the FEBS Journal

Single amino acid repeats in signal peptides

Paweł P. Łabaj[1], Germán G. Leparc[1], Anaïs F. Bardet[1], Günther Kreil[2] and David P. Kreil[1]

1 Chair of Bioinformatics, Boku University Vienna, Austria
2 The Institute of Physiology, Paracelsus Medical University, Salzburg, Austria

Keywords
amino acid compositional bias; protein evolution; protein sequence analysis; signal peptides; single amino acid repeats

Correspondence
D. P. Kreil, Chair of Bioinformatics, Boku University Vienna, 1190 Muthgasse 18, Austria
Fax: +43 47654 6847
Tel: +43 47654 6200
E-mail: saars10@boku.ac.at

(Received 11 January 2010, revised 19 April 2010, accepted 24 May 2010)

doi:10.1111/j.1742-4658.2010.07720.x

There has been an increasing interest in single amino acid repeats ever since it was shown that these are the cause of a variety of diseases. Although a systematic study of single amino acid repeats is challenging, they have subsequently been implicated in a number of functional roles. In general surveys, leucine runs were among the most frequent. In the present study, we present a detailed investigation of repeats in signal peptides of secreted and type I membrane proteins in comparison with their mature parts. We focus on eukaryotic species because single amino acid repeats are generally rather rare in archaea and bacteria. Our analysis of over 100 species shows that repeats of leucine (but not of other hydrophobic amino acids) are overrepresented in signal peptides. This trend is most pronounced in higher eukaryotes, particularly in mammals. In the human proteome, although less than one-fifth of all proteins have a signal peptide, approximately two-thirds of all leucine repeats are located in these transient regions. Signal peptides are cleaved early from the growing polypeptide chain and then degraded rapidly. This may explain why leucine repeats, which can be toxic, are tolerated at such high frequencies. The substantial fraction of proteins affected by the strong enrichment of repeats in these transient segments highlights the bias that they can introduce for systematic analyses of protein sequences. In contrast to a general lack of conservation of single amino acid repeats, leucine repeats were found to be more conserved than the remaining signal peptide regions, indicating that they may have an as yet unknown functional role.

Introduction

The role of single amino acid repeats (SAARs) in human hereditary diseases has become of increasing interest. The best known example is Huntington's disease where a sequence of more than 35–40 glutamines in the protein huntingtin, encoded by a CAG repeat in the corresponding gene, leads to the death of neurones [1,2]. Several other neurodegenerative diseases are similarly caused by CAG repeats coding for polyglutamine stretches in the respective genes [3,4]. Expansion of trinucleotide repeats coding for polyalanine have also been shown to be responsible for diseases [3,5,6]. Remarkably, repeats of aspartic acid as short as five residues have been implicated in disease [7], indicating a functional role. These findings have led to a growing interest in SAARs in general. Which amino acids form unexpected repeats and what is the functional significance of these structural features?

Already in studies of individual species, leucine repeats were observed to be of conspicuously high frequency [8–10]. Remarkably, in a recent systematic survey (COPASAAR) [11], leucine repeats were the

Abbreviations
SAAR, single amino acid repeat.

most frequent type of SAARs in a wide range of species covering all three kingdoms.

Signal peptides, which are located at the amino terminus of the polypeptide chain of secreted and many membrane proteins, have a middle part that is rich in amino acids with hydrophobic side chains (Fig. S1). This part is essential for the interaction with the signal recognition particle and, subsequently, the translocase complex embedded in the membrane of the endoplasmic reticulum. Upon transit through this membrane, the signal peptide is cleaved from the nascent polypeptide chain by signal peptidases and is then in most cases rapidly degraded by other proteases [12–14].

We were interested in determining the extent to which the conspicuously high frequency of leucine repeats could be explained by repeats of hydrophobic amino acids in signal peptides. In the present study, we show that leucine repeats are indeed over-represented in the transient signal peptides of animal species, where they occur in a substantial fraction of all proteins, thus constituting a source of bias in any systematic sequence analysis of the proteome. We further show that, surprisingly, these repeats are more strongly conserved than the surrounding sequence regions, indicating that they may have an as yet unknown function.

Results

Signal peptides and SAARs

Excluding methionine as the initiating amino acid, as well as serine, which is usually found near the cleavage site, the most frequent amino acids in signal peptides are leucine, alanine, valine, phenylalanine and isoleucine (Fig. S1). Our analysis thus focused on these five residues.

In a comparison of the three kingdoms, bacteria and archaea typically had much fewer SAARs than eukaryotes. Although approximately half the eukaryotes had more than ten SAARs of the investigated amino acids per 100 proteins, this could be observed for only 1.0% and 1.9% of *Bacteria* and *Archaea*, respectively (Table S1). We therefore focused on eukaryotic organisms, comparing SAAR frequencies in secreted and type I membrane proteins with and without their respective signal peptides. In view of the under-annotation of signal peptides in the TrEMBL part of UniProt (see annotation of signal peptides in Supporting Information), signal peptides were systematically predicted *de novo* using SIGNALP, version 3.0 [15–17] for an unbiased genome-wide survey.

Repeat location: selective enrichment in signal peptides

We have analyzed the genomes of 102 eukaryotic organisms (Table S2) for repeats of single amino acids with a length of five residues or more (see Materials and methods). Approximately 12% of all proteins had a signal peptide. We found that these proteins, however, contained half of all leucine repeats (49.9%). In comparison, they contained only 10% of all alanine repeats, 16% of valine repeats and 17% each of isoleucine and phenylalanine repeats.

Surprisingly, in these proteins, 82% of leucine repeats were located in the signal peptide, which amounts to $(1/2) \times 82\% = 41\%$ of all leucine repeats. This observation was even more pronounced for some higher organisms, such as human where approximately two-thirds of all leucine repeats were located in signal peptides (Table S2). Much smaller ratios were generally observed for the other four amino acids (Fig. 1).

In view of the high proportion of leucine repeats located in signal peptides, the question arises as to whether this would be expected because of differences in the amino acid composition of these peptides and the mature proteins. The ratio of observed to expected frequencies of repeats in signal peptides for selected model organisms and taxonomical subgroups of eukaryotes (see Materials and methods) is shown in Fig. 2. All enrichment scores of 2 or higher significantly indicated strong enrichment ($P < 5\%$). In general, a strong enrichment was only detected for leucine repeats. In signal peptides of plants, however, an excess of repeats of phenylalanine and valine also was

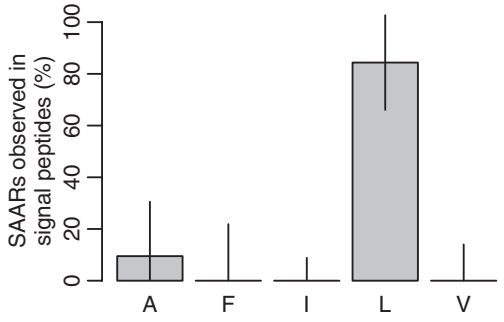

Fig. 1. Fraction of single amino acid repeats in the signal peptide. The percentage of single amino acid repeats in proteins with signal peptides that were found to be located in the signal peptide is shown. For SAARs of A, F, I, L and V, the bars each plot the median percentage for the examined 102 *Eukaryota*, and the whiskers indicate the average absolute deviation (Table S6). For a comparison of observed and expected frequencies, see Fig. 2.

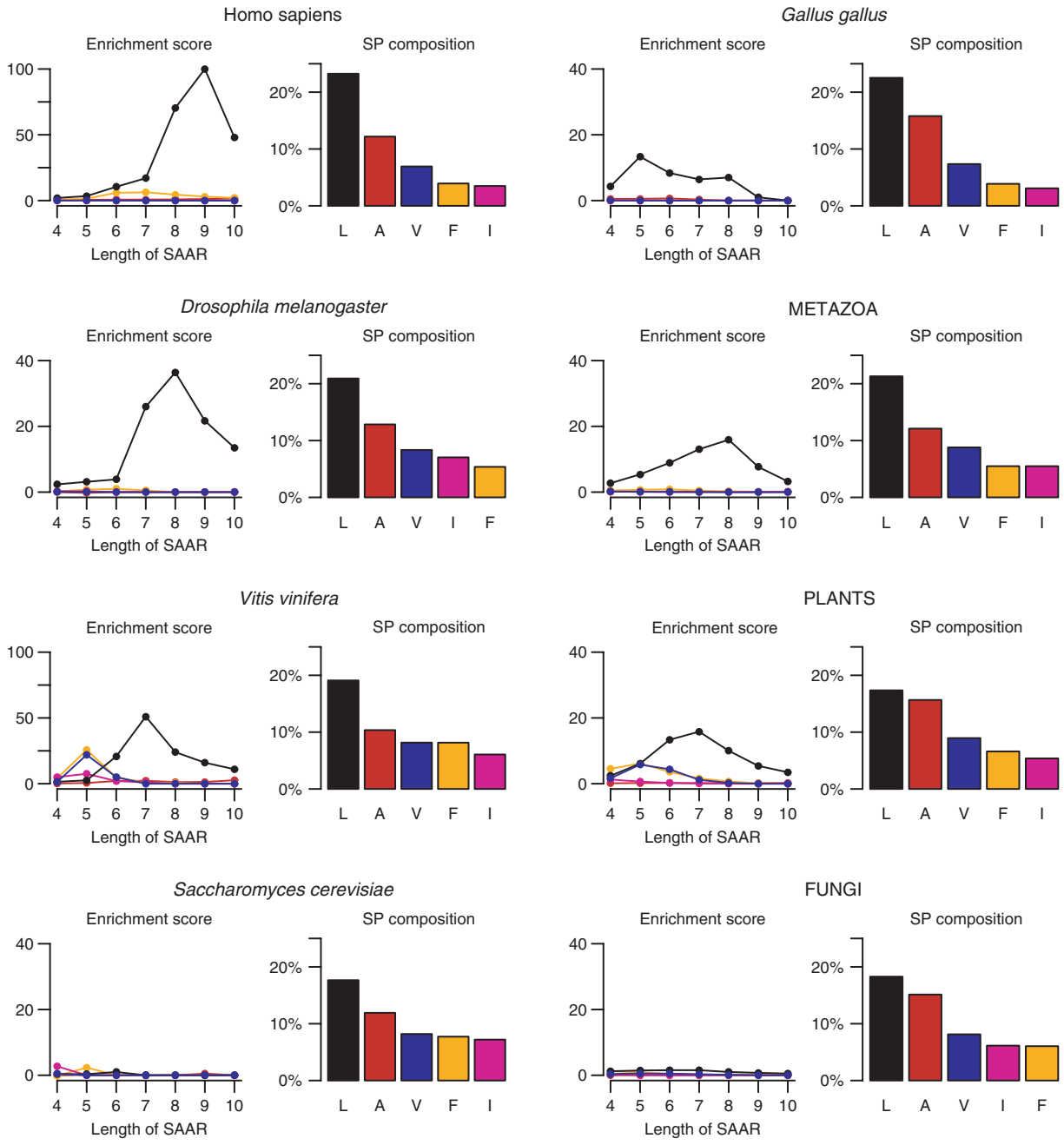

Fig. 2. Enrichment of single amino acid repeats in signal peptides. For each of the selected eukaryotic organisms or taxonomical subgroups, two graphs are presented. Colour codes for residue type (A, F, I, L or V). The left graph shows repeat enrichment scores as a function of repeat length (see Materials and methods). The bar chart displays the average signal peptide amino acid composition.

found. No general trend for any type of repeat enrichment was observed in *Saccharomyces cerevisiae*, and this was also true for fungi in general.

In a detailed examination of individual species (Fig. S2), many species showed a robust enrichment of leucine repeats in signal peptides. This affected 68% of the metazoa and 50% of plants. Remarkably, a clear enrichment of leucine repeats in signal peptides was found in all the tetrapods.

Conservation of leucine repeats

In view of these findings, the question arises as to whether L-repeats in signal peptides have a function in

tetrapods, in which case a conservation of the repeats would be expected. The human proteins with an L-repeat in their signal peptide available from Swiss-Prot (http://www.expasy.org/sprot/) yielded a well annotated nonredundant test set of 225 sequences. We then examined the 1183 orthologues that could be identified in the five tetrapods (i.e. chimpanzee, mouse, cow, chicken and frog) and, for comparison, in five other species: zebrafish, *Ciona intestinalis*, *Drosophila melanogaster*, *Caenorhabditis elegans* and baker's yeast.

Presence of leucine repeats

For each of the considered species, we determined the percentage of orthologues containing an L-repeat in their signal peptides (Fig. 3, black bars). In mammals, the L-repeats were well conserved, with most orthologues featuring a leucine repeat. The percentage of orthologues with L-repeats, however, was much lower

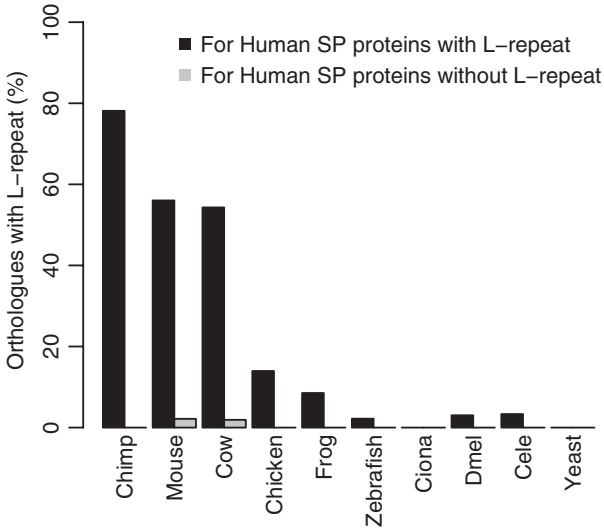

Fig. 3. Fraction of orthologues with an L-repeat in their signal peptides. For each species, the percentage of orthologues identified that contained an L-repeat in their signal peptides is shown. As a reference, the grey bars give the results for orthologues of human proteins with a signal peptide having no L-repeat. The black bars display the results for the test set of human proteins with a signal peptide having an L-repeat, showing a phylogenetic pattern. Many leucine repeats were found within signal peptides in organisms from the *Tetrapodae* group. Moreover, for the mammalian organisms, at least half of the studied leucine repeats within signal peptides were well conserved. Conversely, leucine repeats were not conserved in the other examined eukaryotes, where conservation was observed only for a minority of the 333 studied human proteins, with four in zebrafish, two in *Caenorhabditis elegans*, three in *Drosophila melanogaster*, and none in *Saccharomyces cerevisiae* and *Ciona intestinalis*; in each case comprising no more than 3% of orthologues identified in these organisms.

in other vertebrates and in invertebrates. As a control, orthologues of a set of human proteins with signal peptides but no L-repeat were tested (Fig. 3, grey bars). In the control set, very few orthologues with an L-repeat in their signal peptides were found. This suggests that many leucine repeats in signal peptides appeared during the evolution of higher eukaryotes and may have a functional role.

Stronger conservation of repeat sequences

To test this hypothesis further, we next studied the conservation of individual L-repeats in signal peptides, with a particular interest in a comparson of the conservation of the repeat with that of the remaining signal peptide sequence. Accordingly, the quality of signal peptide sequence alignments with and without the L-repeats in question was assessed (see Materials and methods). Figure 4A plots the average L-repeat conservation scores for the examined organisms. In contrast to the lower, uniform scores of nontetrapods ($40 \pm 3\%$), the scores for tetrapods are clearly distinct and increase from frog (46%) to chimpanzee (94%). This provides a sequence alignment-based confirmation of the phylogenetic pattern observed in the L-repeat frequencies shown in Fig. 3.

The average relative conservation score for each of the examined organisms is shown in Fig. 4B. Here, zero indicates that all regions of the signal peptide were conserved similarly well, whereas an L-repeat that is better conserved than the remaining signal peptide yields a positive score. For the organisms closest to human, some signal peptides were identical. These, however, cannot provide information about the conservation of an L-repeat relative to the remaining signal sequence. Identical signal peptides were found for 82, two and four orthologues in chimpanzee, mouse and cow, respectively. Consequently, we plot an average score where these perfectly conserved signal peptides have been excluded. In conclusion, a relative conservation of L-repeats was clearly observed for all tetrapods. Conversely, the relative conservation scores for nontetrapods were considerably lower.

Test for functional associations

We tested for associations of proteins with leucine repeats in signal peptides to particular Gene Ontology categories. Leucine repeats in signal peptides were respectively observed in 55 (48%), 11 (48%) and 44 (27%) of all level 3 annotation groups for the categories of Biological Process (114 groups), Cellular Component (23 groups) and Molecular Function (164

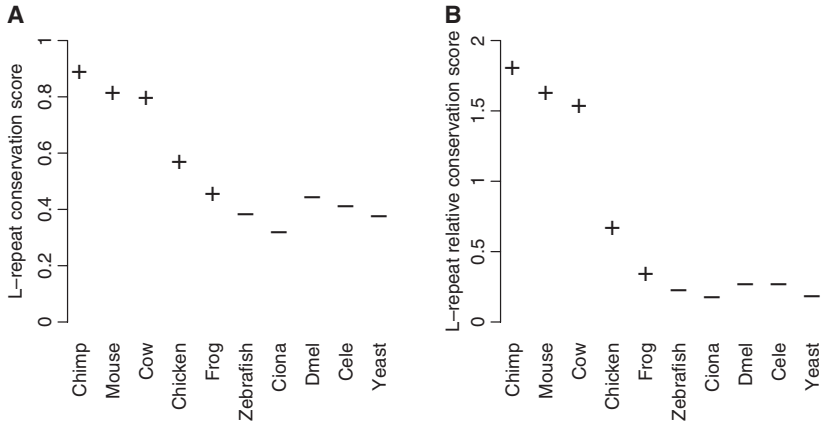

Fig. 4. Sequence alignment-based conservation scores. (A) Average L-repeat conservation scores for the examined species, comprising five tetrapods (+) and five nontetrapods (–). (B) Average L-repeat relative conservation scores (see Materials and methods). For the organisms closest to human, some signal peptides were identical. Perfectly matching signal peptides cannot provide information about the conservation of an L-repeat relative to the remaining signal sequence because both show 100% sequence identity. Consequently, scores excluding these proteins are shown for tetrapods. Similarity-based scores are shown here. Analogous scores based on identity% are shown in Fig. S4.

groups). In comparison, proteins with signal peptides without an L-repeat were present in 71 of 114 groups (62%), 14 of 23 groups (61%) and 79 of 164 groups (48%) of the respective categories. Signal peptides with L-repeats were thus found in a large variety of proteins, similar to signal peptides without an L-repeat. Only a small number of Gene Ontology groups showed a significant over-representation of signal peptides with L-repeats, with no group accounting for more than 4% of the data set (Table S3). For example, there were eight nucleases with an L-repeat in their signal peptides (six related ribonucleases and two from other families), together making up 2% of the test set, versus four nucleases in the 3181 proteins with a signal peptide having no L-repeat.

Discussion

Although SAARs are surprisingly abundant in protein sequences, relatively little is known about their functions. It has been suggested that many may have no effect and might just be tolerated [9,10,18,19]. On the other hand, there is a considerable body of evidence confirming the functional roles of repeats [8,9,19,20] and other compositionally biased or unstructured sequence regions [21]. It is known that repeat expansion beyond specific lengths can be associated with disorders affecting neurological, neuromuscular or developmental processes [3,4]. Conversely, there are instances where a lack of repeats can cause diseases [7]. Apart from the known implications in specific diseases, repeat length can be an important factor in transcriptional regulation [22,23] and protein interaction networks [24], affect morphological changes [25] and may facilitate adaptive processes [26,27]. As a result, a better understanding of SAARs is of considerable interest.

Corroborating earlier studies [8–10], in a recent systematic survey of SAARs [11], a high frequency of leucine repeats was observed throughout the three kingdoms of *Archaea*, *Bacteria* and *Eukaryota*. This is surprising because the cytotoxicity of amino acids repeats is highly correlated with their hydrophobicity and their length [28–30]. The observed toxicity may result from an aggregation of such repeats [29,31,32].

In previous studies, it was shown that SAARs generally show a positional bias for the termini of polypeptides. Leucine repeats, in particular, favour the amino end of proteins [8,32–34]. It is also well known that the amino end of the growing polypeptide chain of secreted and many membrane proteins contains a signal peptide, with a central part rich in hydrophobic amino acids. Consequently, a possible connection with leucine repeats has already been suggested in an earlier study of SAARs [8]. In the present study, we report the results of the first systematic study of SAARs in signal peptides.

Localization of SAARs in signal peptides

In general, archaeal and bacterial species contained few SAARs (Table S1), so the present study focused on eukaryotes. Other types of repeats are also more

common in eukaryotes [18,20,35]. For these organisms, the first striking observation is that secreted and type I membrane proteins, which constitute only 12% of all proteins, already contained half of the leucine repeats. Most of these repeats were located in the signal peptides themselves (82%). This effect was particularly pronounced for some higher organisms; for example, in the human proteome, this figure was 90%, corresponding to two-thirds of all leucine repeats (Table S2). Signal peptides are cleaved off early in the course of translation and then degraded rapidly [36]. This could explain why leucine repeats, which in general are toxic [28–30,32], are tolerated at such high frequencies. Further analysis confirmed a selective enrichment of leucine repeats in signal peptides that could not be explained by differences in the amino acid composition of these segments and the mature proteins. Although present only in some isolated fungi, this effect was fairly common in metazoa and plants (Fig. 2). It was particularly observed in vertebrates, such as in all the tetrapods (Fig. S2). Surprisingly, no enrichment of alanine repeats was detected, although these are the second most abundant type of repeats in signal peptides (Fig. 1). Repeats of valine and phenylalanine are generally rare, yet they were found to be enriched in the signal peptides of some plants (Fig. S2).

Conservation of leucine repeats

A comparison of signal peptides with leucine repeats in human proteins versus their orthologues in tetrapods and in other species confirmed that signal peptides evolve faster [37,38] than the mature parts (Fig. S3). Nevertheless, leucine repeats were conserved in tetrapods. This was particularly clear in mammals. More distantly-related species showed a prominent decline in the number of these repeats (Fig. 3). Conversely, essentially no leucine repeats were seen in orthologues of a control set. A further comparison based on actual signal peptide sequence alignments corroborated these observations (Figs 4A and S6A,C). Moreover, for all tetrapods, the conservation of leucine repeats was clearly higher than the conservation of the remaining sequence of the signal peptide, as indicated by an alignment based relative conservation score (Figs 4B, S6C,D).

In contrast to the instability of SAARs in general [10], the conservation of leucine repeats in signal peptides thus revealed a phylogenetic pattern in all our analyses, suggesting that leucine repeats in signal peptides appeared in the evolution of higher eukaryotes and may have a functional role. This would explain the unexpected high conservation observed for leucine repeats in a comparison of human and mouse proteins, although the repeats were particularly abundant in rapidly evolving genes [19].

Leucine repeats as potential additional secretion signal

It is understood that the hydrophobic region of the signal sequence is required for interaction with the signal recognition particle and studies have demonstrated that the efficiency of the secretion signal can be improved by mutations that increase its hydrophobicity [39] or through the insertion of hydrophobic amino acids [40]. However, recent work has uncovered a surprising complexity of signal sequences [14] and suggested additional functions such as in the modulation of protein biogenesis [41]. The emergence of conserved leucine repeats in signal peptides of higher eukaryotes reported in the present study may thus point to another purpose, although the exact role of the repeats remains to be investigated. In particular, future research needs to address the question of whether they have any positive or negative effects on the interaction efficiency with the signal recognition particle and transport across the membrane of the endoplasmatic reticulum. From our test of an association with particular Gene Ontology categories, which showed that signal peptides with leucine repeats were present in a large variety of proteins, it appears that the repeats may be involved in a general cellular mechanism. Eventually, this raises the question of why leucine repeats in signal peptides have increased in the course of the evolution of mammals.

Unusual transient regions can distort genome-scale sequence analysis

It has been widely recognized that low-complexity regions can generally affect studies of sequence similarity [42–44]. These regions are not always filtered, such as in searches with the current default settings of the BLAST web service at the National Center for Biotechnology Information (http://blast.ncbi.nlm.nih.gov/). Besides the well-known effects of nonspecific matches to biased regions, a decision to filter or not can also considerably affect the assessment of matches within the mature parts. Recent developments, for example, allow more accurate statistics by adjusting for the amino acid composition of proteins [45]. Bias in some sequence regions distorts these statistics. The present survey highlights the scope of this issue by demonstrating that leucine repeats are strongly over-represented in the signal peptides of secreted and type I membrane

proteins, particularly for tetrapods. These proteins constitute a substantial fraction of the proteome (19% in tetrapods). Similarly, the leader sequences of many proteins imported into mitochondria and chloroplasts are removed during post-translational transport into these organelles. As highlighted in the present report, for a systematic study of proteins deduced from genomic data, it is important to consider that transient parts, which are not present in the mature species, need to be separated because they may introduce bias into any subsequent sequence analysis.

Materials and methods

Data integration and feature identification

Primary sequence data, taxonomic information and all results of the analysis were managed in a customized InterMine data warehouse (http://www.intermine.org/) [46]. Protein sequences were loaded from UNIPROT [47], release 13.6 and local warehouse extensions allowed the connection of these sequences to the GeneBank taxonomy from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/Taxonomy/; accessed 8 August 2008). Information about orthologues was extracted from the EnsEMBL database [48] using the BIOMART [49] PERL application programming interface. To minimize artefacts, protein fragments were filtered out because they could particularly affect signal peptide prediction. To allow meaningful analyses for individual organisms, we focused on organisms with approximately 1000 proteins remaining, yielding a reference data set of 1 149 543 full-length proteins from 102 *Eukaryota*, 3 298 615 from 920 *Bacteria* and 118 539 from 57 *Archaea*.

Following Karlin [50], we count a SAAR of length n when an uninterrupted repeat of n identical amino acid residues is observed, independent of the adjacent amino acids. For example, in the amino acid sequence *ACDFLLLLLLLLGWS*, we count four leucine repeats of length five, three repeats of length six, two repeats of length seven and one repeat of length eight. We focus on repeats of five or more residues. Statistically, under the independent and identical distribution model, repeats shorter than five are already expected by random chance for typical protein lengths [50]. Incidentally, this also constitutes the shortest repeat-length that has been implicated in disease [7]. Leucine repeats were identified with custom PERL scripts.

Considering that tools for the prediction of signal peptides in *Archaea* have only recently been developed [51], our analysis focused on eukaryotes and bacteria, for which long established algorithms were available. For a conservative prediction of signal peptides, we combined the neural network and Hidden-Markov-Model predictors of SIGNALP, release 3.0 [15,16], which were applied using their default settings. We required the two prediction models to agree on the location of the cleavage site. Prediction scores had to meet the default threshold for at least one of the methods, with the worse score having to meet at least half the default threshold. Combining scores from both predictors extends sensitivity at the same time as allowing a discrimination of signal anchors, which is provided by the Hidden-Markov-Model component [17].

Feature statistics were computed in the 'R' statistical environment (http://www.r-project.org/). Unless mentioned otherwise, average results for all eukaryotic proteins are reported. Figures for individual organisms as well as summary statistics over all species are provided in Table S2 and Figure S2.

Normalization model for comparative plots

Background model

Longer repeats are much less likely to occur than shorter ones and a survey of amino acid repeats of varying lengths requires a way to adjust for this. It is well known, however, that the traditional assumption of positional independence breaks down in biased protein regions [42] and that the observed frequencies of any amino acid repeat exceed the counts expected from models of independent amino acid residues. To facilitate a comparison of amino acid repeat frequencies in average mature proteins with those in signal peptides, we developed an empirical model for visualization purposes. The model was trained on a set of nonsignal peptide proteins providing a comprehensive unbiased reference sample. Model input parameters comprised protein length, amino acid composition and repeat length. A two-stage approach was found to be efficient. First, a logistic regression is used to model the probability that a protein of given amino acid composition and length has at least one repeat of a particular residue and length. Then, a relevance vector machine [52] predicts the conditionally expected number of repeats. All regression and model testing was performed in the 'R' statistical environment. The standard function for fitting a generalized linear model (glm) was used for logistic regression. The relevance vector machine was trained using the rvm function of the established KERNLAB library [53]. A comprehensive model characterization and further details are provided elsewhere (P. P. Łabaj, P. Sykacek & D. P. Kreil, unpublished results). Model parameters were obtained from proteins without signal peptides or signal anchors.

Scoring over-representation

This model was then applied to calculate the expected frequencies of amino acid repeats in individual secreted proteins, both for the entire proteins (including signal peptide *plus* the mature part) and also for the mature parts alone. We can then easily plot observed counts relative to the

expected frequencies. The comparative graphs (Fig. 2) show differences of such normalized frequencies:

$d_{AA}(l)$ =

[f(entire protein, observed)/f(entire protein, expected)] – [f(mature protein, observed)/f(mature protein, expected)]

for several amino acids, AA, and a range of repeat lengths, l. The difference, d, reflects a relative enrichment of amino acid repeats in signal peptides. We emphasize, however, that the strong enrichment effect shown is observed *independently of the chosen normalization transform*. The discussed effects are already apparent in the raw data (File S1) but, without normalization, cannot be presented well graphically in a single panel for different repeat lengths.

Assessing significance

To assess the significance of the observed relative enrichment scores, d, we calculated empirical P-values. These were computed for the organisms and taxonomical subgroups examined in Fig. 2. For each amino acid repeat type and length, we computed 1 000 000 random scores from the full set of proteins without a signal peptide, providing a comprehensive unbiased reference sample, for which it was known that the observed frequencies would be similar to expected ones (i.e. no significant enrichment would be observed). We could thus quantify the likelihood of scores $d_{AA}(l) > d_0$ by an empirical P-value for each amino acid repeat type and length. As expected, higher scores were less likely by random chance. In particular, all enrichment scores $d_{AA}(l) > 2$ were significant ($P < 0.05$) (Table S4). We can therefore consider observations of strong enrichment $d_{AA}(l) > 2$ as indicators of significant enrichment.

Comparison of species

For the summary shown for every species in Fig. S2, repeats of a particular amino acid were considered to be enriched if the difference between normalized frequencies in the entire and the mature proteins met an arbitrary threshold, $d(l) > 2$, for at least two of the six examined repeat lengths, $l = 5 \ldots 10$, and the average difference also met that threshold, $(1/6)\ \Sigma d(l) > 2$. The reported results were robust under variation of the chosen threshold value.

Study of orthologues

To study the conservation of leucine repeats in signal peptides, a set of human secreted and type I membrane proteins was compiled from Swiss-Prot that had repeats of five or more leucines in their signal peptide. Orthologue and paralogue annotation was obtained from the EnsEMBL database (release 54) and human paralogues were filtered out, leaving a *test set* of 203 genes corresponding to 225 possible gene products (proteins). Similarly, a *reference set* of 1273 human proteins without leucine repeats in their signal peptide was selected, which served as a control.

The conservation of leucine repeats was then tested for the five studied tetrapods: chimpanzee (*Pan trogodlytes*), mouse (*Mus musculus*), cow (*Bos taurus*), chicken (*Gallus gallus*) and the frog *Xenopus tropicalis*, as well as for five more distantly-related organisms in comparison: zebrafish (*Danio rerio*), *C. intestinalis* (a transparent sea squirt), the fruit fly *D. melanogaster*, the worm *C. elegans* and baker's yeast (*S. cerevisiae*).

Repeat presence and sequence conservation

For all species, we compared the number of orthologues containing a leucine repeat in their signal peptide, looking for a phylogenetic pattern in the test set compared to the control. Then, the conservation of individual leucine repeats was assessed by sequence alignment of the signal peptides of each human protein and its orthologues. Considering the faster evolutionary rates observed in signal peptides [37,38], alignments were performed at the peptide level to avoid saturation. Manual inspection of multiple sequence alignments, which are the basis of multi-species, tree-based conservation scores [54], reflected the difficulty encountered when applying standard multiple alignment tools to this type of sequence. These regions evolve relatively fast, and they are of very unusual amino acid composition. Many common assumptions of established algorithms, such as the positional independence of residues, break down in compositionally biased regions [50]. We have therefore developed and validated an *ad hoc* approach based on pairwise sequence alignments, which were less severely affected by indels. This allowed species specific summaries to be made.

EMBOSS NEEDLE software (http://www.emboss.org/) implementing the Needleman–Wunsch global alignment algorithm [55] was run with default settings, using the EBLOSUM62 substitution matrix. The conservation of the signal peptide sequence without the L-repeat in question was measured by the human versus orthologue alignment score after removing the L-repeat (masking). Normalization relative to the perfect human–human alignment yields a standardized conservation score in the range from $x = 0$ (no conservation) to $x = 1$ (perfect conservation). Similarly, the conservation of L-repeats was assessed by comparing the human versus orthologue alignment scores for the entire signal peptide sequence and the sequence where the L-repeat had been removed (masked). This score difference (unmasked – L-repeat masked) was normalized relative to the perfect human–human alignment (unmasked – L-repeat masked), giving a standardized L-repeat conservation score in the range from $y = 0$ (no conservation) to $y = 1$ (perfect conservation). With these measures, we can introduce a relative conservation score, r, to test the conservation of L-repeats relative to the conservation of the hosting signal peptide:

$$r = \log_{10}[(1 - x)/(1 - y)]$$

In this context, a perfectly conserved signal peptide ($x = 1$, $y = 1$) gives a limit of $r = 0$. Similarly, $r = 0$ if the conservation of the L-repeat and the remaining signal peptide deteriorate equally ($x = y$). If, however, the L-repeat is conserved better (y is closer to 1) than the surrounding signal peptide (x is less close to 1), then r becomes positive. Conversely, a negative score would indicate that the L-repeat were less well conserved than the surrounding signal peptide sequence. Defining the score on a log-scale makes it symmetrical to changes in either x or y. Finally, two variants of the scores x, y and r were considered: one set computed for an EBLOSUM62-based alignment similarity score and one set based on alignment identity%.

A number of analyses were performed in order to demonstrate that our results were not biased by typical signal peptide sequences being similar to one another. In particular, we wanted to refute the conjecture that an observed strong conservation of L-repeats might just be the result of a high likelihood of finding leucines in the hydrophobic cores of signal peptides, or that these hydrophobic cores might inflate conservation measures. Accordingly, we compared the observed y-scores with corresponding scores calculated from alignments of human signal peptides from the test set with random signal peptides from the other ten organisms. The susceptibility of the similarity score to substitutions by similar amino acids also was not responsible for observed conservation signals, as demonstrated by a comparison of the analysis obtained results based on similarity scores with the obtained results based on an alignment identity% score instead. For both scoring variants, scores from signal peptides of orthologues were clearly distinct from alignments of random signal peptides (Wilcoxon signed rank test for paired samples, $P \ll 10^{-87}$). Furthermore, for more than 98% the proteins, scores from alignments of true orthologues were higher than the mean ± 2SD scores for arbitrary signal peptides. The results obtained after 100 random draws are shown for each species in Table S5. Score distributions are compared in Fig. S5.

By focussing our comparison of proteins on annotated clear orthologues, we introduced a selection bias towards more strongly conserved proteins, with an average similarity score of 0.7 ± 0.1 for the mature part. This does not affect our analysis, however, because signal peptides evolve much faster (Fig. S3), resulting in an unbiased and more widely spread similarity score of 0.5 ± 0.2 (Fig. S6).

Test for functional associations

Finally, the FatiGO [56] web service (http://babelomics3.bioinfo.cipf.es/, release 3.2) was employed to test for associations of proteins with leucine repeats in their signal peptides to particular Gene Ontology categories. Accordingly, all human secreted and type I membrane proteins were considered, comparing those with leucine repeats in their signal peptide against those without any leucine repeats.

Acknowledgements

The authors gratefully acknowledge helpful discussions with Arndt von Haeseler (University of Vienna). This work was supported by the Vienna Science and Technology Fund (WWTF), Baxter AG, Austrian Research Centres Seibersdorf and the Austrian Centre of Biopharmaceutical Technology.

References

1 Gusella JF & Macdonald ME (2006) Huntington's disease: seeing the pathogenic process through a genetic lens. *Trends Biochem Sci* **31**, 533–540.
2 Spires TL & Hannan AJ (2007) Molecular mechanisms mediating pathological plasticity in Huntington's disease and Alzheimer's disease. *J Neurochem* **100**, 874–882.
3 Siwach P & Ganesh S (2008) Tandem repeats in human disorders: mechanisms and evolution. *Front Biosci* **13**, 4467–4484.
4 Hands S, Sinadinos C & Wyttenbach A (2008) Polyglutamine gene function and dysfunction in the ageing brain. *Biochim Biophys Acta* **1779**, 507–521.
5 Brown LY & Brown SA (2004) Alanine tracts: the expanding story of human illness an cleotide repeats. *Trends Genet* **20**, 51–58.
6 Albrecht A & Mundlos S (2005) The other trinucleotide repeat: polyalanine expansion disorders. *Curr Opin Genet Dev* **15**, 285–293.
7 Delot E, King LM, Briggs MD, Wilcox WR & Cohn DH (1999) Trinucleotide expansion mutations in the cartilage oligomeric matrix protein (COMP) gene. *Hum Mol Genet* **8**, 123–128.
8 Karlin S, Brocchieri L, Bergman A, Mrazek J & Gentles AJ (2002) Amino acid runs in eukaryotic proteomes and disease associations. *Proc Natl Acad Sci USA* **99**, 333–338.
9 Alba MM & Guigo R (2004) Comparative analysis of amino acid repeats in rodents and humans. *Genome Res* **14**, 549–554.
10 Siwach P, Pophaly SD & Ganesh S (2006) Genomic and evolutionary insights into genes encoding proteins with single amino acid repeats. *Mol Biol Evol* **23**, 1357–1369.
11 Depledge DP & Dalby AR (2005) COPASAAR – a database for proteomic analysis of single amino acid repeats. *BMC Bioinformatics* **6**, 196.
12 von Heijne G (1983) Patterns of amino acids near signal-sequence cleavage sites. *Eur J Biochem* **133**, 17–21.

13 von Heijne G (1990) The signal peptide. *J Membr Biol* **115**, 195–201.
14 Hegde RS & Bernstein HD (2006) The surprising complexity of signal sequences. *Trends Biochem Sci* **31**, 563–571.
15 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.
16 Bendtsen JD, Nielsen H, von Heijne G & Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783–795.
17 Nielsen H & Krogh A (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc Int Conf Intell Syst Mol Biol* **6**, 122–130.
18 Huntley M & Golding GB (2000) Evolution of simple sequence in proteins. *J Mol Evol* **51**, 131–140.
19 Mularoni L, Veitia RA & Alba MM (2007) Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics* **89**, 316–325.
20 Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, de la Banda MG & Whisstock JC (2005) Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res* **15**, 537–551.
21 Russell RB & Gibson TJ (2008) A careful disorderliness in the proteome: sites for interaction and targets for future therapies. *FEBS Lett* **582**, 1271–1275.
22 Gerber H, Seipel K, Georgiev O, Hofferer M, Hug M, Rusconi S & Schaffner W (1994) Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263**, 808–811.
23 Brown L, Paraso M, Arkell R & Brown S (2005) *In vitro* analysis of partial loss-of-function ZIC2 mutations in holoprosencephaly: alanine tract expansion modulates DNA binding and transactivation. *Hum Mol Genet* **14**, 411–420.
24 Hancock JM & Simon M (2005) Simple sequence repeats in proteins and their significance for network evolution. *Gene* **345**, 113–118.
25 Fondon JW & Garner HR (2004) Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci USA* **101**, 18058–18063.
26 Caburet S, Cocquet J, Vaiman D & Veitia RA (2005) Coding repeats and evolutionary 'agility'. *Bioessays* **27**, 581–587.
27 Kashi Y & King DG (2006) Simple sequence repeats as advantageous mutators in evolution. *Trends Genet* **22**, 253–259.
28 Dorsman JC, Pepers B, Langenberg D, Kerkdijk H, Ijszenga M, den Dunnen JT, Roos RA & van Om-men GJ (2002) Strong aggregation and increased toxicity of polyleucine over polyglutamine stretches in mammalian cells. *Hum Mol Genet* **11**, 1487–1496.
29 Oma Y, Kino Y, Sasagawa N & Ishiura S (2004) Intracellular localization of homopolymeric amino acid-containing proteins expressed in mammalian cells. *J Biol Chem* **279**, 21217–21222.
30 Oma Y, Kino Y, Sasagawa N & Ishiura S (2005) Comparative analysis of the cytotoxicity of homopolymeric amino acids. *Biochim Biophys Acta* **1748**, 174–179.
31 Oma Y, Kino Y, Toriumi K, Sasagawa N & Ishiura S (2007) Interactions between homopolymeric amino acids (HPAAs). *Protein Sci* **16**, 2195–2204.
32 Siwach P, Sengupta S, Parihar R & Ganesh S (2009) Spatial positions of homopolymeric repeats in the human proteome and their effect on cellular toxicity. *Biochem Biophys Res Commun* **380**, 382–386.
33 Zhang L, Yu S, Cao Y, Wang J, Zuo K, Qin J & Tang K (2006) Distributional gradient of amino acid repeats in plant proteins. *Genome* **49**, 900–905.
34 Huntley MA & Clark AG (2007) Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. *Mol Biol Evol* **24**, 2598–2609.
35 Marcotte EM, Pellegrini M, Yeates TO & Eisenberg D (1999) A census of protein repeats. *J Mol Biol* **293**, 151–160.
36 Simon SM & Blobel G (1993) Mechanisms of translocation of proteins across membranes. *Subcell Biochem* **21**, 1–15.
37 Williams EJB, Pal C & Hurst LD (2000) The molecular evolution of signal peptides. *Gene* **253**, 313–322.
38 Li YD, Xie ZY, Du YL, Zhou Z, Mao XM, Lv LX & Li YQ (2009) The rapid evolution of signal peptides is mainly caused by relaxed selection on non-synonymous and synonymous sites. *Gene* **436**, 8–11.
39 von Heijne G, Liljeström P, Mikus P, Andersson H & Ny T (1991) The efficiency of the uncleaved secretion signal in the plasminogen activator inhibitor type 2 protein can be enhanced by point mutations that increase its hydrophobicity. *J Biol Chem* **266**, 15240–15243.
40 Riedl E, Koeppel H, Brinkkoetter P, Sternik P, Steinbeisser H, Sauerhoefer S, Janssen B, van der Woude FJ & Yard BA (2007) A CTG polymorphism in the CNDP1 gene determines the secretion of serum carnosinase in cos-7 transfected cells. *Diabetes* **56**, 2410–2413.
41 Gouridis G, Karamanou S, Gelis I, Kalodimos CG & Economou A (2009) Signal peptides are allosteric activators of the protein translocase. *Nature* **462**, 363–367.
42 Kreil DP & Ouzounis CA (2003) Comparison of sequence masking algorithms and the detection of biased protein sequence regions. *Bioinformatics* **19**, 1672–1681.
43 Wootton J & Federhen S (1993) Statistics of local complexity in amino-acid-sequences and sequence databases. *Comput Chem* **17**, 149–163.
44 Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, Sander C & Ouzounis CA (2000) CAST: an iterative algorithm for the complexity analy-

sis of sequence tracts. Complexity analysis of sequence tracts. *Bioinformatics* **16**, 915–922.
45 Yu YK, Wootton JC & Altschul SF (2003) The compositional adjustment of amino acid substitution matrices. *Proc Natl Acad Sci USA* **100**, 15688–15693.
46 Lyne R, Smith R, Rutherford K, Wakeling M, Varley A, Guillier F, Janssens H, Ji W, Mclaren P, North P *et al.* (2007) FlyMine: an integrated database for *Drosophila* and *Anopheles* genomics. *Genome Biol* **8**, R129.
47 The UniProt Consortium (2008) The universal protein resource (UniProt). *Nucleic Acids Res* **36**, D190–D195.
48 Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L *et al.* (2009) Ensembl 2009. *Nucleic Acids Res* **37**, D690–D697.
49 Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G & Kasprzyk A (2009) BioMart – biological queries made easy. *BMC Genomics* **10**, 22.
50 Karlin S (1995) Statistical significance of sequence patterns in proteins. *Curr Opin Struct Biol* **5**, 360–371.
51 Bagos PG, Tsirigos KD, Plessas SK, Liakopoulos TD & Hamodrakas SJ (2009) Prediction of signal peptides in archaea. *Protein Eng Des Sel* **22**, 27–35.
52 Tipping M (2000) The relevance vector machine. *Adv Neural Inf Process Syst* **12**, 652–658.
53 Tipping M (2001) Sparse Bayesian learning and the relevance vector machine. *J Mach Learn Res* **1**, 211–244.
54 Chica C, Labarga A, Gould CM, Lopez R & Gibson TJ (2008) A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC Bioinformatics* **9**, 229.
55 Needleman SB & Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443–453.
56 Al-Shahrour F, Minguez P, Tarraga J, Montaner D, Alloza E, Vaquerizas JM, Conde L, Blaschke C, Vera J & Dopazo J (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res* **34**, W472–W476.

Supporting information

The following supplementary material is available:
Fig. S1. Compositional profile of a eukaryotic signal peptide.
Fig. S2. SAAR enrichment by species.
Fig. S3. Plots of conservation scores for mature sequences versus sequences of signal peptides.
Fig. S4. Leucine repeat conservation analysis variants.
Fig. S5. Distribution of observed and random conservation scores.
Fig. S6. Conservation score distributions for signal peptides vs mature proteins.
Table S1. Abundance of the SAARs studied.
Table S2. Overview of eukaryotic organisms.
Table S3. Over-representation of Gene Ontology terms for L-repeats in signal peptides.
Table S4. Enrichment significance: empirical *P*-values.
Table S5. Comparison of L-repeat conservation scores for true orthologues and random orthologues.
Table S6. Leucine repeats in *Eukaryota*.
File S1. File containing the tab-separated raw results (raw_results.tsv) showing the observed and predicted counts of SAARs in mature and whole proteins.

This supplementary material can be found in the online version of this article.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.