Simple methods of finding short protein coding sequences using multiple species alignments

Hanno Hinsch[1,] and Artemis Hatzigeorgiou[1,2]*

[1]Center for Bioinformatics, Department of Genetics, School of Medicine; [2]Computer and Information Science, School of Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA

ABSTRACT

Eukaryotic genomes contain many conserved regions of unknown function. Accurately assessing the protein coding potential of these regions is a key step in annotation. We develop three protein coding measures that directly assess conserved regions in multiple sequence alignments of many species: one based on phase-shifts induced by alignment gaps, another based on the 3rd position mutation asymmetry in codons, and a third based on nucleotide composition asymmetry. The methods are easy to implement and require no training. Using a human-chimp-rat-mouse-chicken multiple alignment, these measures can classify coding regions as short as 30nt with greater specificity than single-genome measures using 120nt. Results from human-mouse and human-chicken alignments can be further improved by considering additional species; only the chimp genome proved uninformative. The phase-shift method is especially accurate.

Contact: agh@pcbi.upenn.edu, hannoh@seas.upenn.edu

1 INTRODUCTION

The genomes of many organisms, including humans, have a large number of highly conserved regions whose function is not readily apparent (Margulies 2003). A standard first step is to assess the possibility that the conserved region is in fact an unknown exon. Thus a contemporary geneticist is frequently faced with the problem of determining the coding potential of conserved regions.

A variety of tools and algorithms are available to aid in this determination. They vary in several aspects, but can be roughly classified by the nature of the input data and the assumed gene structure, and the prediction biases these two characteristics imply.

Ab initio gene prediction programs, eg GenScan (Burge 1997), require only a single genome but derive their statistical power in part by modeling the composition and lengths of exons and introns. This introduces the possibility of bias against short exons, and genes comprised of either fewer or more exons than typical. Alternative or suboptimal gene structures are typically not predicted by these programs.

Several newer ab initio gene finders, eg Twinscan (Korf 2001) and SGP2 (Parra 2003) use a second genome to further improve prediction accuracy, and Exoniphy (Siepel 2004) uses multiple genomes. However, like GenScan, they model the lengths and distributions of exons and thus share the presumed bias against atypical gene structure. For a recent review of ab initio gene finders, including a discussion of biases, see (Brent 2004).

Alternative gene finding strategies based on cDNA and EST evidence, for example GenomeWise (Birney 2004), can not detect modestly or selectively expressed genes for whom evidence is simply unavailable, although they are quite successful at finding very short exons (Volfovsky 2004) when evidence permits.

Many model-free single-genome methods have been developed; see (Fickett 1992) for a classic evaluation and review. Unlike methods that fully elucidate a gene structure, these methods directly estimate sequence protein coding potential without reference to any particular gene model, and should therefore show no structural bias. This theoretical advantage, though, has been accompanied by a lack of statistical power, as the absence of a gene model leads to more false positive predictions.

The appeal of model-free multiple-genome methods, as considered in this study, is simple: if they can provide the discriminative power of the full gene finders without introducing structural bias, they will allow the geneticist to more accurately determine the function of short conserved regions. As the number of available genomes and alignments continues to increase, this is becoming ever more desirable. To this end several two-genome methods have been presented (Nekrutenko

* To whom correspondence should be addressed.

2003, Mignone 2003). These studies readily confirm the advantage of using an additional genome to estimate coding potential.

We are not, however, aware of any studies that generalize such measures to more than two genomes, or assess the value of doing so.

In this study we developed three model-free methods to estimate the protein coding potential directly from multiple sequence alignments of more than two species. Our results show that these new measures are substantially more discriminative than those using only one or two species, thus allowing the accurate classification of significantly shorter conserved coding regions than currently possible.

2 METHODS

We developed and assessed three measures that predict protein coding potential from an n-way multiple alignment, where n can be arbitrarily large.

2.1 Phase shift measure

A Thymine in the first position of a codon does not have the same biological meaning as a Thymine in the second position of a codon, thus our phase shift measure is based on the observation that if aligned nucleotide sequences translate into even remotely similar proteins, then a majority of the aligned nucleotides must appear in the same phase. The measure calculates the percentage of nucleotides in the alignment that are phase-shifted with respect to the reference sequence. Such phase shifts are of course induced by alignment gaps, whose predictive power has been exploited in earlier measures. See (Kellis 2003) for an application of pairwise alignments in the context of gene discovery in yeast, and also prediction programs based on extended hidden Markov models (HMMs) that either explicitly (Noguchi 2004) or implicitly (Siepel 2004) use gap information as an input. Our measure generalizes to multiple genomes, is simple to implement, and, unlike the HMM-based methods, requires no training.

We calculate the score by:

(1) Calculating the percentage of nucleotides in the alignment that are "unshifted", where

(2) A nucleotide is unshifted if the number of gaps preceding the nucleotide in its sequence is the same as the number of preceding gaps in the reference sequence (modulus 3).

True protein coding sequences have a very high percentage of "unshifted" nucleotides.

2.2 Composition asymmetry measure

Protein coding sequences show a marked compositional asymmetry – there is not the same number of A, T, C, or G's in each of the three codon positions. Our multi-genome composition asymmetry measure generalizes a similar single-genome measure (see Fickett 1992 for a discussion of several such composition measures) by summing the A, T, C, and G's in each codon position in a window over all aligned sequences, and then calculating a Pearson chi square statistic of homogeneity from the resulting table of counts.

More precisely, we calculate the score by:

(1) Creating a 4 by 3 table of counts in which each row is a nucleotide (A, T, C, G) and each column a phase (0, 1, 2).

(2) Filling the table by iterating through all nucleotides in the alignment, incrementing the appropriate table cell count for each nucleotide.

(3) Calculating the Pearson chi-square statistic of homogeneity for the completed table.

True protein coding sequences will tend to have an asymmetric composition, hence a high score.

2.3 Mutational asymmetry measure

Finally, our mutational asymmetry measure is informed by the relative preponderance of mutations in the third nucleotide position (the "wobble" position) of codons in protein coding sequences. The number of mutations in each position is calculated, then an F-statistic is calculated to estimate the probability that the three positions are equally likely to mutate.

There are several different ways to calculate the number of mutations at a site for this mutational asymmetry measure. Counting the number of different nucleotides that appear at a site is one, another is to count the number of sequences in which the nucleotide differs from the reference sequence. If the phylogenetic tree is known, a third possibility is to calculate the parsimony score of the site; this score is the minimum lower bound on the number of mutations that could explain the observed pattern of nucleotides (Fitch 1971).

We calculate the mutational asymmetry score by:

(1) Calculating the mutation count of each column in the alignment (either the "raw" total number of nucleotides that differ from the reference, or the parsimony score).

(2) Partitioning the columns into three groups (phase 0, 1, 2).

(3) Calculating a conventional F-statistic to determine if the three groups share a common mean mutation rate.

True protein coding sequences tend to have more mutations in the third phase, and thus a higher F-statistic.

2.4 Single genome measures

For the sake of comparison, we also assessed the effectiveness of two single genome measures: the composition asymmetry measure calculated from only the human genome and a boolean open reading frame score.

2.5 Assessment

Developing a meaningful assessment of any gene prediction tool is a challenge, and assessing these measures is no exception. On the one hand lies the imperative to evaluate the tool within the context in which it will ultimately be used, on the other hand the necessity of working with test data whose nature is fully understood.

As we hope that these measures will prove useful in elucidating conserved regions even within or around known or predicted genes, presumably those in which the nature of short conserved elements is not apparent, we chose to evaluate how well the measures distinguish coding sequence from adjacent intronic and untranslated sequence. We consider this more stringent than a comparison against intergenic or randomized sequence.

The test procedure used the UCSC July 2003 5-way Multiz (Blanchette 2004) alignment (human hg16, chimp panTro1, mouse mm3, rat rn3, chicken galGal2 assemblies) and the corresponding RefSeq genes for human chromosome 21 (http://genome.ucsc.edu). We chose chromosome 21 because we believe it to be relatively well annotated.

We first selected regions that were aligned in all five species and also overlapped the coding/non-coding junction (splice site, start codon, or stop codon) of a RefSeq exon. From each 5' junction we extracted two samples: a 5' noncoding window directly upstream of the junction, and a coding sequence window directly downstream. Similar samples were extracted from the 3' junctions. This balanced set of aligned sequence data abutting the junctions formed our basic test set.

Using this basic set, we assessed the measures using various window sizes from 15nt to 120nt, and various species subsets including human-mouse, human-chicken, human-mouse-chicken, and others.

This scheme gave us high confidence that the coding and non-coding labels were accurately assigned. Unfortunately, the region abutting a junction is probably not representative of aligned non-coding sequence in general. To assess the impact of comparing regions abutting junctions, which are known to be very highly conserved, we ran additional tests which excluded the junctions. These tests compared a non-coding sequence window offset away from the junction with a coding sequence window also offset away from the junction.

Although any number of test statistics could arguably be used to evaluate prediction measures, historically the usefulness of these measures has been limited by their lack of specificity – the large number of false predictions. We chose therefore to calculate each measure's rate of false positives in various situations. The false positive (FP) rate is defined as the number of false positives divided by the sum of the false positives and true negatives; interpret it as the frequency with which a non-coding region will falsely be judged to be coding.

We evaluated the performance of each measure independently using logistic regression in the JMP-IN statistical package (SAS Institute). The false positive rate corresponding to a sensitivity of 90% was observed directly from the receiver operating characteristics table calculated by the software.

3 RESULTS

3.1 Phase shift measure

The phase shift measure was the unequivocal champion under all circumstances excepting human-chimp alignments. This very simple measure, requiring no training set and easily implemented in only a few lines of code, has a false positive rate of 1.28% in a 5-species window of 60nt (Table 1)

Table 1. Effect of window size on FP rate

Measure	120 nt	90 nt	60 nt	30 nt	15 nt
Phase shift	0.000	0.003	1.28	8.42	31
Composition asymmetry	7.19	17.55	17.34	46.48	59.75
Mutation asymmetry (parsimony)	0.60	2.98	15.09	49.5	73.00
Mutation asymmetry (raw)	1.2	5.63	19.42	40.44	69.24
Composition asymmetry (human only)	55.69	75.17	66.45	80.60	87
ORF yes/no (human only)	59.88	69.54	92.3	100	100

FP rate using 5-species alignments of 15nt to 120nt. Shows FP rate at 90% sensitivity on human chr21 alignments for all 5 species and indicated window lengths. FP rate = 100 * (FP / (FP+TN)) A False Positive rate of 20% means that for every 100 negative samples, 20 would be incorrectly identified as positive. The ORF test has 100% sensitivity (all true CDS regions have an open reading frame). Some values were interpolated.

Its specificity increased to 0.49% (using a 60nt window) as we sampled farther away from junctions. Using a window of 120nt it reported no false positives. Using only human-mouse alignment it achieved a rate of 8.35%; this dropped to 1.44% when chicken was added.

In comparison to the single genome composition asymmetry measure, which achieved an FP rate of 55.69% using a window of 120nt, the phase shift measure was better even using a window less than one tenth the length (15nt -- a short exon) with an FP rate of 31%.

The biological basis for this performance is of course clear – frame shifts matter – yet the degree of specificity still surprised us. It seems alignments of non-coding sequences almost invariably show at least one phase-shifting gap some place or another, and that single gap alone is sufficient information to effectively distinguish between the two classes of sequence.

Table 2 Effect of including additional species on FP rate

Measure	Hg-mm-gal-rn-pan	Hg-mm-gal-rn	Hg-mm-gal	Hg-gal	Hg-mm	Hg-pan	Mixed
Phase shift	1.28	1.44	1.44	3.53	8.35	87.96	10.73
Composition asymmetry	17.34	10.91	23.60	45.59	31.14	64.69	37.65
Mutation asymmetry (parsimony)	15.09	16.05	23.76	37.60	59.71	75.00	53.17
Mutation asymmetry (raw)	19.42	23.25	21.99	37.60	59.71	75.00	59.95
Composition asymmetry (human only)	66.45	66.45	66.45	66.45	66.45	66.45	68.21
Orf yes/no (human only)	92.3	92.3	92.3	92.3	92.3	92.3	93.29

FP rate based on 60nt alignments of various species. Shows FP rate at sensitivity of 90% on human chr21 alignments using a window size of 60 bp and the indicated species. The "Mixed" column reflects performance on the full mixed-species set of human chromosome 21 alignments with no filtering. Some values were interpolated.

3.2 Mutation asymmetry measure

The results for the mutation asymmetry measure are not quite as good, but still far outstrip the single genome measures. Using a 60nt window of the 5-species alignment we saw a FP rate of 15.09%. This dropped to 0.6% using a window of 120nt. This measure is sensitive to the number of species in the alignment: when human-mouse is used instead of all 5 species, the FP rate jumps from 15.09% to 59.71%.

Counting mutations using parsimony improves results for certain tree topologies, such as those incorporating both mouse and chicken, presumably because it avoids double-counting on shared branches. The parsimony score may provide a more significant increase in power with other sets of species (ie, other phylogenetic trees). The addition of the chimpanzee genome was modestly helpful to this measure.

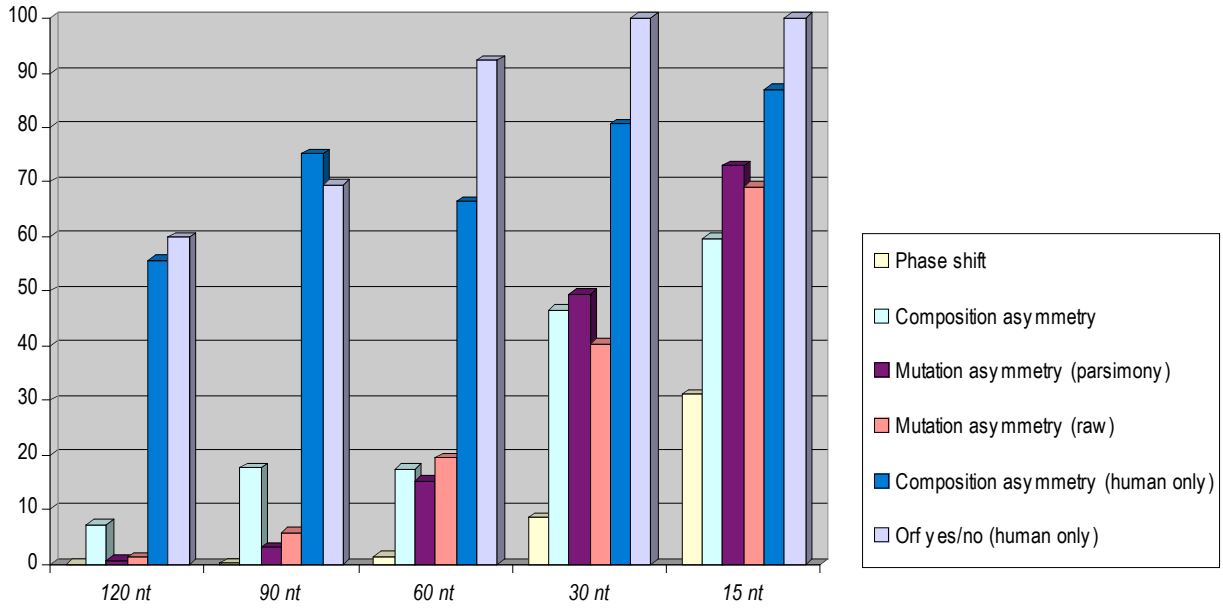

Fig 1 Effect of window size on FP rate

3.3 Composition asymmetry measure

The results for the generalized version of the composition asymmetry measure, as one might guess, show that its effectiveness is roughly proportional to the number of nucleotides considered. In the 5-species alignment, the FP rate varies from 7.19% using 120nt, to 17.34% using 60nt, and 46.48% using 30nt. Considering 60nt windows as species are added, the FP rate drops from 64.69% when only human-chimpanzee is considered, to 31.14% for human-mouse, to a more reasonable 10.91% for human-mouse-rat-chicken. The addition of the chimpanzee genome seems to merely confound matters – at 17.91%, the results for human-chimpanzee-chicken-rat-mouse are distinctly worse.

Table 3. Effect of junction proximity on FP rate

Measure	Offset 0	Offset +/-20	Offset +/-50
Phase shift	1.28	1.18	0.49
Composition asymmetry	17.34	8.49	15.53
Mutation asymmetry (parsimony)	15.09	9.91	8.25
Mutation asymmetry (raw)	19.42	10.14	10.68
Composition asymmetry (human only)	66.45	66.27	69.42
Orf yes/no (human only)	92.3	85.61	81.55

Effect of junction proximity on FP rate using the 5 species alignment with a window of 60nt. The power of all measures increases as the window is offset 20nt away from the coding/non-coding junctions. The difference between an offset of 20nt and 50nt is mixed, and less pronounced.

3.4 Effect of additional genomes

Since the composition asymmetry measure was also calculated for a single genome, we can see here the full value of incorporating additional genomes (Table 2). Using a 60nt window, an FP rate of 66.45% was recorded using human sequence only. This dropped to 31.14% using human-mouse, to 23.60% using human-mouse-chicken, and to a low of 10.91% using human-mouse-rat-chicken.

The mutation asymmetry measure improved markedly when a third species (chicken-mouse) was added, and also a fourth (rat) when using parsimony scores.

The phase shift measure reached a near optimum with the human-mouse-chicken set.

We conclude that generally more is better, even if two of the genomes are at roughly equal evolutionary distance (rat and mouse), except when the genome is chimpanzee (perhaps because it is very close, or possibly because of a difference in quality).

3.5 Effect of junction proximity

Finally, our test for the effect of junction proximity (Table 3) clearly shows that distinguishing the coding and noncod ing from one another at the exon junction is more difficult
than if the sample sequences are offset away from the junc tion. This is no surprise – splice junctions are known to be exceptionally well conserved. The noteworthy point is that our results for specificity, which are calculated from windows around the junction, are likely to be worst-case estimates easily bettered in actual use.

4 DISCUSSION

Direct measurement of protein coding potential remains

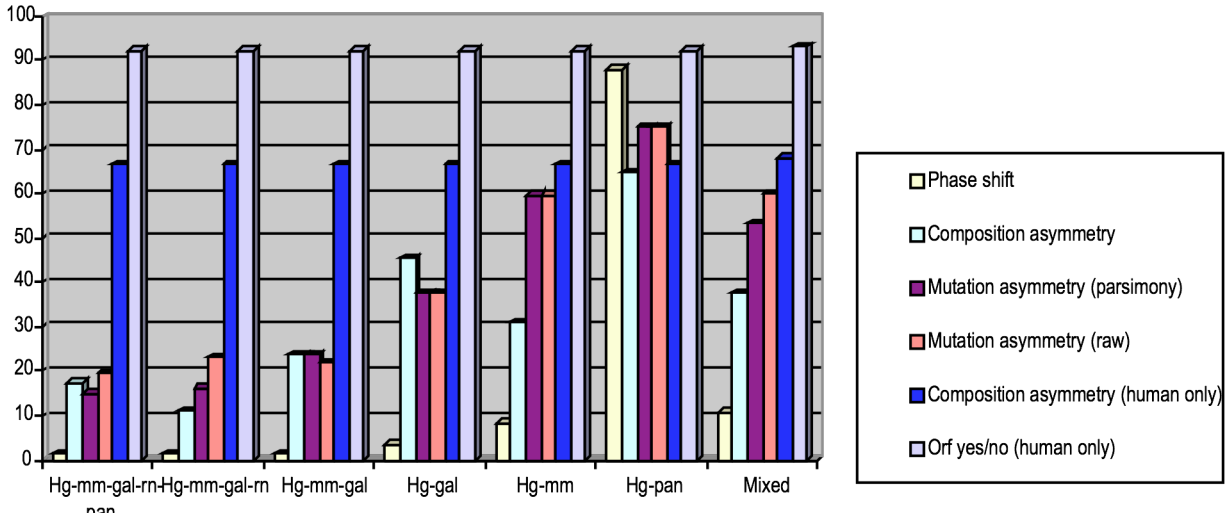

Fig 2 Effect of including additional species on FP rate

an invaluable adjunct to full gene finders, which develop much of their statistical power by making strong assumptions about the underlying gene structure. A raw predictor of coding potential, on the other hand, loses this power but provides an unbiased view of possible protein coding regions. When only one genome is available for use, experience has shown the tradeoff favors the complete gene finders – tools that examine only coding potential without an assumption of gene structure generate too many false positives.

We demonstrate the use of multiple genomes reduces the false positives of protein coding measures to much lower levels, which should increase their usefulness.

Multiple-genome methods derive their statistical power from the depth of evolutionary data at each nucleotide site, and thus need to examine far fewer sites for any given level of confidence. The search for short, structurally atypical elements such as nested genes, genes with short exons or introns, single exon genes, and genes with many equally optimal splice sites should benefit most from their use.

ACKNOWLEDGEMENTS

A.H. is partially supported by the NSF career award DBI-0238295.

REFERENCES

Birney, E., Clamp, M. & Durbin, R. (2004). GeneWise and Genomewise, *Genome Research*, **14**, 988-95.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., Haussler, D. & Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner, *Genome Research*. **14**, 708-15.

Brent, M.R. & Guigo, R. (2004). Recent advances in gene structure prediction, *Current Opinion in Structural Biology,* **14**, 264-72.

Burge, C. & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA, *Journal of Molecular Biology,* **268**, 78-94

Fickett, J.W. & Tung, C.S. (1992). Assessment of protein coding measure, *Nucleic Acids Research,* **20**, 6441-50.

Fitch, W.M. (1971). Toward defining the course of evolution: Minimum change for a specified tree topology, *Systematic Zooogy***, 20,** 406–416.

Kellis, et al. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements, Nature. 2003 May 15;**423**:241-54

Korf, I., Flicek, P., Duan, D. & Brent, M.R. (2001). Integrating genomic homology into gene structure prediction, *Bioinformatics,* **17 Suppl 1**, S140-8.

Margulies, E.H., Blanchette, M., Haussler, D. & Green, E.D. (2003). Identification and characterization of multi-species conserved sequences, *Genome Research*. **13**, 2507-18.

Mignone, F., Grillo, G., Liuni, S. & Pesole, G. (2003). Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis, *Nucleic Acids Research*, **31**, 4639-45.

Nekrutenko, A., Chung, W.Y. & Li, W.H. (2003). An evolutionary approach reveals a high protein-coding capacity of the human genome*, Trends in Genetics*, **19**, 306-10.

Noguchi, H., Yada, T. & Sakaki, Y.(2002). A novel index which precisely derives protein coding regions from cross-species genome alignments*, Genome Inform Ser Workshop Genome Inform*, **13**, 183-91.

Parra, G., Agarwal P., Abril, J.F., Wiehe, T., Fickett, J.W. & Guigo, R. (2003). Comparative gene prediction in human and mouse, *Genome Research*, **13**, 108-17.

Siepel, A. & Haussler, D. (2004). Computational identification of evolutionarily conserved exons, *J.RECOMB'04*, March 27–31, San Diego, California, USA.

Volfovsky, N., Haas, B.J. & Salzberg, S.L. (2003). Computational discovery of internal micro-exons, *Genome Research*, **13**, 1216-21.

5 APPENDIX A

Pseudocode for the three measures.

5.1 Method phase_shift_score()

```
(1)  numShifted= 0;
(2)  unShifted= 0;
(3)
(4)  refseq= first seq in alignment
(5)  for each further seq in alignment
(6)     refPhase= 0;
(7)     seqPhase= 0;
(8)
(9)     skip initial columns that are gaps in both seq
        and refseq
(10)
(11)    for each remaining column in alignment
(12)       if refseq[column] is a gap
(13)          //increment phase mod 3
(14)          refPhase= (refPhase+1)%3;
(15)
(16)       if seq[column] is a gap
(17)          //increment phase mod 3
(18)          seqPhase= (seqPhase+1)%3;
(19)
(20)       //count in-phase and out-of-phase columns
(21)       if seq[column] is NOT a gap
(22)          if seqPhase != refPhase
(23)             numShifted++;
```

```
(24)          else
(25)            unShifted++;
(26)
(27) // return percent unshifted
(28) return unShifted / (unShifted + numShifted);
```

5.2 Method composition_asymmetry_score()

```
(1)  create 4 by 3 contingency table of integers
(2)  //one row for each nucleotide (ATCG)
(3)  //one column for each phase (0,1,2)
(4)
(5)  for each sequence in alignment
(6)     phase= 0;
(7)     for each column in sequence
(8)        if seq[column] is not a gap
(9)           table[ character at seq[column], phase]++;
(10)          phase= (phase+1)%3;
(11)
(12) return Pearson chiSquare value of table;
```

5.3 Method mutation_asymmetry_score()

```
(1)  //given an integer vector of mutation counts
(2)  //(either raw or parsimony) where each element is the
(3)  //score of the corresponding column of the alignment
(4)
(5)  SSwithin=0;
(6)  MSwithin=0;
(7)  SSbetween=0;
(8)  MSbetween=0;
(9)
(10) //overall and phase means
(11) overallMean=  totalMutations /totalColumns;
(12) means[0] = phase 0 mutations /phase 0 columns;
(13) means[1] = phase 1 mutations /phase 1 columns;
(14) means[2] = phase 2 mutations /phase 2 columns;
(15)
(16) //SSwithin
(17) phase= 0;
(18) for each element in mutation vector
(19)    SSwithin+= square( element - means[phase] );
(20)    phase = (phase + 1) % 3;
(21)
(22) MSwithin= SSwithin / (totalColumns - 3);
(23)
(24) //SSbetween
(25) for each phase
(26)    SSbetween+= (number of columns in phase) *
(27)                square (means[phase] – overallMean );
(28)
(29) MSbetween= SSbetween / (3 - 1);
(30)
(31) if( MSwithin == 0 )
(32)    return 0;
(33) else
(34)    return ( MSbetween / MSwithin );
```