Identifying Chimerism in Proteins Using Hidden Markov Models of Codon Usage

Lawrence Hunter
National Library of Medicine
Building 38A, 9th Floor
Bethesda MD 20894 USA
email: hunter@nlm.nih.gov

Barry Zeeberg
4378 North Pershing Drive #1
Arlington VA 22203 USA
phone: (703) 525-7036
email: brz@gwis2.circ.gwu.edu

Abstract

Protein chimerism is a phenomenon involving the combination of multiple ancestral sequences into a single, multi-domain protein through evolution. We propose a novel method for detecting chimeric proteins by analyzing their nucleotide sequence. The method tests for differences in the distributions of synonymous (isoaccepting) codons in different regions of the protein. The test involves the comparison of the ability of varying size hidden Markov models (HMMs) of codon usage to fit the natural sequence, relative to a set of randomized controls. We demonstrate the method on the families of yeast nuclear and mitochondrial amino-acyl tRNA synthetases. The method is potentially useful for the automated screening of entire genomes or large databases.

Introduction

In this paper, we present a novel method for testing evolutionary chimerism, and apply the method to a demonstration case. Our example is the characterization of an ancient family of proteins known to be chimeric, the tRNA synthetases (Burbaum and Schimmel 1991).

Chimerism is the result of an evolutionary event which brings together two previously independent proteins or protein domains to form a new, combined protein. Many proteins appear to be composed of multiple domains, and may hence be chimeric. For example, Li and Graur (1991) discuss tissue plasminogen activator (TPA), which consists of four domains apparently homologous to separate domains from fibronectin, epidermal growth factor, trypsin and plasminogen (Patthy 1985). In TPA, the domain boundaries coincide with boundaries between introns and exons, suggesting a possible evolutionary mechanism for the creation of such chimeric proteins.

It is possible that multidomain proteins contain traces of their evolutionary history in their nucleotide sequence. We explored the ability of differences in the usage of isoaccepting synonymous codons (that is, ones which code for the same amino acid) to support inference of distinct evolutionary history for putative protein domains. In essence, our method tests whether it is likely that the observed distribution of synonymous codons arose from the concatenation of multiple, distinct distributions.

In order for this method to detect chimeric proteins, different portions of such proteins would have to demonstrate different distributions of synonymous codons. There are several reasons to believe this requirement could be met. First, it has been demonstrated that there are clear differences in the usage of synonymous codons between genes within a single genome (Li and Graur 1991). This within-genome difference could arise by several different mechanisms. For genes with high expression levels, a positive correlation exists between the relative abundance of tRNA species and the relative frequencies of the corresponding synonymous codons. In genes with relatively low expression, however, this correlation is weak or nonexistent. Another possible mechanism involves constraints on DNA other than protein coding, such as the stability of its 3D structure, interactions with histones, or recognition by DNA-binding regulatory proteins. In addition to these within-genome factors, there are further possible mechanisms to explain differences in synonymous codon usage within chimeric proteins. It is possible that the different domains of a chimeric protein arose at different times during the evolutionary history of the organism, and during the evolution of one of the domains there was an environmental stress (e.g., temperature extremes or scarcity of a trace metal required for the biosynthetic pathway of one of the nucleotides) that resulted in selective pressure for the preferential usage of one particular synonymous codon.

Even assuming that the domains that were joined to create a chimeric protein did reflect different distributions of synonymous codons, it is possible that such distributional differences are lost in the modern protein, having been overwhelmed by noise from silent nucleotide substitutions since the chimerism event. Due to both the possibility of uniform distributions of synonymous codons in the original domains, and the possibility that silent substitutions may overwhelm original differences, this method can only be used to positively infer a chimeric origin; the absence of a signal does not imply that a protein is not chimeric.

We chose to test our method on the amino-acyl tRNA synthetases, which are among the most ancient of proteins. The amino-acyl tRNA synthetases represent a family of approximately 20 proteins per organism. Because of their key role in the translation machinery, collectively they embody the genetic code of an organism or organelle. These synthetases have been studied extensively by Schimmel and colleagues (e.g., Burbaum and Schimmel 1991), whose corpus of published work conclusively demonstrate that the individual members of the family are

each chimeric, based on structural similarities determined by X-ray crystallography and subsequence conservation (Schimmel, Shepard, and Shiba 1992).

We suggested above that protein chimerism can be detected by testing for a difference in the distribution of synonymous codon usage in the different domains that make up a chimeric protein. However, since we do not know ahead of time where the domain boundaries are, we use a hidden Markov model (HMM) of codon usage, where the hidden states identify the boundaries of the domains. We then compare the goodness of fit of HMMs of varying numbers of hidden states to the distributions of the codons in the sequence being tested. If a 2 state HMM fits the sequence significantly better than a 1 state HMM, this would provide evidence that the sequence is chimeric. If the 2 state model is not significantly better than the 1 state model, there may be several reasons: (1) the protein is not chimeric, (2) the ancestral domains had homogeneeous synonymous codon usage distributions, (3) homogenizing point mutations occurred after the two ancestral domains joined, or (4) our detection method is insufficiently sensitive.

Methods

We compare the goodness of fit of a 1 state versus a 2 state hidden Markov model to sequence data to test the hypothesis of chimerism in a protein or a family of proteins. In our models, a state represents a particular distribution of codons in a portion of a nucleotide sequence. A state specifies, for each codon, the probability of observing that codon in the sequence covered by that state. Transitions between states indicate changes in the expected distribution of codons. A self-transition indicates that the next codon in the sequence was drawn from the same distribution. All transitions are unidirectional; once a transition is made from a prior state to a subsequent one, the prior state cannot be returned to. This captures the fact that chimeric events splice together sequences, and do not interleave them.

For each sequence to be tested for chimerism, we find the optimal parameter values for a set of HMMs of 1 and 2 states, using the full Baum-Welch algorithm as described by Rabiner (1989), incorporating unpublished corrections independently discovered by Rabiner and by Trebbe (personal communication from L. R. Rabiner to B. Zeeberg). We then calculate the probability of each model, given the data. These fits need to be controlled for two factors: first, since the HMMs with 2 states have more free parameters, they are expected to show better fits for a given dataset, a priori. Second, we care only about the possibility of differing distributions of synonymous codons, not all codons. We thus control for amino acid compositional bias, since amino acid bias is likely to have a functional rather than evolutionary significance. To address these issues, we devised a randomized control condition which provides an estimate of the expected fit of these HMMs to the data if there were no contribution from multiple distributions of synonymous codons.

To calculate the expected fit of a model, we create a control set of sequences derived from the natural sequence by randomly permuting synonymous codons. The amino acid sequence of the randomized sequences are identical with the original sequence, as is the number of each type of codon. The only difference between the sequence under test and the set of randomized sequences is the order in which the synonymous codons appear. The studies used 1100 randomized sequences per test. Each randomized sequence is used to train the various sized HMMs, and the mean of the goodness of fit of each HMM to the randomized sequences is used as a baseline. The goodness of fit for the 1 state model is unchanged by the randomization procedure, since permutation does not affect the distribution if there is only a single domain.

The significance of an apparent improved fit of a 2 state model is evaluated in two ways. The first method is based upon the previous observation that the distribution of the log likelihood ratio for the randoms, was essentially normal (Zeeberg and Hunter 1997). Thus we calculate a Z-score of the difference between the log likelihood ratio for the natural and the mean of the log likelihood ratios for the randoms (that is, that difference divided by the standard deviation of the log likelihood for the randoms).

Our second measure, empirical significance testing, does not depend on assumptions of normality. We calculate the percentage of randomized sequences for which the difference between the fit of the 2 state model and the 1 state model trained on randomized data was less than the equivalent difference for the natural sequence. The two measures are correlated; we use the Z score as an indication of the magnitude of a difference, and the empirical significance measure as a threshold for determining that the magnitude is unlikely to have been observed by chance.

The method was implemented in C code on an Apple Macintosh Quadra 950. Nucleotide sequences were obtained for the amino-acyl tRNA synthetases of the nuclear and mitochondrial synthetases of the yeast S. cerevisiae (Goffeau et al. 1996; MIPS database).

Results and Discussion

We have previously reported a test of the two state model for all of the annotated S. cerevesiae nuclear synthetases (Zeeberg and Hunter 1997). The goal of the current effort is to extend the previous work to include all the annotated yeast amino-acyl tRNA synthetases, both nuclear and mitochondrial, and compare the evidence for chimerism in corresponding nuclear and mitochondrial synthetases.

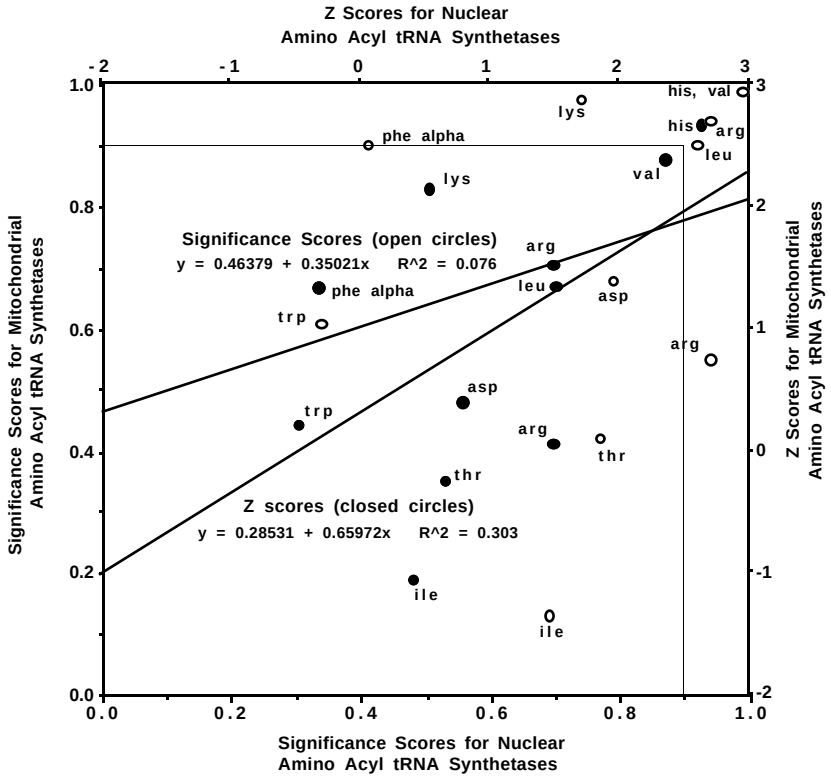
Several reviews are available that describe the (apparently incompletely known) molecular biology of the yeast mitochondrial amino-acyl tRNA synthetases (Costanzo and Fox 1990; Gillham 1994). Briefly, both the mitochondrial and nuclear synthetases are encoded by the nuclear genome, the mitochondrial synthetases being subsequently exported to the mitochondria, presumably as the result of a leader sequence which specifically targets gene products for transport to the mitochondria. There is,

Table 1. Significance and Z Scores for Mit and Nucl Amino Acyl tRNA Synthetases

	nuclear		mitochondrial	
synthetase	**significance**	**Z score**	**significance**	**Z score**
ala	0.75	0.64		
arg	0.94	1.51	0.94, 0.55	1.51, 0.04
asn	0.85	1.01		
asp	0.79	0.81	0.68	0.38
cys	0.88	1.15		
glu			0.43	-0.22
glu/pro	1.00	4.23		
gln	0.86, 0.97	1.06, 2.06		
gly	0.01, 0.94	-2.09, 1.70		
his	0.99	2.65	0.99	2.65
ile	0.69	0.43	0.13	-1.08
leu	0.92	1.53	0.90	1.34
lys	0.74	0.55	0.98	2.12
met			0.40	-0.30
phe alpha	0.41	-0.30	0.90	1.32
phe beta	0.93	1.70		
pro	0.97	2.05		
ser	0.82, 0.15	0.88, -1.01		
thr	0.77	0.68	0.42	-0.27
trp	0.34	-0.46	0.61	0.20
tyr			0.27	-0.64
val	0.99	2.37	0.99	2.37

The synthetases for which there are both nuclear and mitochondrial entries are shown as a scatter plot in Fig. 1. For the nuclear gln, gly, and ser synthetases there are 2 entries, presumably as a result of gene duplication in the yeast chromosome. None of these appear in Fig. 1, since the corresponding mitochondrial synthetases are missing. For the mitochondrial arg synthetase there are 2 entries, and both appear in Fig. 1. For arg, his, and val synthetases, the identical values for the nuclear and mitochondrial results are due to the fact that these are overlapping genes; these results therefore do not represent independent computations.

Figure 1. Significance and Z Scores for Mit and Nucl Amino Acyl tRNA Synthetases



however, apparently only an incomplete set of yeast amino-acyl tRNA synthetases annotated in the MIPS database, and this incompleteness limited the scope of the comparison we wished to make between the mitochondrial and nuclear synthetases.

At the 90% level of significance, 9 of the annotated nuclear synthetases and 6 of the annotated mitochondrial synthetases (Table 1) were found to be chimeric by our method. Despite the fact that some nuclear and mitochondrial synthetases for a given tRNA show significant amino acid sequence similarities (results of BLAST search, unpublished), no correlation was observed between the significance scores or the Z scores for the mitochondrial synthetases and the corresponding nuclear synthetases (see Fig. 1, which shows only results for which both nuclear and mitochondrial synthetases are annotated). These synthetases are all known to be chimeric (Schimmel, Shepard, and Shiba 1992), so the failure of our method to identify all them as such suggests that the method may not be as sensitive as is desirable, or that a higher state (> 2 state) HMM may be required. However, the chimeric event that led to the formation of these proteins is extremely ancient; if silent nucleotide substitutions are going to overwhelm our approach in some situations, it is not surprising that it may have occurred in this family of proteins.

It is intriguing that the 2 genes that are annotated as encoding the nuclear gly synthetase (for which we have found that CLUSTALW analysis confirms significant amino acid homology) were found to be at the extreme opposite ends of the spectrum of chimerism using our 2 state HMM (Table 1). We speculate that an ancient gene duplication event resulted in one of the two genes becoming a pseudogene which underwent substantial homogenizing nucleotide mutations which masked the original evidence of chimerism.

Finally, in order to test the method for false positives, a set of proteins would be needed that are known to be nonchimeric. Since this is not feasible in practice, the alternative would be to generate a synthetic nonchimeric protein and use this as if it were the natural protein in our HMM analysis. Of course, any of the 1100 randoms that we generate in each experiment could be thought of as being this synthetic nonchimeric protein, and the other 1099 randoms that had been generated from the natural protein could be thought of as the set of randoms that would have been generated from the synthetic nonchimeric protein. Thus, there is a 10% chance that a protein randomly chosen (from the 1100 randoms) to be the synthetic nonchimeric protein would yield a false positive result. While this may seem distressing when studying a single protein, studying a family of proteins is much more reliable: for example, our finding of chimerism for 9 out of 20 nuclear synthetases would happen by random chance with a probability of about $(0.10)^9\ (0.9)^{11}\ 20!/(9!\ 11!) = 5.27 \times 10^{-5}$. Similarly, the probability of a single false positive out of 20 synthetases would be about 0.27.

Conclusions

We propose a novel method for identifying chimeric nucleotide sequences. Although the method is potentially prone to false negatives, positive conclusions are statistically well founded. We tested the technique on all known S. cerevisiae tRNA synthetases. Establishing that a protein is chimeric by traditional methods is labor intensive, generally involving deducing an evolutionary tree for the individual modules of the chimeric protein, as Patthy (1985) did for TPA. Our present method is somewhat less sensitive than an experimental approach, but we believe it could be useful for the automated preliminary screening of entire genomes or gene databases.

References

Burbaum, J.J. and Schimmel P. 1991. Assembly of a Class I tRNA Synthetase from Products of an Artificially Split Gene. Biochem. 30:319-324.

Costanzo, M.C. and Fox, T.D. 1990. Control of Mitochondrial Gene Expression in Saccharomyces Cerevisiae. Annu. Rev. Genet. 24:10.

Gillham, N.W. 1994. Organelle Genes and Genomes, Oxford University Press, p. 313.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldman, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S.G. 1996. Life with 6000 Genes. Science 274:546-567.

Li, W-H. and Graur, D. 1991. Fundamentals of Molecular Evolution , Sinauer Associates, Inc.

Patthy, L. 1985. Evolution of the Proteases of Blood Coagulation and Fibrinolysis by Assembly from Modules. Cell 41:657-663.

Rabiner, L.R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 77:257-285.

Schimmel, P. 1991. Classes of Amino-acyl tRNA Synthetases and the Establishment of the Genetic Code. Trends Biochem. Sci. 16:1-3.

Schimmel, P., Shepard, A., and Shiba, K. 1992. Intron Locations and Functional Deletions in Relation to the Design and Evolution of a Subgroup of Class I tRNA Synthetases. Protein Sci. 1:1387-1391.

Zeeberg, B. and Hunter, L. 1997. Characterization of a Family of Chimeric Proteins, the Amino-acyl tRNA Synthetases, by Determining Differential Codon Usage using One and Two State HMMs. Math. Model. Sci. Comp. vol. 8. Forthcoming.