ELSEVIER

Gene 253 (2000) 313–322

GENE
AN INTERNATIONAL JOURNAL ON GENES, GENOMES AND EVOLUTION

www.elsevier.com/locate/gene

The molecular evolution of signal peptides

Elizabeth J.B. Williams, Csaba Pal, Laurence D. Hurst *

Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK

Received 3 March 2000; received in revised form 12 May 2000; accepted 24 May 2000
Received by W.-H. Li

Abstract

Signal peptides direct mature peptides to their appropriate cellular location, after which they are cleaved off. Very many random alternatives can serve the same function. Of all coding sequences, therefore, signal peptides might come closest to being neutrally evolving. Here we consider this issue by examining the molecular evolution of 76 mouse–rat orthologues, each with defined signal peptides. Although they do evolve rapidly, they evolve about half as fast as neutral sequences. This indicates that a substantial proportion of mutations must be under stabilizing selection. A few putative signal sequences lack a hydrophobic core and these tend to be more slowly evolving than others, indicating even stronger stabilizing selection. However, closer scrutiny suggests that some of these represent mis-annotations in GenBank. It is also likely that some of the substitutions are not neutral. We find, for example, that the rate of protein evolution correlates with that of the mature peptide. This may be a result of compensatory evolution. We also find that signal peptides of immune genes tend to be faster evolving than the average, which suggests an association with antagonistic co-evolution. Previous reports also indicated that the signal peptide of the imprinted gene, *Igf2r*, is also unusually fast evolving. This, it was hypothesized, might also be indicative of antagonistic co-evolution. Comparison of *Igf2r*'s signal peptide evolution shows that, although it is not an outlier, its rate of evolution is comparable to that of many of the faster evolving immune system signal sequences and 5/6 of the amino acid changes do not conserve hydrophobicity. This is at least suggestive that there is something unusual about *Igf2r*'s signal sequence.

Keywords: *Igf2r*; Molecular evolution; Mouse–rat orthologues; Neutral evolution; Signal peptides

1. Introduction

There is considerable variation both within and between proteins in the rate of evolution. What are the causes of this variation? Most proteins in the mouse–rat comparison (Wolfe and Sharp, 1993; Makalowski and Boguski, 1998; Hurst and Smith, 1999) show evidence of stabilizing selection and hence evolve at a slow rate, compared with the underlying mutation rate (as assayed by the K_A/K_S ratio, where K_A and K_S are the rates of non-synonymous and synonymous DNA changes per site, respectively). However, there remain a significant proportion of proteins (or sub-domains of proteins) that show relatively high rates of evolution.

Numerous analyses have indicated that many fast evolving proteins, or sub-domains of proteins, are probably engaged in some form of selectively driven antagonistic co-evolution. For example, host immune system genes (Hughes et al., 1990; Hurst and Smith, 1999) and parasite antigens (Hughes, 1992) show rapid evolution, especially at the sites of mutual binding (Hughes et al., 1990; Hughes, 1991, 1992). Similarly, genes involved in bacterial antagonistic interactions (Tan and Riley, 1997), as well as those potentially involved in both inter-sexual conflict, for example some of those of *Drosophila*'s seminal fluids (Aguade et al., 1992; Tsaur and Wu, 1997), and parent–offspring conflict, for example numerous placentally expressed genes in mammals (e.g., placental lactogens) (Hurst and McVean, 1998), also show unusually high rates of evolution.

However, other high rates of evolution are indicative of neutral evolution. The most convincing example comes from analysis of pseudogenes (Li et al., 1981). However, there has yet to be identified a class of protein coding genes (or sub-domains) that are dominantly

Abbreviations: lgf2r, insulin like growth factor type II receptor; lgf2, insulin like growth factor 2; Q, glutamine; L, leucine; P, proline; R, arginine; V, valine.

* Corresponding author. Tel.: +44-1225-826424; fax: +44-1225-826779.
E-mail address: l.d.hurst@bath.ac.uk (L.D. Hurst)

0378-1119/00/$ - see front matter
PII: S0378-1119(00)00233-X

neutrally evolving. Here we examine the molecular evolution of signal peptides and ask whether these might serve as good paradigmatic examples of neutral evolution, as from knowledge of their biochemistry this might be suspected.

Signal peptides are short N terminal (genic 5′) parts of a protein whose function it is to direct the peptide to its appropriate cellular location. After having delivered the mature peptide to this location, the signal peptide is cleaved and is presumed to be digested. The fact that it is cleaved allows us to suppose that signal peptides have one and only one function, to deliver the mature peptide to its appropriate location.

Signal peptides often show little evidence of sequence similarity. The lack of identity among these sequences implies that numerous forms of sequence can serve the very same role and are sufficient for membrane transport (see, e.g., Izard and Kendall, 1994). It has often been argued that the only requirement for proper functioning of the signal peptide is to contain a hydrophobic core consisting exclusively or largely of hydrophobic amino acids. As a support for the theory, Kaiser et al. (1987) found that about 20% of random sequences can act as functional signal sequences. Furthermore, it is also known that amino acids with similar hydrophobicity are coded by neighbouring codons (see for references Freeland and Hurst, 1998). Therefore, most non-synonymous mutations conserve hydrophobicity. More generally then, if signal sequences are not neutrally evolving, it is hard to imagine a class of coding sequences (as opposed to pseudogenes) that are wholly neutral (but see Dickerson, 1971 on fibrinopeptides).

There are, however, other features of signal peptides that are common, although the relationship between structure and function is not transparent. Often there is a leucine-rich region. Typical signal peptides also have a positively charged n-region and a neutral but polar c-region. Positions −3 and −1 from the cleavage site must be small and neutral for cleavage to occur correctly (for analysis of other diagnostic features/methods see Nielsen et al., 1997; Ladunga, 1999).

While the previously reported low sequence similarity is consistent with neutral evolution, other data suggest that the pattern might be more complicated. From a sample size of three, it was reported that some genes may have relatively low rates of evolution in the signal sequence (Li et al., 1985), but that this may be an artifact of the small size of the signal peptide. Further, Smith and Hurst (1998) reported a strong correlation between the rate of evolution of the signal peptide and that of the complete gene. This is hard to understand from a neutralist perspective. However, this previous analysis permitted non-independence, in that the signal peptide was allowed as part of the complete coding sequence. So it remains to be established whether, if one controls for non-independence, the correlation of rates still exists.

Here, then, we shall ask just how fast signal sequences evolve and whether their evolution is simply neutral. To this end we shall also ask whether the rate of evolution in the signal peptide is correlated with that in the mature peptide. Additionally, we shall ask whether there are any biochemical aspects of signal peptides that in any way explain the variance in their rate of evolution.

We shall also examine a specific selectionist explanation for the evolution of one particular signal peptide, that of the insulin like growth factor type II receptor (*Igf2r*). *Igf2r* is an imprinted gene expressed off the maternally derived chromosome in rodent embryos. It is a transmembrane receptor that binds to the paternally expressed Igf2 (and numerous other ligands) to target them to the lysosomes. This is interpreted by the conflict hypothesis for the evolution of imprinting (Moore and Haig, 1991) as an antagonistic interaction. McVean and Hurst (1997) found that, in the mouse–rat comparison, at the position where Igf2r binds to Igf2 the protein shows an especially low rate of evolution, indicating stabilizing selection, rather than the directional selection expected if the interaction is antagonistic. Smith and Hurst (1998) showed that the same was also true in the human–cow comparison.

The latter analysis, however, also reported that in both the human–cow and the mouse–rat comparisons, the signal sequence has a high K_A/K_S ratio. It was speculated, therefore, that there might be a conflict concerning the cellular localization of Igf2r. Hence, the unusual property of Igf2r's signal sequence may be explained by strong directional selection driven by antagonistic co-evolution. As the previous analysis used only nine other genes, here we aim to ask whether Igf2r's signal peptide really is an outlier by comparing it with a much larger set of other signal peptides.

2. Materials and methods

We compiled a dataset of mouse and rat orthologues in which the signal peptide has been annotated in at least one of the two GenBank entries. NCBI Entrez (http://www.ncbi.nlm.nih.gov/Entrez/), and ACNUC software at the UK HGMP Resource Centre (http://www.hgmp.mrc.ac.uk/) was used to search and extract rat and mouse complete coding sequences with annotated signal peptide regions. This resulted in a list of nearly 400 genes.

Each of these was scrutinized in the HOVERGEN database (Duret et al., 1994) to find mouse–rat orthologues. Genes were accepted as othologues if, and only if, the mouse–rat sequences had no other non-rodent sequence between them and at least one non-rodent sequence appeared as a sister group. This resulted in a data set of 80 gene pairs.

GENETRANS was used to automatically extract

complete coding sequences, while GBPARSE (available from http://sunflower.bio.indiana.edu/~wfischer/Perl_Scripts/) was used to automatically extract signal peptide regions, using annotations in the GenBank entries. Mature peptides (complete sequence minus signal peptide) were analysed by editing out the signal peptide from the alignment files.

DNA alignments of signal sequence and entire protein coding sequences were carried out by PILEUP, using the default settings. The alignments were checked by eye and modified if necessary. Signal sequences were checked to ensure that they aligned perfectly against themselves within the complete gene alignment. For four genes we were unable to find unambiguous alignments, and these were excluded. This resulted in a dataset of 76 genes, including *Igf2r*. The genes and their Accession Nos. are given in Appendix A.

Substitution rates were estimated by the package DIVERGE (available at HGMP). The program is based on the method described by Li (1993) using the Pamilo and Bianchi modifications (Pamilo and Bianchi, 1993), and applies Kimura's two-parameter method to correct for multiple hits and to account for the difference in substitution rates for transitions and transversions. These data are also reported in Appendix A.

The hydrophobicity of all the signal peptides was examined using PepPlot within GCG at HGMP. Mouse Genome Informatics (http://mgd.hgmp.mrc.ac.uk/) and SWISS-PROT were used to find immune related genes of our sample. A gene was classified as an immune gene if either of the two entries specifically mentioned involvement in immune response or expression in immune specific cell types.

3. Results

3.1. Signal peptides have a fast rate of evolution, but many non-synonymous mutations are under stabilizing selection

Signal peptides do appear to evolve faster than mature peptides, although by how much depends precisely on the statistic used (see Table 1). If we calculate a mean K_A for both mature and signal peptide, then we find that on average signal peptides evolve a little under twice as fast. Allowing for underlying mutation rate differences by using the K_A/K_S ratio suggests that signal peptides evolve a little over twice as fast. By comparison, if we consider the mean value of the paired ratios per gene (e.g., K_A signal peptide/K_A mature peptide), then the signal peptides on the average evolve over five times faster than the flanking mature peptide. Allowing for the underlying mutation rate, we find a comparable figure. The paired test is possibly the least accurate as the ratios have an extremely high variance which most likely reflects the effects of the small size of the signal peptide. Similarly, if K_S is unusually low, the K_A/K_S ratio becomes extraordinarily (and probably unrealistically) high. That signal peptides are fast evolving, none the less seems clear: out of 76 genes 53 had a higher K_A/K_S in the signal sequence than in the gene as a whole. This is highly significantly different from null expectations (χ^2, $P<0.001$).

However, by none of these measures is the rate of signal sequence evolution as high as would be expected were the sequences neutrally evolving. Mean K_A is half the value of mean K_S, and mean K_A/K_S for each signal sequence is 0.63 ± 0.114 which is very much more than two standard errors away from unity, the figure expected if sequence evolution is perfectly neutral.

3.2. Unusual signal sequences evolve unusually slowly

Some signal peptides appear to evolve relatively slowly. Is this chance variation or might these sequences be functionally unusual as well? We have examined the hydrophobicity plot of all of the signal peptides. At least six of our genes do not have the typically hydrophobic signal peptide, i.e. they lacked a hydrophobic core (NB. Igf2r is normal). These are indicated in Appendix A. Although the sample size is limited, we find that these six tend to evolve slowly for signal peptides (they evolve at about a third of the rate of others), although the statistics are marginal (Table 2).

Table 1
Basic statistics of the K_A, K_S and K_A/K_S for mature and signal peptides ($N=76$)

	K_A	K_S	K_A/K_S
Mature peptide (mean ±S.E.M.)	0.05 ± 0.006	0.198 ± 0.010	0.249 ± 0.028
Signal peptide (mean ±S.E.M.)	0.09 ± 0.012	0.181 ± 0.019	0.628 ± 0.114[b]
Signal/mature (paired)	5.23 ± 1.60[a]	0.990 ± 0.108	5.61 ± 1.08[c]
Rank correlation between mature and signal, P value from slope of regression of ranks	$r^2=0.136$ $P=0.001$ Slope = 0.368	$r^2=0.023$ $P=0.193$ Slope = 0.151	$r^2=0.033$ $P=0.12$ Slope = 0.18

[a] Omits four data points in which mature $K_A=0$.
[b] Omits three data points in which $K_S=0$.
[c] Omits five data points in which the ratio is infinite.

Table 2
Comparison of the six unusual signal peptides with those with normal hydrophobicity plots

	K_A	K_S	K_A/K_S	Signal peptide size (nt)
Unusual ($N=6$)	0.032 ± 0.014	0.123 ± 0.045	0.219 ± 0.065	310 ± 161
Normal ($N=70$)	0.096 ± 0.012	0.186 ± 0.021	0.665 ± 0.123[a]	76 ± 3.39
Mann–Whitney U test for difference	$P=0.0506$	$P=0.3808$	$P=0.066$	$P=0.365$

[a] Omits three data points in which $K_S=0$.

This finding obviously tempts the question as to whether there are different classes of signal peptide that have different rates of evolution (and if so why) or whether these six do not really have signal peptides at all and are mis-annotated in the GenBank entry?

To address this issue further, we examined the Swiss-Prot entries for these six proteins. We have also examined the sequences using Sigcleave at EMBOSS (http://www.hgmp.mrc.ac.uk/Registered/Option/emboss.html). Two of these, Acetyl Co-A (Swiss-Prot Acc: P45952, mouse; P08503, rat) and sterol carrier protein 2 (Swiss-Prot Acc: P32020, mouse; P11915, rat) had no signal peptide mentioned in Swiss-Prot. Sigcleave failed to identify any signal peptide cleavage sites. Sigcleave correctly identifies 95% of signal peptides, and rejects 95% of non-signal peptides. The cleavage site should be correctly predicted in 75–80% of cases. Given this, the GenBank annotation is likely to be misleading.

Of the remaining four, both Sigcleave and Swiss-Prot agreed that a signal peptides might be present. However, Ephrin B1 (Swiss-Prot Acc: P52795, mouse; P52796, rat) and Coagulation factor III (Swiss-Prot Acc: P20352, mouse; P42533, rat) have only weakly defined cleavage sites under the Sigcleave analysis. Ephrin B1 also lacks the usual leucine-rich domain. It is therefore possible that these proteins do not have signal peptides.

Inhibin beta A (Swiss-Prot Acc: P18331, mouse; Q04998, rat) has an unusual Swiss-Prot entry, as the rat protein had been annotated as having a signal peptide and a propeptide; however, it was not known where one stopped and the other started. The GenBank entry may well then be a combination of signal peptide and propeptide. This was supported by Sigcleave analysis. This method found a cleavage site at only 21 amino acids, where the GenBank annotation indicates a signal peptide in excess of 200 amino acids long. The size defined by Sigcleave is around the mean for the remaining 'normal' signal peptides. Inhibin alpha (Swiss-Prot Acc: Q04997, mouse; P17490, rat) likewise has a huge signal peptide according to GenBank, but both Swiss-Prot and Sigcleave agree that the signal peptide is cleaved at amino acid 21. This could have caused the appearance of a slow rate of evolution. However, inhibin beta A shows neither synonymous nor non-synonymous evolution ($K_A=K_S=0$). Inhibin alpha shows a high K_A/K_S ratio ($K_A=0.07$, $K_S=0.04$).

It appears, then, that there is some degree of mis-annotation in GenBank. This issue can, however, only be addressed definitively by detailed biochemical analysis of the genes concerned, analysis which, as yet, appears not to have been done.

3.3. Mitochondrial signal sequences are longer but evolve at a normal rate

Signal sequences are known to direct the transport of proteins across different types of membranes (e.g., endoplasmic reticulum, Golgi-network, mitochondria). Therefore, it is reasonable to ask whether the variation in the rate of evolution is explained by the location to which the signal peptides direct the protein. In order to address this issue, we have compared the evolution of signal sequences that direct the import of mitochondrial proteins encoded in the nucleus to the remaining others (Table 3).

In our original sample there is only one sequence that was annotated as being a nuclear-encoded mitochondrial protein. Therefore, we compiled a new dataset of mitochondrial proteins. This we did by examining NCBI Entrez using 'mitochondrial' as key word and

Table 3
Comparison of mitochondrial and non-mitochondrial signal peptides

	K_A sig	K_S sig	K_A/K_S sig	Signal peptide size (nt)
Mitochondrial ($N=8$)	0.0572 ± 0.016	0.102 ± 0.015	0.727 ± 0.251	107.6 ± 12
Non-mitochondrial ($N=75$)	0.089 ± 0.012	0.18 ± 0.02	0.608 ± 0.114[a]	94.2 ± 14.1
Mann–Whitney U test for difference	$P=0.44$	$P=0.19$	$P=0.56$	$P=0.0029$

[a] Omits three data points in which $K_S=0$.

then checking to ensure the genes were nuclear. Although several mitochondrial genes with annotated signal (or transit) peptides can be found in the databank, only eight genes have been found with rat orthologues (see Appendix B). The analyses of these genes have detected no significant difference in the rate of evolution of mitochondrial signal sequences compared with that of non-mitochondrial ones. However, we have to emphasize that the failure to notice any differences may be a pitfall of the low sample size of mitochondrial proteins.

Although there is no sign of unusual evolution, it is still possible that mitochondrial signal peptides are functionally different. We find that mitochondrial genes are significantly longer than non-mitochondrial ones (Mann–Whitney test, $P<0.01$). This result is not surprising, as precursor proteins are imported into the mitochondria in a multistep process mediated by translocation systems of the outer and inner membrane (Gillham, 1995). Hence, pre-sequences of mitochondrial proteins are expected to contain multiple signal elements to reach their appropriate locations (Gillham, 1995). We have also examined the secondary structure of mitochondrial pre-sequences. None of them show signs of unusual hydrophobicity.

3.4. Are the substitutions due to selection or drift?

While signal sequences as a whole are not perfectly neutrally evolving, we can also ask about the substitutions that are seen. Are these the result of drift or might positive selection be suspected? We cannot answer this question definitively, but can ask whether (a) the substitutions greatly affect hydrophobicity, (b) whether the rates of evolution of signal and mature peptides are correlated (which is not obviously consistent with neutral expectations) and (c) whether genes involved in antagonistic interactions (immune genes and Igf2r) show fast evolving signal peptides.

3.5. Is hydrophobicity conserved?

A neutralist model for the evolution of signal sequences would predict that the non-synonymous substitutions conserve the hydrophobicity of the amino acid. This test does not discriminate selectionist and neutralist explanations, as selectionist explanations might also require conservation of hydrophobicity. However, it has the potential of falsifying a neutralist hypothesis.

Knowing whether more of the substitutions do this than expected is not, however, trivial. The code is arranged such that point mutations tend to conserve hydrophobicity (for quantification see Haig and Hurst, 1991). A bias to conservation is therefore expected from the null model that all non-synonymous mutations are equally likely to be fixed, regardless of hydrophobicity. Given, too, an ambiguity regarding transition/transversion rates (and hence the expected rate of different non-synonymous changes), predicting the null expectation for the degree of conservation is hard to do unambiguously. However, here we perform a simple, albeit rough, alternative test. It is known that mutations at the first site in a codon tend to conserve hydrophobicity where those at the second site do not (Haig and Hurst, 1991). Assuming that there is no reason to expect more mutations at the first rather than the second site, the neutralist model would be falsified by not finding an excess of mutations at the first rather than at the second site.

We have done this for all the non-synonymous substitutions in each signal peptide. We find that of 275 mutations, 159 affect the first site, while 116 affect the second, a significantly greater excess ($P<0.01$), consistent with expectations.

3.6. Rates of protein evolution in signal peptide and mature peptide are correlated

An earlier study of signal sequences (Smith and Hurst, 1998) indicated that the rate of evolution of the entire peptide may well be correlated with the rate of evolution of the signal peptide. In our dataset as well the K_A/K_S of the signal peptide strongly co-varies with the K_A/K_S of the entire peptide (r^2 ranked data $=0.122$, rank correlation $P=0.002$).

However, in this analysis [and the previous one (Smith and Hurst, 1998)] the signal sequence is included within the entire peptide, so introducing a non-independence. If we analyse the rates of evolution of signal peptides and compare them with those of the mature peptides, this non-independence is removed. We now fail to find a strong correlation between K_A/K_S of the mature and signal peptides, although there might be a tendency (ranked data $r^2=0.03$, $P=0.12$) (Table 1). Similarly, we find that the K_S values do not correlate ($P=0.19$). However, we find a strong positive correlation between the absolute rate of evolution of signal sequences (K_A) and that of mature peptides (regression of ranks: $P=0.001$).

3.7. Signal sequences of immune genes are fast evolving

Previous analyses have indicated that throughout their sequence, immune system genes are fast evolving (Hurst and Smith, 1999). Is the same true of their signal peptides? From a neutralist perspective it is hard to see why selection acting on the mature peptide should affect substitution rates in the signal peptide.

Comparing signal peptides of immune system genes ($N=14$) with non-immune genes, we found that the former tended to be faster evolving (assayed by K_A/K_S: Mann–Whitney U test, $P=0.03$; see also Appendix C). The immune genes' mature peptides are also faster evolving ($P<0.001$). Given that high rates of evolution through the rest of the sequence are most likely a result of antagonistic co-evolution, this finding is consistent with some high rates of evolution in signal peptides being associated (directly or indirectly) with the same. Analysis of intra-populational variation would be helpful to clarify the issue.

3.8. *Igf2r's signal peptide evolves at a rate comparable with that of many immune genes*

To determine whether the signal peptide of Igf2r evolves at a faster rate than other signal peptides, the signal peptide K_A/K_S values were ranked. The K_A/K_S of *Igf2r*'s signal peptide is ranked 67 out of 76 (higher ranks being faster evolving). In large part this is because the K_S of the signal peptide is low: taking the K_A alone, it was ranked 50 out of all the 76 signal peptides. Neither statistic suggests that it is an outlier as previously indicated on the basis of a sample size an order of magnitude smaller.

We can also ask, given the rate of evolution of the mature peptide, does *Igf2r* have an unusually fast rate of evolution? In order to take this covariance into consideration, the difference in rank of the signal peptide K_A/K_S and the mature peptide was examined. *Igf2r* was found to have a positive difference in rank, but 11 genes had a higher difference, i.e. a higher K_A/K_S given the K_A/K_S of the mature peptide. This shows that, for the local K_A/K_S, *Igf2r*'s signal sequence is not an outlier. Likewise, its K_A, controlling for the K_A of the mature peptide is not unusual (25 have a greater K_A in the signal peptide given the K_A of the mature peptide).

Could the size of signal peptides be affecting this result? There is a much higher variation in signal peptide K_A/K_S (S.E.M. = 0.114, omitting three with a ratio of infinity) compared with entire peptide K_A/K_S (S.E.M. = 0.023). This could have been due to signal peptides being small and hence providing misleading estimates. If *Igf2r* has an unusually sized signal peptide, this might in part explain the findings. However we cannot substantiate this hypothesis.

By splitting the data set into two sets, one with large signal peptides and one with small, we found that there was no appreciable difference in K_A/K_S in the two sets (using two-tailed Mann–Whitney U test on signal peptide K_A/K_S, $P>0.05$). It was also shown that there was no difference in the variation between the two sets of data. Taking the squares of the residuals from the regression line (which was flat), we find the large and the small set are no different (two-tailed Mann–Whitney, $P>0.5$). These results all indicate that although the signal peptides are small, there is no trend with respect to their size within the group of signal peptides. Size effects are therefore unlikely to explain the K_A/K_S of *Igf2r* given the local rate of evolution. This result is further strengthened by noting that *Igf2r* itself has a signal peptide size of 96 bp, which is very close to the mean of the signal peptides in this data set (mean = 94.6).

All these results suggest that *Igf2r* signal peptide is probably not an outlier in our sample. However, as established, immune genes tend to have fast evolving signal sequences. Perhaps importantly, then, only four out of 14 immune genes have a higher K_A/K_S ratio in their signal peptides than *Igf2r* (Appendix C). This suggests that the rate of evolution of Igf2r's signal peptide may be, as originally claimed, unusually high for a non-immune gene.

That something unusual is going on is further supported by the finding that of six non-synonymous changes in the signal peptide, five occur at the second site, and do not conserve hydrophobicity. Four of the five reverse the hydrophobicity, as measured on the White interface scale (http://blanco.biomol.uci.edu/hydrophobicity_scales.html), the other causes a proportionally large change (the five second site changes are, in order 5′→3′: Q↔L, L↔P, R↔P, P↔L and L↔V). The number of non-synonymous changes at the first and second site is significantly different to that found in the other 75 genes taken in total (*G*-test of independence with Williams Correction = 4.06, $P<0.05$).

4. Discussion

This analysis set out to answer the following four questions:

1. Do signal peptides have rates of evolution expected of sequences that are perfectly neutral?
2. Does the rate of evolution of the signal peptide correlate with that of the mature peptide, possibly indicating a non-neutral force on peptide evolution?
3. Do all signal peptides evolve at the same rate?
4. Can we substantiate the claim that *Igf2r* has an especially high rate of evolution in its signal peptide?

As regards the first issue, it appears that signal peptides do evolve faster than the mature peptide, although by how much is dependent upon the measure. If we use paired samples, then they evolve on average between five and six times faster. However, if instead we take an average for all signal peptides and compare that with an average for all mature peptides, then they appear to evolve about twice as fast. Either way, a significant fraction of non-synonymous mutations must be under stabilizing selection. For the six lacking the usual hydrophobic core and that were on average significantly better conserved, the fraction must be much

higher. The functioning of the signal peptides of these unusual proteins is worthy of further investigation. Removal of the six signal peptides without the usual hydrophobic core does not affect the conclusion that signal sequences have many non-synonymous mutations under stabilizing selection (see Table 2).

The results above suggest that there could be different classes of signal peptides with different rates of evolution. However, although mitochondrial signal peptides are generally longer than non-mitochondrial ones, we failed to detect significant difference in the substitution rates of the two groups.

But what of the substitutions that we see, are these neutral? We could not falsify the hypothesis that the majority of non-synonymous mutations conserve hydrophobicity. We cannot therefore falsify the hypothesis that the substitutions are neutral. But neither does this permit us to falsify the hypothesis that they are under selection.

That neutral evolution may not be the only process going on in signal sequences is suggested by the fact that the absolute rate of protein evolution is correlated in mature and signal peptides. The best interpretation of the data that we can imagine is that there is some form of compensatory evolution going on: a change to the amino acids in the signal sequence might select for a change in the mature peptide or vice versa. While the activity of the mature peptide is independent of the signal peptide after the signal has been cleaved, prior to cleavage there may be selection on, for example, secondary or tertiary structure. There may, for example, be changes in the signal peptide that affect the activity of the mature peptide and/or vice versa. These would have to act prior to delivery of the mature peptide. Alternatively, the correlation between K_A for the mature and signal peptides might indicate genomic regional variation in the strength of the stabilizing selection.

We also find that signal peptides of immune system genes have unusually high rates of evolution. This is consistent with the hypothesis that some of the substitutions are driven by (or associated with) antagonistic co-evolution. It has been previously shown that coding regions of immune system genes tend to have high K_A/K_S ratios (Kuma et al., 1995; Hurst and Smith, 1999), a result that we can confirm. This can be accounted for by arguing that at least some part of the genes are under strong directional selection driven by host–parasite coevolution. It is perhaps surprising that the signal peptides of immune specific genes also evolve at an unusually high rate. This might however indicate, as before, that some of the frequent adaptive changes in the mature peptide regions cause slight disruptions in the secondary or the tertiary structure, that might select for compensatory changes in the signal sequences. Alternatively, one might speculate that the optimal cellular location (e.g., cytoplasm or membrane) of immune genes has been changed regularly as a response to new parasites.

Finally, we examined a particular selectionist hypothesis for the evolution of the signal sequence of *Igf2r*. Generally, were one to find rapid evolution (i.e. a high K_A/K_S ratio) of imprinted genes, it would provide reasonable support (McVean and Hurst, 1997) for the conflict theory for the evolution of imprinting, given that so many conflicts, for example maternal–foetal conflict (Hurst and McVean, 1998), do result in rapid evolution (but see also Haig, 1997). A previous study (McVean and Hurst, 1997) revealed that seven imprinted genes are not especially fast evolving. Further analysis by Smith and Hurst (1999) of 15 imprinted genes supported this broad conclusion, while noting that *Mash2* did have a rate of evolution comparable with immune system genes.

An earlier analysis (Smith and Hurst, 1998) indicated that Igf2r's signal peptide was an outlier, given the rate of evolution of the complete gene. We could find no evidence to indicate that this was an outlier in this larger data set, although its K_A/K_S was in the top 15% or so. However, for a non-immune gene it does appear to be fast evolving. Igf2r appears to have a signal peptide whose rate of evolution is higher than the majority of immune genes and comparable to that of the faster evolving ones. Given, too, that in the human–cow comparison the signal sequence also shows fast evolution (eliminating statistical artifact as an explanation), this result suggests that the rate of evolution of Igf2r's signal sequence might need special explanation. Examination of intra-population variation should help establish whether selection is acting on this sequence.

5. Conclusion

In summary, then, despite the fact that many random sequences function as signal peptides, we can certainly rule out the notion that signal peptides are paradigms of neutral evolution. Perhaps this is not surprising in retrospect, given that they are functional. Those putative signal peptides lacking the hydrophobic core evolve slowly at rates comparable to mature peptides. In part, this may be more a case of mistaken identity and an artifact of mis-annotation in GenBank. The remainder may be more nearly neutrally evolving than most sequences, but the unexpected correlation between the rate of protein evolution in the mature and the signal peptide suggests the unexpected possibility of compensatory evolution, suggesting that some of the non-synonymous substitutions could be the result of selection. This is supported by the finding that immune genes have high rates of evolution in their signal peptides and by the finding that Igf2r also has a fast evolving signal peptide for a non-immune gene.

Appendix A: The 76 mouse–rat orthologues and their substitution rates

Gene name (mouse)	Mouse Accession No.	Rat Accession No.	Mature peptide			Signal peptide			Signal peptide size	Rat cds size
			K_A/K_S	K_S	K_A	K_A/K_S	K_S	K_A		
Sterol carrier protein 2, liver[a]	M91458	M62763	0.068	0.291	0.020	0.45	0.057	0.026	60	432
Acetyl coenzyme A dehydrogenase, medium chain[a]	U07159	J02791	0.042	0.349	0.015	0.113	0.085	0.010	75	1266
Inhibin alpha[a]	X69618	M36453	0.103	0.228	0.024	0.17	0.166	0.028	699	1101
Inhibin beta-A[a]	X69619	M37482	0	0.112	0.000	0.27	0.111	0.030	924	1275
Coagulation factor III[a]	M26071	U07619	0.502	0.201	0.101	0.309	0.317	0.010	81	888
Ephrin B1[a]	U12983	U07560	0.061	0.138	0.008	0	0	0	23	1038
Small inducible cytokine A11[b]	U26426	U96637	0.189	0.075	0.014	1.596	0.041	0.065	69	294
Small inducible cytokine B subfamily, member 5[b]	u27267	u90448	0.694	0.474	0.683	0.719	0.163	0.117	120	393
Oxytocin[b]	m88355	m67442	0.076	0.194	0.015	0	0.131	0	57	378
Interleukin 4 receptor, alpha	m29854	x69903	0.708	0.179	0.127	0.544	0.152	0.083	75	2412
Low density lipoprotein receptor	x64414	x13722	0.263	0.238	0.063	0.374	0.124	0.046	63	2640
Tumour necrosis factor receptor superfamily, 1a	M60468	M63122	0.49	0.197	0.097	0.302	0.064	0.019	87	1386
Calreticulin	x14926	x53363	0.067	0.109	0.007	0	0.059	0	51	1251
Glutamate dehydrogenase	x57024	x14223	0.022	0.212	0.005	0.129	0.159	0.021	159	1677
Cathepsin E	x97399	D38104	0.241	0.152	0.036	0.878	0.295	0.259	57	1098
Insulin-like growth factor binding protein 5	x81583	m62781	0.015	0.105	0.002	0.556	0.051	0.028	57	816
Thyroid stimulating hormone receptor	u02602	m34842	0.13	0.218	0.028	0.761	0.185	0.141	63	2295
Gamma-aminobutyric acid receptor, subunit gamma 2[b]	m86572	l08497	0.019	0.219	0.004	0.504	0.048	0.024	114	1401
Gamma-aminobutyric acid receptor, subunit alpha 1[b]	m86566	l08490	0	0.148	0.000	0	0.165	0	141	1368
Myelin/oligodendrocyte glycoprotein (MOG)	u64572	m99485	0.108	0.203	0.022	0.442	0.268	0.118	81	738
Activin A receptor type II-like kinase 1	l48015	l36088	0.085	0.154	0.013	0.593	0.205	0.122	63	1515
Activin A receptor type II-like kinase	L15436	L19341	0.056	0.189	0.011	∞	0	0.63	45	1530
KGF-7[b]	u58503	x56551	0.123	0.124	0.015	0.264	0.168	0.044	75	585
Insulin like growth factor 2 receptor[b]	u04710	u59809	0.176	0.188	0.033	1.424	0.069	0.098	96	7449
Decay accelerating factor 1[b]	l41366	af039583	0.813	0.229	0.186	0.663	0.265	0.176	102	1200
Beta-glucuronidase structural	J02836	m13962	0.253	0.237	0.060	0.683	0.251	0.171	66	1947
Endothelin-1	D43775	m64711	0.214	0.262	0.056	0.504	0.062	0.031	51	609
Glycoprotein hormones, alpha subunit	M22992	j00757	0.077	0.227	0.017	1.535	0.044	0.067	69	363
Carboxyl ester lipase[b]	u33169	m69157	0.228	0.195	0.044	0	0.136	0	60	1839
Surfactant associated protein D	l40156	m81231	0.231	0.188	0.044	0.222	0.235	0.052	57	1125
5′ nucleotidase	L12059	J05214	0.149	0.185	0.028	1.094	0.118	0.129	84	1731
Secretory granule neuroendocrine protein 1, 7B2 protein	X15830	M63901	0.018	0.114	0.002	0.673	0.148	0.010	72	633
Insulin receptor	J05149	M29014	0.024	0.19	0.005	0.46	0.98	0.451	78	4152
Lysosomal membrane glycoprotein 1	M25244	M34959	0.465	0.203	0.094	1.249	0.217	0.271	63	1224
Luteinizing hormone receptor	M81310	M26199	0.148	0.1901	0.028	0.685	0.131	0.090	78	2103
Leukemia inhibitory factor	X12810	M32748	0.209	0.265	0.055	0.194	0.127	0.025	66	609
Mannose binding lectin, serum (C)	D11440	M14103	0.645	0.202	0.131	0.117	0.817	0.096	54	735
Myelin-associated glycoprotein	M31811	M16800	0.096	0.165	0.016	0.135	0.231	0.031	48	1881
Matrix gamma-carboxyglutamate (gla) protein	D00613	J03026	0.412	0.174	0.072	∞	0	0.03	57	312
Leptin	U18812	D45862	0.184	0.101	0.019	0	0.217	0	63	504
Secreted phosphoprotein 1	X16151	M14656	0.397	0.214	0.085	0.74	0.163	0.121	66	954
Transthyretin	D00073	K03252	0.155	0.261	0.040	0.249	0.217	0.054	60	444
Pancreatitis-associated protein	D13509	M55149	0.166	0.3	0.050	0.177	0.632	0.112	78	528
Parathyroid hormone receptor	X78936	M77184	0.049	0.132	0.006	0.358	0.071	0.025	63	1776
Uteroglobin	L04503	J05536	0.624	0.111	0.069	0.338	0.063	0.021	57	291
Pancreatic polypeptide	M18208	M13588	0.195	0.493	0.096	0.607	0.087	0.053	87	297
Prolactin receptor	L13593	M57668	0.319	0.148	0.047	0.237	0.297	0.07	114	1833
Selectin, platelet	M87861	L23088	0.228	0.227	0.052	0.326	0.271	0.088	123	2307
Rat regenerating islet-derived, mouse homologue 1	D14010	M62930	0.439	0.157	0.069	0.281	0.455	0.128	63	498
Insulin-like growth factor binding protein 6	X81584	M69055	0.224	0.13	0.029	0.132	0.314	0.041	75	681
CD1d1 antigen	M63695	D26439	0.397	0.221	0.086	1.801	0.057	0.103	54	1011
Cytochrome C oxidase, subunit Vb	X53157	D10952	0.19	0.134	0.025	0.632	0.169	0.107	93	390
Secreted acidic cysteine rich glycoprotein	X04017	D28875	0.11	0.136	0.015	0	0	0	51	906
Mast cell protease 7	L00653	D38455	0.302	0.643	0.201	0.781	0.417	0.326	57	825
Granzyme B	X04072	M34097	0.542	0.204	0.111	1.232	0.134	0.165	60	747
Kallikrein-3, plasma	M58588	M58590	0.275	0.189	0.052	0.2	0.254	0.051	57	1917
Receptor tyrosine kinase	U18933	D37880	0.119	0.125	0.015	0.119	0.13	0.015	93	2643
Thrombopoietin	L34169	D32207	0.621	0.131	0.081	0.932	0.185	0.172	63	981
CD3 antigen, zeta polypeptide	J04967	D13555	0.217	0.118	0.026	0.214	0.431	0.092	63	495
TGF-alpha	M92420	M31076	0.03	0.112	0.003	1.038	0.025	0.026	114	480
Acid phosphatase 5, tartrate resistant	M99054	M76110	0.098	0.25	0.025	0.689	0.243	0.167	63	984
UDP-glucuronosyltransferase 1 family, member 1	U09930	J02612	0.191	0.184	0.035	0.058	0.232	0.013	210	1590
Mouse vasopressin-neurophysin II	M88354	M25646	0.103	0.176	0.018	0.314	0.168	0.053	57	495
Lymphocyte antigen 84	Y07519	U04319	0.587	0.176	0.103	0.69	0.194	0.134	78	1011
Small inducible cytokine A5	X70675	U06436	0	0.161	0.000	∞	0	0.109	66	279
Follistatin-like polypeptide	M91380	U06864	0.122	0.16	0.020	0.271	0.09	0.024	54	921
Carbonic anhydrase 5, mitochondrial	X51971	U12268	0.345	0.216	0.0750	2.085	0.072	0.150	102	915
Biglycan	L20276	U17834	0.021	0.108	0.002	0.683	0.046	0.031	57	1110
Immunglobulin CTLA-4	X05719	U37121	0.233	0.157	0.037	1.79	0.054	0.010	111	672
Acetyl coenzyme A dehydrogenase, short chain	L11163	J05030	0.039	0.261	0.010	0.261	0.422	0.110	78	1245
Orosomucoid 1	M27008	J00696	0.677	0.24	0.162	0.443	0.279	0.124	54	618
Islet amyloid polypeptide	M25389	J04544	0.11	0.243	0.027	0.612	0.159	0.097	111	282
Apolipoprotein A-IV	M64249	M00002	0.449	0.235	0.106	1.2	0.068	0.082	60	1176
Calcium binding protein, intestinal	J05186	M86870	0.133	0.168	0.022	0	0	0	72	1932
Casein kappa	M10114	K02598	1.5	0.102	0.153	0	0.275	0	63	537
Matrix metalloproteinase 7	L36244	L24374	0.284	0.223	0.063	7.771	0.026	0.202	60	804

[a] The first six entries are those genes with signal peptides with unusual hydrophobicity plots.

[b] The mouse signal peptide was used for the analysis, due to the lack of annotated signal peptide region of the rat orthologue.

Appendix B: Mitochondrial signal peptides

Gene name (mouse)	Mouse Accession No.	Rat Accession No.	K_A/K_S mature peptide	K_S mature peptide	K_A/K_S signal peptide	K_S signal peptide	Signal peptide size
Carbonic anhydrase	X51971	U12268	0.345	0.216	2.085	0.072	102
Ornithine carbamoyltransferase	M17030	K00001	0.079	0.108	1.49	0.06	96
ATP synthase alpha subunit	L01062	J05266	0.026	0.306	0.809	0.056	99
ATP synthase coupling factor 6	U77128	M73030	1.469	0.04	0.58	0.116	96
Malate dehydrogenase	M16229	X04240	0.049	0.206	0.224	0.095	72
Aspartate aminotransferase isoenzyme	J02622	J02622	0.029	0.183	0.203	0.177	87
ATP synthase subunit c	L19737	D13123	0.062	0.11	0.163	0.141	183
FAD-linked glycerol-3-phosphate dehydrogenase (Gdm1)	U60987	U08027	0.097	0.249	0.264	0.097	126

Appendix C: The 10 fastest evolving (highest K_A/K_S) signal peptides

Ranked K_A/K_S of signal peptide regions	Name	Expression pattern/effect of mutation	Classification
67	Igf2r	Embryo, placenta, nervous system, etc.	Biochemical: receptor
68	Glycoprotein hormones, alpha subunit	Produced in both gonadotrophs and thyrotrophs/ endocrine defects, growth defects, obesity	Physiological: glycoprotein hormone
69	Small inducible cytokine A11	Immune system	
70	Immunglobulin CTLA-4	Immune system (immunoglobulin superfamiliy)	Glycoprotein
71	CD1d1 antigen	Immune system (CD1 antigen)	Surface glycoprotein
72	Carbonic anhydrase 5, mitochondrial	Housekeeping gene	Mitochondrial biochemical enzyme
73	Matrix metalloproteinase 7	Thymus, speen, liver, placenta, uterus mammalian gland	Biochemical: enzyme
74	Activin A receptor type II-like kinase 2	Embryo (growth factor receptor?)	Biochemical: receptor
75	Matrix gamma-carboxyglutamate (gla) protein	Osteoblasts during embryogenesis	Biochemical: emzyme
76	Small inducible cytokine A5	Immune system	

Acknowledgement

We wish to thank the ESF's TBA program for providing resources for C.P. L.D.H. is funded by the Royal Society.

References

Aguade, M., Miyashita, N., Langley, C.H., 1992. Polymorphism and divergence in the mst26a male accessory-gland gene region in *Drosophila*. Genetics 132, 755–770.

Dickerson, R.E., 1971. The structure of cytochrome c and the rates of molecular evolution. J. Mol. Evol. 1, 26–45.

Duret, L., Mouchiroud, D., Gouy, M., 1994. HOVERGEN — a database of homologous vertebrate genes. Nucleic Acid. Res. 22, 2360–2365.

Freeland, S.J., Hurst, L.D., 1998. Load minimization of the genetic code: history does not explain the pattern. Proc. R. Soc. Lond. B 265, 2111–2119.

Gillham, N.W., 1995. Organelle Genes and Genomes. Oxford University Press, Oxford.

Haig, D., 1997. Parental antagonism, relatedness asymmetries and genomic imprinting. Proc. R. Soc. Lond. B 264, 1657–1662.

Haig, D., Hurst, L.D., 1991. A quantitative measure of error minimization in the genetic code. J. Mol. Evol. 33, 412–417.

Hughes, A.L., 1991. Circumsporozoite protein genes of malaria parasites (*Plasmodium* spp.): evidence for positive selection on immunogenic regions. Genetics 127, 345–353.

Hughes, A.L., 1992. Positive selection and interallelic recombination at the merozoite surface antigen-1 (msa-1) locus of plasmodium-falciparum. Mol. Biol. Evol. 9, 381–393.

Hughes, A.L., Ota, T., Nei, M., 1990. Positive Darwinian selection promotes charge profile diversity in the antigen binding cleft of class I MHC molecules. Mol. Biol. Evol. 7, 515–524.

Hurst, L.D., McVean, G.T., 1998. Do we understand the evolution of genomic imprinting? Curr. Opin. Genet. Dev. 8, 701–708.

Hurst, L.D., Smith, N.G.C., 1999. Do essential genes evolve slowly? Curr. Biol. 9, 747–750.

Izard, J.W., Kendall, D.A., 1994. Signal peptides — exquisitely designed transport promoters. Mol. Microbiol. 13, 765–773.

Kaiser, C.A., Preuss, D., Grisafi, P., Botstein, D., 1987. Many random sequences functionally replace the secretion signal sequence of yeast invertase. Science 235, 312–317.

Kuma, K., Iwabe, N., Miyata, T., 1995. Functional constraints against variations on molecules from the tissue-level — slowly evolving

brain-specific genes demonstrated by protein-kinase and immunoglobulin supergene families. Mol Biol Evol. 12, 123–130.

Ladunga, I., 1999. PHYSEAN: PHYsical SEquence ANalysis for the identification of protein domains on the basis of physical and chemical properties of amino acids. Bioinformatics 15, 1028–1038.

Li, W.-H., 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. 36, 96–99.

Li, W.-H., Gojobori, T., Nei, M., 1981. Pseudogenes as a paradigm of neutral evolution. Nature 292, 237–239.

Li, W.-H., Wu, C.-I., Luo, C.-C., 1985. A new method for estimating synonymous and non-synonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. 2, 150–174.

Makalowski, W., Boguski, M.S., 1998. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. Proc. Natl. Acad. Sci. USA 95, 9407–9412.

McVean, G.T., Hurst, L.D., 1997. Molecular evolution of imprinted genes: no evidence for antagonistic coevolution. Proc. R. Soc. Lond. B 264, 739–746.

Moore, T., Haig, D., 1991. Genomic imprinting in mammalian development: a parental tug-of-war. Trends Genet. 7, 45–49.

Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G., 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. 10, 1–6.

Pamilo, P., Bianchi, N.O., 1993. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol. Biol. Evol. 10, 271–281.

Smith, N.G.C., Hurst, L.D., 1998. Molecular evolution of an imprinted gene: repeatability of patterns of evolution within the mammalian insulin-like growth factor type II receptor. Genetics 150, 823–833.

Smith, N.G.C., Hurst, L.D., 1999. The causes of synonymous rate variation in the rodent genome: Can substitution rates be used to estimate the sex bias in mutation rate? Genetics 152, 661–673.

Tan, Y., Riley, M.A., 1997. Positive selection and recombination: major molecular mechanisms in colicin diversification. Trends Ecol. Evol. 12, 348–351.

Tsaur, S.C., Wu, C.I., 1997. Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of *Drosophila*. Mol. Biol. Evol. 14, 544–549.

Wolfe, K.H., Sharp, P.M., 1993. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J. Mol. Evol. 37, 441–456.