees into 4 subtrees, computes the quality for all 3 possible configurations of subtrees and selects the best configuration. 5-optim works in the same way, but divides the tree into 5 subtrees and evaluates all 15 possible configurations. The 4-/5-optim are applied iteratively on all possible nodes until the tree is no longer improved.

Since different initial random trees may generate different results, we ran the algorithm 1000 times, each with a different random tree, and chose the best result.

2.8 Tree quality index

The different methods are evaluated using a tree quality index in combination with visual inspection. The tree quality index is the least squares approximation error, suitably normalized. A drawback of this index is that it does not directly measure the similarity between our tree and the true tree, but instead measures how well the tree fits the pairwise distances given by the distance functions. However, since we usually do not know the true tree (topology and branch lengths), it is in general not possible to directly measure how well our tree corresponds to the true tree.

A good distance function will generate pairwise distances with few conflicts, which will give a low least squares approximation error. A bad distance function will typically generate pairwise distances with more conflicts, resulting in a high least squares approximation error. However, it is possible, theoretically, that a bad distance function produces erratic distances in a systematic way, resulting in a low least squares approximation error. To prevent that this causes problems in our analysis, we evaluate each distance function on a set of sequences for which the phylogenetic tree is known. The topological distance (Penny and Hendy, 1985) (Nei and Kumar, 2000) between trees constructed by each method and the true tree is computed. Only the functions that perform reasonably well are considered in the least squares error analysis.

3 Information vs error

It is clear that increasing the order k of the k-nucleotides increases the selectivity of the pattern, and hence the distances should be more accurate. So in principle, the higher the k value, the more information we capture and the more reliable the results should be. However, this is counterbalanced by the background noise of the distances, which is proportional to $O(4^k)$. In every component of the vector there is some expected error. In the case of the Poisson model, this error has variance 1. As we compute the distance between two vectors, these errors accumulate, and because of the exponential size of the vectors, the error becomes very significant for $k > 3$. The additional noise, which acts as an added value to each distance, is reflected in the trees themselves which become more corona-shaped (Figure 2). Furthermore, the additional noise, which makes all the distances larger by an additive constant, makes the tree fitting by least squares numerically better. This is quite disconcerting, as a better fit means worse trees.

[Figure 2 about here.]

As a measure of signal to noise ratio, we use:

$$s/n = \frac{v(real)}{v(rand)} \tag{13}$$

where v is defined as the *within row variance* of the k-nucleotide matrix:

$$v = \frac{1}{4^k g - g} \sum_{i=1}^{4^k} \sum_{j=1}^{g} (x_{ij} - \overline{x}_i)^2 \tag{14}$$

where g is the number of columns, 4^k is the number of rows, x_{ij} is the element in row i and column j, and $\overline{x}_i$ is the mean of row i. $v(rand)$ is the variance of random data and $v(real)$ is the variance of real data.

The s/n measure follows the F-distribution, so we can easily obtain the significance level.

4 Results

The first notable result is that the methods split themselves consistently into two classes, 2.2.2, 2.2.3, 2.2.4 and 2.2.5 which are very good, and all the rest which are very poor. In this version of the paper we explore further the good methods only.

4.1 RNA viruses

4.1.1 Information vs error

After comparing phylogenetic trees based on di-, tri- and tetra- and pentanucleotides, we choose the trinucleotide results as the best compromise between information and noise. For the pentanucleotide composition, with $4^5 = 1024$ different combinations on genomes of average length 15000 nucleotides, the count matrix became too sparse. Several pentanucleotides did not occur a single time in the genomes. Also, in the case of tetranucleotides, one position in the count matrix was empty for a few organisms, and the signal-to-noise ratio was quite low (Table 1). The s/n of trinucleotides is 7.1, which is significant with $p < 0.01$.

[Table 1 about here.]

4.1.2 Tree topology analysis

All the methods (2.2 - 2.6) were used to create phylogenetic trees from the trinucleotides of the RNA viruses. In order to compare the different methods, the topological distance (Penny and Hendy, 1985) (Nei and Kumar, 2000) between pair of trees constructed by the different methods is computed (Table 2). The first four methods (2.2.2, 2.2.3, 2.2.4 and 2.2.5) produce trees that are quite similar to each other (Table 2), and they also correspond to the known classification of the RNA viruses. The trees constructed by the other methods deviated significantly from the first four ones and from the known classification. As a consequence, we will focus on those four methods that produced the best trees.

[Table 2 about here.]

[Figure 3 about here.]

4.1.3 Creating distances

The Euclidean distance ($p = 2$) produced the best trees (based on trinucleotides, Table 3). In all cases, $p = 3$ gave essentially similarly good results. We prefer $p = 2$ as it is a well understood norm.

4.1.4 Tree quality index

The quality index is the least squares approximation error, suitably normalized.

Comparing the trinucleotide results from the Euclidean distances of the three methods (Table 3), the odds ratios and odds differences produced the best results (quality index 0.0085). Within a fixed set of sequences and a fixed k, it makes sense to use the error of the least squares approximation as a measure of quality. For a perfect distance function and for perfect data, this error would be zero. In Table 3 this means that we can select the best method by looking for the smallest error on each column.

[Table 3 about here.]

4.1.5 Classification of the SARS virus

The coronavirus group was always clustered together in the RNA virus tree for all combinations of different k-nucleotides and methods. However, the relative topology of the coronaviruses differed slightly from tree to tree. To obtain a more reliable tree for this virus group, we repeated the analysis for the 7 coronaviruses alone. The resulting trees all had the same topology, no matter which k-nucleotide composition or method was used. We only show the tree from the trinucleotide composition using the odds ratio bias function (Figure 4). The SARS virus is classified to the group I coronaviruses. This last result is the main contribution of the paper (Tobler et al., 2003) and differs from the original classification of SARS.

[Figure 4 about here.]

4.2 Known tree of 11 vertebrates

As a second example, we study the mitochondrial genome (sense strand) of 11 vertebrates for which the phylogenetic tree is already known (Figure 5) (Russo et al., 1996).

We choose dinucleotides as the best compromise between information and noise. The s/n of dinucleotides is 3.6, which is significant with $p < 0.05$.

Topological distances between trees created using the different methods revealed that the trees based on odds ratios, logarithm of odds ratios, Poisson distributions and odds differences were similar to the known tree (Table 4). The other methods produced trees that were significantly different from the known tree. The trees produced by the first four methods were also similar to each other, like in the case of RNA viruses. This indicates that these four methods are generally the best ones, and that they perform similarly.

When comparing the quality index of trees, the best results were produced by the odds ratio method (Table 5), like in the case of the RNA viruses. Also, the Euclidean distance ($p = 2$) produced better results than the other distances.

[Table 4 about here.]

Since the mitochondrial DNA is double stranded, we also studied the sense strand concatenated with its reverse complement as suggested in (Karlin and Burge, 1995). However, the resulting trees were of lower quality than those based on one sense strand only.

[Figure 5 about here.]

[Table 5 about here.]

5 Discussion

The phylogenetic trees of the RNA viruses (Figure 3) and the coronaviruses (Figure 4) combined with the trees from the 11 vertebrates indicate that building trees from k-nucleotides is a good alternative to the traditional approach of building trees from multiple sequence alignments. Compared to the MSA approach, our method has the advantage that it considers whole genomes, using all the information available, and it can compare genomes of different lengths without problems. Also, it does not suffer from gene rearrangements, which are common in viruses and cause havoc to MSA methods. MSA methods should give better results in general, but when the MSAs are difficult to obtain or of dubious quality, our method performs better.

We have formalized the method of going from k-nucleotide counts to phylogenetic trees (Figure 1). Defining the distance matrix as a fixed intermediate step, we explored an extensive number of methods of arriving at this distance matrix from k-nucleotide counts. By consistently evaluating the methods using a tree quality index and s/n ratio, we found out that the odds ratio method works best. We believe that our results will have significance for future phylogenetic analysis based on k-nucleotide frequencies.

From the phylogenetic trees of the 11 vertebrates and the 49 RNA viruses, we see that the first four methods (odds ratios, logarithm of odds ratios, Poisson distributions and odds differences) perform similarly. The best method in general considering both quality index and topological distance is the Euclidean distance ($p = 2$) of odds ratios. For double stranded DNA, we suggest to study the sense strand only, since the tree quality degraded when we included the reverse complement. Whether to study dinucleotides or trinucleotides depends on the lengths of the sequences. Because the error increases exponentially with increasing dimension k of the k-nucleotides, few applications would benefit from studying the tetra- or higher dimensions.

MSA based classification of SARS indicates that it belongs to a new group of coronaviruses (Marra et al., 2003) (Rota et al., 2003) or to the group II coronaviruses (Snijder et al., 2003). Here we suggest an alternative classification. The closest neighbors of SARS in the coronavirus tree (Figure 4) are HCoV-229E, PEDV and TGEV, which all belong to group I of coronaviruses. From this tree, we draw the conclusion that SARS also belongs to this group. SARS is in our tree closest to TGEV, so we hypothesize a close relationship to the TGEV virus. Notably, this classification is supported by experimental evidence, which showed that group I coronaviruses (transmissible gastroenteritis virus, TGEV) specific antibodies were able to recognize antigens in SARS-CoV infected cultured cells (Ksiazek et al., 2003).

References

Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T., and Ikemura, T. (2002). A novel bioinformatic strategy for unveiling hidden genome signatures of eukaryotes: self-organizing map of oligonucleotide frequency. *Genome Inform Ser Workshop Genome Inform*, 13:12–20.

Brocchieri, L. (2001). Phylogenetic inferences from molecular sequences: review and critique. *Theor Popul Biol*, 59:27–40.

Campbell, A., Mrazek, J., and Karlin, S. (1999). Genome signature comparisons among prokaryote, plasmid, and mitochondrial dna. *Proc Natl Acad Sci U S A*, 96:9184–9189.

Chuzhanova, N. A., Jones, A. J., and Margetts, S. (1998). Feature selection for genetic sequence classification. *Bioinformatics*, 14:139–143.

Friberg, M. (2003). Virus classification using k-nucleotide frequencies. Technical Report 401, Informatik, ETH, Zurich, http://cbrg.inf.ethz.ch/bio-recipes/VirusClassification/code.html.

Gentles, A. J. and Karlin, S. (2001). Genome-scale compositional comparisons in eukaryotes. *Genome Res*, 11:540–546.

Gonnet, G. H., Hallett, M. T., Korostensky, C., and Bernardin, L. (2000). Darwin v. 2.0: an interpreted computer language for the biosciences. *Bioinformatics*, 16:101–103.

Karlin, S. and Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet*, 11:283–290.

Karlin, S. and Ladunga, I. (1994). Comparisons of eukaryotic genomic sequences. *Proc Natl Acad Sci U S A*, 91:12832–12836.

Ksiazek, T. G., Erdman, D., Goldsmith, C. S., Zaki, S. R., Peret, T., Emery, S., Tong, S., Urbani, C., Comer, J. A., Lim, W., Rollin, P. E., Dowell, S. F., Ling, A. E., Humphrey, C. D., Shieh, W. J., Guarner, J., Paddock, C. D., Rota, P., Fields, B., DeRisi, J., Yang, J. Y., Cox, N., Hughes, J. M., LeDuc, J. W., Bellini, W. J., and Anderson, L. J. (2003). A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med*, 348:1953–1966.

MacKay, D. J. C. (2003). Textbook: Information theory, inference, and learning algorithms. *Cambridge University Press*.

Marra, M. A., Jones, S. J., Astell, C. R., Holt, R. A., Brooks-Wilson, A., Butterfield, Y. S., Khattra, J., Asano, J. K., Barber, S. A., Chan, S. Y., Cloutier, A., Coughlin, S. M., Freeman, D., Girn, N., Griffith, O. L., Leach, S. R., Mayo, M., McDonald, H., Montgomery, S. B., Pandoh, P. K., Petrescu, A. S., Robertson, A. G., Schein, J. E., Siddiqui, A., Smailus, D. E., Stott, J. M., Yang, G. S., Plummer, F., Andonov, A., Artsob, H., Bastien, N., Bernard, K., Booth, T. F., Bowness, D., Czub, M., Drebot, M., Fernando, L., Flick, R., Garbutt, M., Gray, M., Grolla, A., Jones, S., Feldmann, H., Meyers, A., Kabani, A., Li, Y., Normand, S., Stroher, U., Tipples, G. A., Tyler, S., Vogrig, R., Ward, D., Watson, B., Brunham, R. C., Krajden, M., Petric, M., Skowronski, D. M., Upton, C., and Roper, R. L. (2003). The genome sequence of the sars-associated coronavirus. *Science*, 300:1399–1404.

Nakashima, H., Nishikawa, K., and Ooi, T. (1997). Differences in dinucleotide frequencies of human, yeast, and escherichia coli genes. *DNA Res*, 4:185–192.

Nakashima, H., Ota, M., Nishikawa, K., and Ooi, T. (1998). Genes from nine genomes are separated into their organisms in the dinucleotide composition space. *DNA Res*, 5:251–259.

Nei, M. and Kumar, S. (2000). Molecular evolution and phylogenetics. *Oxford University Press*, 1st ed.

Nussinov, R. (1984). Strong doublet preferences in nucleotide sequences and dna geometry. *J Mol Evol*, 20:111–119.

Penny, D. and Hendy, M. D. (1985). The use of tree comparison methods. *Syst. Zool.*, 34:75–82.

Rice, J. A. (1995). Mathematical statistics and data analysis. *Duxbury Press*, 2nd ed.

Rota, P. A., Oberste, M. S., Monroe, S. S., Nix, W. A., Campagnoli, R., Icenogle, J. P., Penaranda, S., Bankamp, B., Maher, K., Chen, M. H., Tong, S., Tamin, A., Lowe, L., Frace, M., DeRisi, J. L., Chen, Q., Wang, D., Erdman, D. D., Peret, T. C., Burns, C., Ksiazek, T. G., Rollin, P. E., Sanchez, A., Liffick, S., Holloway, B., Limor, J., McCaustland, K., Olsen-Rasmussen, M., Fouchier, R., Gunther, S., Osterhaus,

A. D., Drosten, C., Pallansch, M. A., Anderson, L. J., and Bellini, W. J. (2003). Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science*, 300:1394–1399.

Russo, C. A., Takezaki, N., and Nei, M. (1996). Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol Biol Evol*, 13:525–536.

Snijder, E. J., Bredenbeek, P. J., Dobbe, J. C., Thiel, V., Ziebuhr, J., Poon, L. L., Guan, Y., Rozanov, M., Spaan, W. J., and Gorbalenya, A. E. (2003). Unique and conserved features of genome and proteome of sars-coronavirus, an early split-off from the coronavirus group 2 lineage. *J Mol Biol*, 331:991–004.

Tobler, K., Friberg, M., Gonnet, G. H., and Ackermann, M. (2003). K-nucleotide frequency distances indicate the positioning of the sars-coronavirus among type i coronaviruses. *submitted*.

List of Figures

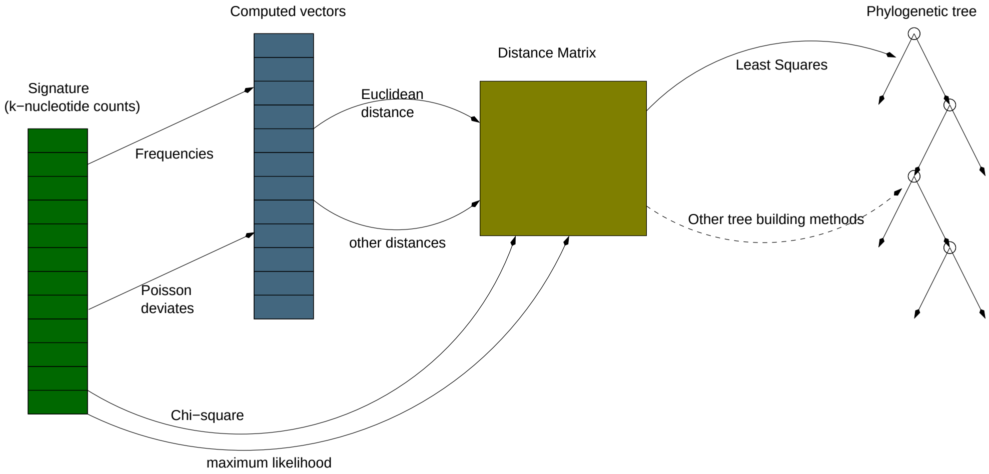

Figure 1: Alternative ways of computing a phylogenetic tree from k-nucleotide counts

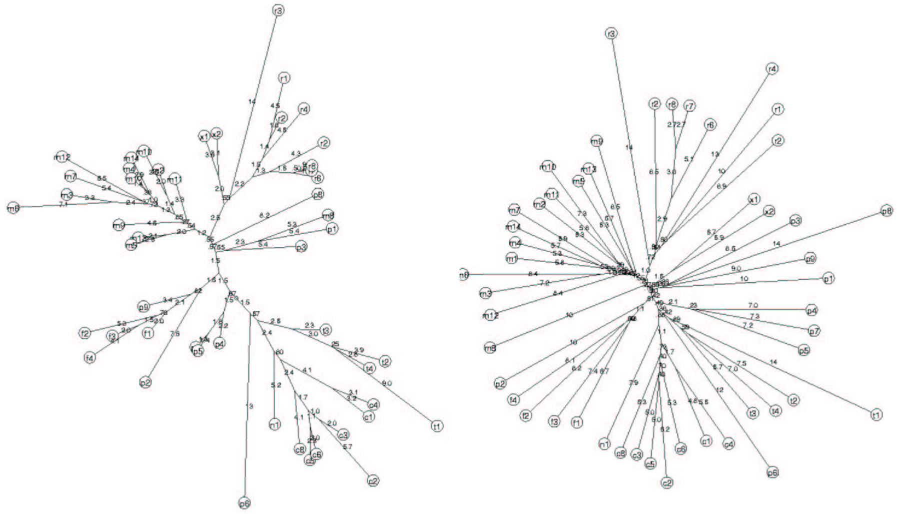

Figure 2: Trees of RNA viruses based on dinucleotides (left) and pentanucleotides (right). The additive noise makes the right tree corona-like.

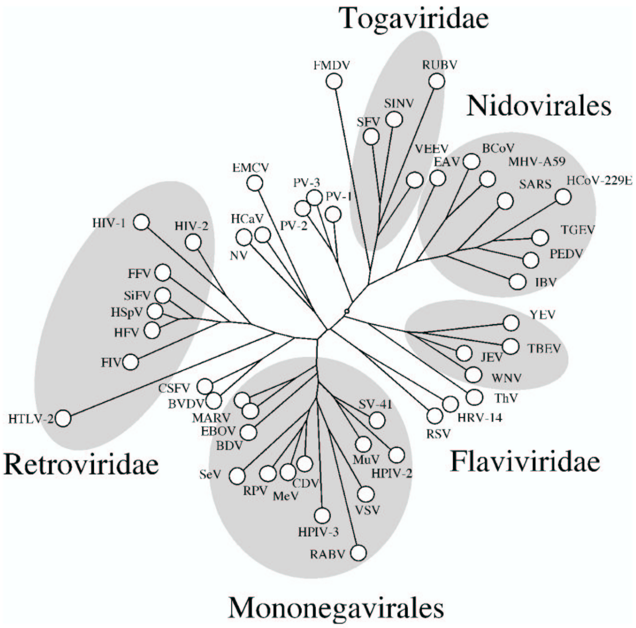

Figure 3: Unrooted tree of 49 RNA Viruses based on the Euclidean distance of odds ratios of trinucleotides.

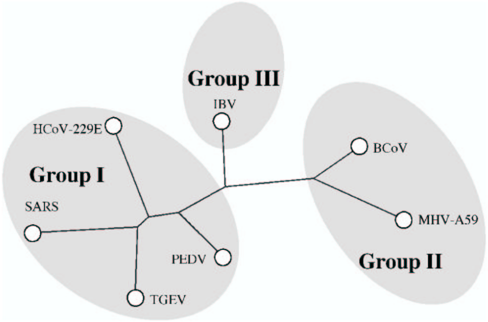

Figure 4: Unrooted tree of 7 coronaviruses based on the Euclidean distance of odds ratios of trinucleotides.

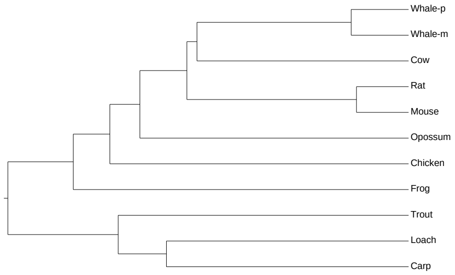

Figure 5: Phylogram of the 11 vertebrates

List of Tables

di-nuc	tri-nuc	tetra-nuc	penta-nuc
15.5	7.1	3.3	1.9

Table 1: Signal-to-noise ratios of the 49 RNA viruses for different k-nucleotides

	OddsRatio	LogOdds	Poisson	OddsDiff	RelEntropy	ChiSq	MLDist
OddsRatio		54	48	46	76	76	84
LogOdds			64	72	80	78	88
Poisson				64	82	82	90
OddsDiff					84	84	84
RelEntropy						48	70
ChiSq							50
MLDist							

Table 2: Topological distance between RNA virus trees constructed using different methods. The first four methods produce trees that are similar to each other (distances between 46 and 64). The last three methods deviate significantly from the first four ones.

Method	Dist func?	di-nuc	tri-nuc	tetra-nuc
OddsRatio p=1	Yes	0.0173	0.0104	0.0057
OddsRatio p=2	Yes	0.0153	*0.0085*	0.0046
OddsRatio p=3	Yes	0.0171	0.0085	0.0048
OddsRatio p=inf	Yes	0.0313	0.0196	0.0170
LogOddsRatio p=1	Yes	0.0189	0.0122	0.0066
LogOddsRatio p=2	Yes	0.0201	*0.0120*	0.0063
LogOddsRatio p=3	Yes	0.0244	0.0124	0.0065
LogOddsRatio p=inf	Yes	0.0366	0.0224	0.0214
PoissonDistr p=1	Yes	0.0179	0.0112	0.0070
PoissonDistr p=2	Yes	0.0159	*0.0105*	0.0060
PoissonDistr p=3	Yes	0.0178	0.0107	0.0060
PoissonDistr p=inf	Yes	0.0309	0.0174	0.0138
OddsDiff p=1	Yes	0.0185	0.0100	0.0050
OddsDiff p=2	Yes	0.0167	*0.0085*	0.0044
OddsDiff p=3	Yes	0.0178	0.0085	0.0046
OddsDiff p=inf	Yes	0.0303	0.0185	0.0147

Table 3: Quality index from tree construction of 49 RNA viruses. The best result from each method, based on quality index together with s/n ratio, is marked in italics.

	known	OddsRatio	LogOdds	Poisson	OddsDiff	RelEntropy	ChiSq	MLDist
known		8	8	8	6	12	12	16
OddsRatio			2	2	4	12	12	16
LogOdds				0	4	10	10	14
Poisson					4	10	10	14
OddsDiff						12	12	16
RelEntropy							0	4
ChiSq								4
MLDist								

Table 4: Topological distance between trees of the 11 vertebrates, constructed using different methods. The known tree is also included.

Method	Dist func?	di-nuc	tri-nuc	tetra-nuc
OddsRatio p=1	Yes	0.0080	0.0057	0.0021
OddsRatio p=2	Yes	*0.0052*	0.0042	0.0048
OddsRatio p=3	Yes	0.0071	0.0077	0.0118
OddsRatio p=inf	Yes	0.0188	0.0307	0.0313
LogOddsRatio p=1	Yes	0.0085	0.0060	0.0025
LogOddsRatio p=2	Yes	*0.0062*	0.0040	0.0019
LogOddsRatio p=3	Yes	0.0065	0.0044	0.0030
LogOddsRatio p=inf	Yes	0.0116	0.0185	0.0224
PoissonDistr p=1	Yes	0.0086	0.0059	0.0022
PoissonDistr p=2	Yes	*0.0053*	0.0044	0.0052
PoissonDistr p=3	Yes	0.0067	0.0080	0.0123
PoissonDistr p=inf	Yes	0.0166	0.0313	0.0326
OddsDiff p=2	Yes	*0.0102*	0.0105	0.0068
RelativeEntropy min	No triang.	*0.0113*	0.0074	0.0046
Chi-Sq sqrt	No triang.	*0.0088*	0.0084	0.0060
ML	No triang.	*0.1238*	0.0785	0.0398

Table 5: Quality index from tree construction of 11 vertebrates. The best result from each method, based on quality index and s/n ratio, is marked in italics.