Architecture, function and prediction of long signal peptides

Jan A. Hiss and Gisbert Schneider

Submitted: 5th April 2009; Received (in revised form): 18th May 2009

Abstract
Protein targeting in eukaryotic cells is vital for cell survival and development. N-terminal signal peptides guide proteins to the membrane of the endoplasmic reticulum (ER) and initiate translocation into the ER lumen. Here, we review the status of signal peptide architecture and prediction with an emphasis on exceptionally long signal peptides, which often escape the notion of the currently available prediction methods. We benchmark publicly available prediction methods for their ability to correctly identify exceptionally long signal peptides. A set of 136 annotated eukaryotic signals served as reference data. The best prediction tool detected only 63%. A potential reason for the poor performance is the domain architecture of long signal peptides, whose structural peculiarities are insufficiently considered by current prediction algorithms. To overcome this limitation, we motivate a general domain view of long signal peptides, which becomes detectable when both the overall length and secondary structure of long signal peptides are taken into consideration. This concept provides a structural framework for identifying and understanding multiple targeting and post-targeting functions.

Keywords: bioinformatics; machine learning; organelle; protein targeting; signal sequence; transit peptide

INTRODUCTION
In 1975, Blobel and Dobberstein demonstrated that N-terminal signal peptides guide proteins to the endoplasmic reticulum (ER) [1, 2]. With an average length of approximately 22 residues signal peptides share only low pair-wise amino acid identity and are therefore difficult to predict by sequence alignment methods [3–5]. Motivation for reliable prediction systems comes from the fact that signal peptides share regions of conserved features [6]: an often positively charged and in length strongly varying n-region, a less varying hydrophobic core region (h-region) and a more conserved signal peptidase cleavage site (c-region) [7]. Conserved residue patterns in the c-region render this portion particularly accessible to residue-based predictions [8]. In addition to the existence of a signal peptide, three components are essential for ER targeting of proteins: the cytosolic signal recognition particle (SRP) [9, 10], the ER membrane-bound SRP receptor [11] and the ER membrane-spanning hetero trimeric Sec61 complex consisting of Sec61, Sec62 and Sec63 [12]. Recent studies by Kalies and co-workers present evidence that a single Sec61 complex is sufficient for protein translocation in the ER [13].

For a subgroup of proteins, an additional ER membrane protein, TRAM (translocating chain-associated membrane protein), has been shown to be essential [14, 15]. Features of the signal sequence such as its length play a role in TRAM recruitment [16]. For the prion protein yet another additional protein TRAP (translocon-associated protein) is necessary for efficient translocation into the ER [17]. TRAP has also been shown to be associated with Sec61 and non-translating mammalian ribosomes [18]. There is strong evidence that both proteoliposomes [19] in eukaryotes as well as lipids in bacteria [20, 21] also play a role in signal recognition and protein translocation. Jungnickel and Rapoport...
demonstrated that under certain experimental conditions ER targeting can occur without involvement of the SRP [22]. Normally, SRP binds to the hydrophobic stretches \((h\text{-region})\) of signal peptides in newly translated proteins, arrests translation and guides the ribosome to the SRP receptor in the ER membrane [9, 10]. Sec61p forms a channel through which the nascent precursor protein can enter the ER lumen [23]. The signal peptide is then cleaved off by signal peptidase I (SP-I) [24], thereby initiating the co-translational protein translocation.

After signal peptide processing by SP-I the signal peptide is released laterally into the ER membrane where it is further processed by signal peptide peptidase (SPP) [25]. Deviating examples exist, where the signal peptide is not degraded but kept stable after SP cleavage, thus being able to perform post-targeting functions [26, 27]. In the case of preprolactin an intermediate form of the signal peptide is found in the cytosol [28]. For some of these cases, additional post-targeting function has been confirmed experimentally:

(i) UL40 protein from human cytomegalovirus possesses a 37-residue-long signal peptide, which contains an epitope for presentation by HLA-E encompassing positions 15–23. This mimotope sequence is identical to the natural epitope derived from the HLA-C signal peptide. Presentation of the HLA-C signal peptide epitope on HLA-E allows the immune system to monitor the level of expressed HLA-C. A reduced level of HLA-E on the cell surface leads to apoptosis induced by natural killer cells [29, 30]. In infected cells, the virus-encoded mimotope is loaded on HLA-E, thereby maintaining a stable level of HLA-E on the cell surface. This allows the virus to simultaneously downregulate HLA-C to avoid detection by cytotoxic T-cells, and prevent natural killer cell-induced apoptosis [29].

(ii) Lassa virus glycoprotein-C contains a long signal peptide (58 residues), which is required for generation of the mature functional protein. The signal peptide itself is stable after cleavage by SP-I and integrated into the virion. While its replacement by a short signal peptide still leads to correct ER targeting, functional glycoprotein-C is no longer obtained. Co-expression with the long wild-type signal restores maturation of functional glycoprotein-C. These studies by Eichler et al. clearly demonstrate that long signal peptides can possess additional functions that are unrelated to protein targeting [31, 32]. The identical observations were made for the long signal peptide (58 residues) of glycoprotein-C from lymphocytic choriomeningitis virus (LCMV) [33]. Both Lassa virus and LCMV belong to the arena virus family.

(iii) The long signal peptide (98 residues) of REM protein from mouse mammary tumor virus (MMTV) remains after processing by SP-I. It contains a nuclear localization signal and was shown to accumulate in the nucleoli [34].

(iv) Human interleukin-15 precursor exists in two isoforms. One isoform contains a short signal peptide encompassing 29 residues; the other isoform has a longer signal peptide (48 residues). Notably, these two signals are encoded on different exons (short signal: exons 4a+5; long signal: exons 3–5). It has been shown that the length of the signal peptide affects the efficiency of ER targeting, glycosylation and secretion [35].

(v) Murine C4b-binding protein, which is part of the complement system possess an experimentally validated 56-amino-acid long signal peptide. It has been argued that the long signal peptide may play a role in retarding or otherwise modulating the folding pathway of C4b [36].

Besides the aspect of a post-targeting function these signal peptides have an additional feature in common, namely their exceptional length. Long signal peptides have often been considered as 'exceptions from the rule' and treated as signal peptides with the sole function of ER targeting. From the reported post-targeting functions the question arises whether long signal peptides are rather isolated instances or represent a common scheme sharing additional features. We have investigated this subject by comprehensive sequence analysis and come to the conclusion that long signal peptides are abundant in eukaryotic proteomes and possess several structurally and functionally independent domains. While increased length alone does not provoke additional function \(per se\), additional functions may require increased length. We assume that the majority of long signal peptides escaped
previous notion as a consequence of insufficiencies of current automated signal peptide prediction software.

ARCHITECTURE AND FUNCTION OF LONG SIGNAL PEPTIDES

Analysis of the length distribution of eukaryotic signal sequences in UniProtKB [37] (release 14.2; 7964 entries with signal peptide annotation) revealed the anticipated mean of 23 ± 6 residues. Only proteins with evidence for their existence at protein level were taken into account (this means evidence for the protein to exist, but not for the signal peptide to be functional) (Figure 1A).

A different picture presents itself for the length distributions of mitochondrial targeting peptides (mTPs, Figure 1B) and chloroplast transit peptides (cTPs, Figure 1C). Proteins containing thylakoid signals are in this context regarded as chloroplast signals. UniProtKB contains 1101 non-putative proteins with annotated mTPs. The average mTP length is 36 ± 18 residues. In the case of cTPs, 514 entries were retrieved with an average length of 56 ± 19 residues. These values are in perfect accordance with earlier reports on the length distribution of cTPs [38]. Notably, the average cTP has 1.5 times the length of an mTP, which itself is approximately 50% longer than a standard export signal. This means that the potential coding capacity of the signals increase in the order SP < mTP < cTP, and more functions could be encoded in the organelle targeting sequences than in a standard ER-targeting signal.

Viral signal peptides display a length distribution resembling the one of eukaryotic signal peptides (compare Figure 1A and D, N = 213). The average length of viral signal peptides is 25 ± 16 residues. The longest example found is the signal peptide of the fusion glycoprotein-F0 of the canine distemper virus strain Onderstepoort with 135 residues [39]. It should be mentioned that it does not represent a classical signal sequence. Its cleavage occurs post-translational and before the mature protein is transported to the cell surface. This unusual signal sequence has a negative regulatory effect on F protein activity [39].

Due to the observed length distribution of known signal peptides, it presents a challenge to impartially define a general length cut-off above which a signal peptide may be termed ‘long’. We propose a signal

![Figure 1](https://academic.oup.com/bib/article-abstract/10/5/569/214121)

**Figure 1:** Length distribution of signal peptides annotated in UniProtKB (A–C: version 13.6, D: version 14.0). Putative proteins were excluded. (A) Eukaryotic signal peptides (N = 7539), (B) mitochondrial transit peptides (N = 1101), (C) chloroplast transit peptides (N = 514), (D) viral signal peptides (N = 213).
group-specific threshold twice the average length. Taking into account a reduction of six residues for the signal peptidase cleavage site, one obtains a threshold of 40 residues for ER signal peptides, 66 residues for mTPs and 112 residues for cTPs. According to these thresholds there are 136 (2%) ER, 81 (7%) mTPs, 7 (1%) cTPs and 11 (5%) vSP ‘long’ examples among the currently known, annotated signal sequences.

One may argue that these long signal sequences might simply represent a tolerated variance in length, and discrimination between long and short signal peptides solely based on length is artificial. It might therefore be reasonable to consider long single peptides as special instances of standard signal sequences within an extended sequence space that may code for additional functions. Do these signal peptides have other common aspects? As a first approach we analyzed the occurrence of predicted $\beta$-turns [40] as a secondary structure aspect reported to be important in long eukaryotic signal peptides [41]. We distinguished between short signal peptides (1–40 residues, \( N = 7828 \)), long signal peptides (>40 residues, \( N = 136 \)) and the middle portion (central 20 residues) of those long signal peptides. On average, short signal peptides were predicted to possess 0.9 ± 1.4 $\beta$-turns, long signal peptides 4.7 ± 3.2 and the middle portion of long signal peptides 1.9 ± 2.3. $\beta$-turns are predicted to occur more than twice as often in the middle portion of long signal peptides compared to short ones and more than five times as frequent in long signal peptides compared to short ones.

We then analyzed the distribution of $\beta$-turn positions in long signal peptides (Figure 2). Since signal peptides do not share apparent sequence homology [3–5], an alignment-based comparison is not feasible. To overcome this obstacle, we divided each sequence into three equal parts. Then, we checked whether these signal portions contain at least one $\beta$-turn in a certain third. A signal peptide can therefore produce a maximum of three ‘votes’, which was regarded as 100%. One signal peptide portion receives a vote if it contains at least one potential $\beta$-turn; additional $\beta$-turns in the same portion cause no additional vote. This allows for an alignment-independent comparison of positional preferences for $\beta$-turns in short and long signal peptides. While 7% of the short signal peptides possess at least one $\beta$-turn in the first portion, long signal peptides contain $\beta$-turns three times more often (23%) in their first portion. Also short eukaryotic signal peptides appear to prefer turns in portions 1 and 3, long signal sequences apparently favor turns in portions 1 and 2 (Figure 2). From this preliminary study, we conclude that there might be structural differences between long and short signal sequences:

(i) Short signal peptides contain the majority of the predicted $\beta$-turns in the last portion, whereas long signal peptides in the first portion of the signal peptide.

(ii) On average, long signal peptides possess five times more predicted $\beta$-turns than short signals.

In conjunction with the proposed length cutoffs and the predicted structural difference, additional functions, respectively, their existence would provide a meaningful criterion for the definition of long signal peptides. While it is undisputable that such additional function exist an automatic prediction and therefore a systematic count is not feasible due to various reasons:

(i) lack of appropriate software tools that are able to recognize unusually long signals,

(ii) potential complexity of the additional function (e.g. multiple protein–protein interactions) and

(iii) multiple possibilities for additional functions (e.g. gene regulation, protein–protein interaction, alternative targeting).

Figure 2: Sequence position of $\beta$-turns. Black: short signal peptides (\( N = 7828 \)); Gray: long signal peptides (\( N = 136 \)). Portions 1–3 indicate whether the predicted turn was located in the respective third of the signal sequence (equal length of each portion).
Instead of predicting the additional function itself we therefore propose to predict the potential for such additional function as a criterion. In a previous publication [41], we presented an algorithm (‘NtraC’ prediction system, publicly available from http://gecco.org.chemie.uni-frankfurt.de/NtraC/NtraC.html) for the detection of domains within long ER signal peptides. One of these domains contains a standard ER-targeting signal, while the other is free to perform distinct functions. Here, we used the NtraC algorithm to analyze the 7828 short and 136 long signal peptides for a potential additional function by predicting their domain architecture. Only 3% (239) of the 7828 short signal peptides, but 61% (83) of the 136 long signal peptides were predicted to be domain organized. This suggests that long signal peptides may actually form a subgroup of signal peptides, which is distinguishable not only by length and structure but also by the potential to possess an additional function.

The extraordinary length of these signal peptides needs to be discussed from two different perspectives:

First, are those elongated examples simply tolerated because of lacking selective pressure? If this is true, then modules of shorter ‘ancestral’ signals should exist. An argument in favor of this hypothesis is the conservation of targeting function to the secretory pathway independent from the length of a signal peptide. It is known that small domains containing a secretion signal are present in many long signal sequences [41].

Second, do these signal peptides harbor additional co- or post-targeting functions? In other words: Is there a selective pressure to preserve long signal peptides for maintaining the integrity of essential additional functions? An argument for this perspective is the conservation of long signal peptides in ortholog proteins. For example, human tumor necrosis factor receptor superfamily, member 10d precursor (UniProtKB ID: Q9UBN6) contains a long signal peptide, which is conserved in the predicted orthologs in chimpanzee (96%; XP_528087) and rhesus monkey (80%; XP_001107922). A further aspect is the recruitment of co-factors such as TRAM, which has been shown to depend on the length of the signal peptide [16].

It is important to realize that long signal peptides must be distinguished from tandem signals (also called ‘twin-targeting signals’) which lead to a dual targeting, e.g. toward mitochondria and chloroplasts [42–44] (Figure 3A). While tandem signals exhibit two sequentially processed and distinctively recognizable signals, long signal peptides are regarded as one single unit. Tandem signals represent a sequential combination of individual, sequentially decoded signals [42], which are separated by a signal peptidase cleavage site. Long signal peptides on the other hand contain only one signal peptidase cleavage site and should therefore be regarded as a unit with one primary targeting function. Long signal peptides must also be distinguished from ‘ambiguous signals’ [42], which promote dual targeting by a mixture of two signals (Figure 3B). There are several known examples for dual targeting to ER and mitochondria in vertebrata, e.g. γ-secretase subunit nicastrin [45] and cytochromeP450 2B1 [46]. Both contain an ER-targeting signal with a signal peptidase cleavage site that is directly followed by a mitochondria-targeting signal lacking a cleavage site (Figure 3C). In contrast, long signal peptides combine two (or more) signals N-terminal of the single signal peptidase cleavage site (Figure 3D).

### PREDICTION OF LONG SIGNAL PEPTIDES

Exceptionally long functional signal peptides undoubtedly exist, but are hard to detect. One reason for the limited number of annotated long signal peptides in sequence databases might be the difficulty to predict them in silico. Long signal peptides are often not recognized by existing software tools or regarded as artifacts and consequently ignored. For example, SignalP penalizes signal peptide cleavage sites that are more distant from the N-terminus than some average value [5]. This sometimes leads to the preference of degenerate signal peptidase cleavage sites closer to the N-terminus. As a consequence, truncated signal peptides are predicted. An example is the long signal peptide (58 residues) of Lasa virus glycoprotein-C which is predicted to be cleaved at position 34 but was experimentally shown to be cleaved at position 58 [32]. It is important to keep in mind that penalizing cleavage sites far downstream from the N-terminus is well motivated, as the overwhelming majority of signal peptides follows the standard length distribution (Figure 1A). Therefore, all training data-based prediction tools will tend to regard a long signal peptide as an outlier. For the development of SignalP 3.0, sequences that are shorter than 15 and...
longer than 45 residues were removed from the training data [5], which led to improved prediction accuracy for standard signal peptides but hampers the detection of long signals.

The prediction tool Phobius combines transmembrane segment prediction and signal peptide prediction [47]. Its strength lies in the discrimination of signal peptides and transmembrane segments, which can be easily mistaken for one another due to their hydrophobic nature. The hidden Markov model implemented in Phobius uses a fixed-size window for the \( h \)-region (6–20 residues) and the \( c \)-region (4–18 residues) but allows an arbitrary length of the \( n \)-region. While this should enable the system to detect \( n \)-regions that contain more than 10 residues, such findings are exponentially penalized per additional residue. For example, Phobius detects the long signal peptide (43 residues) of shrew-1 [48] but fails for the Lassa virus glycoprotein-C [32] (signal peptide: 58 residues).

To approach this problem systematically we probed the predictive performance of six software tools that are publicly available on the World Wide Web (SignalP [5], Phobius [47], Philius [49], PrediSi [50], Signal-CF [51], Signal-3L [52]), taking the sequences of 136 long signal peptides found in this study as test cases (Table 1). We considered only methods that are able to predict the signal peptidase cleavage site position, because we focused on the detection of a signal peptidase cleavage site and the resulting correct prediction of the length of the signal sequence. For example, WoLF PSORT [53], which only states whether a sequence might be secreted but gives no signal peptide length, was excluded from our study. The program TargetP [38] was not taken into account since its prediction for the signal peptides cleavage site is based on SignalP. The tool SPOCTOPUS [54] was only used for a representative set of nine examples (Table 2) but not for the systematic approach (Table 1). SPOCTOPUS had been developed for topology prediction and requires a multiple sequence alignment for multiple sequence input, which is meaningless for isolated signal sequences. Notably, no program was capable of correctly recognizing all nine representative signals. For this group of selected sequence, examples PrediSi [50] and Signal-CF [51] performed best, detecting seven out of nine examples. In the systematic approach SignalP [5] and Signal-CF performed best detecting 61% and 63%.
respectively of the 136 long signal peptides with their
correct length (Table 1).

As a consequence from this benchmark, one
should be cautious when trying to predict long
signal peptides using software tools with the
default parameter settings. For manual detection of long
signal peptides we recommend to use not a single
program but a combination, for example, Signal-CF
[51] and SignalP 3.0 [5]. While Signal-CF performed
best in our preliminary test, SignalP presents the
S-score (signal peptide probability) and C-score
(signal peptidase cleavage site probability) distributions.
These can, properly considered, give hints
about the existence of a long signal peptide.

Long signal peptides also can have recessed
h-regions. Due to their often-weak signal, prediction
tools tend to overlook them (e.g. in ADA25_MOUSE),
consider them as transmembrane sequences (e.g. ATS4_HUMAN, CGB2_HUMAN) or
a signal anchor (e.g. IGF2B_XENLA, NCKX2_CHICK, DCBD2_HUMAN, GUC2D_BOVIN, Table 2).

**WHY DO SOME PROTEINS CONTAIN LONG SIGNAL PEPTIDES?**

One may ask the question whether long signal peptides
represent a variation of short signal peptides or
represent a functional subgroup. If the signal peptides
share a common origin, where the long signal peptides
might represent a modification of the standard
signal concept, the distribution of amino acids should
differ between signals.

We therefore tested the null hypothesis assuming
identical distributions, which could hint toward a
common sequence origin. Since we cannot assume
Gaussian distributions, we used the non-parametric
two-sided Kolmogorov-Smirnov test [55]. For an
alpha-level of 5% remaining error probability, the
null hypothesis was not rejected for the amino-acid
composition of short and long SP, mTP, cTP
and vSP. The null hypothesis was accepted for
any combination with \( P \) between 28% and 97%. In
other words, based on the overall amino-acid com-
position, no statistically significant difference is
observed between these sets of sequences. We con-
clude that the amino-acid composition alone does
not represent the decisive feature for the distinction
between cellular compartments. Still, a prominent
role might be attributed to the distribution of posi-
tively charged arginine residues. They are least abun-
dant in SP (3 ± 4%) and most abundant in mTP
(13 ± 5%). Their relative frequency increases with
the signal length and the complexity of targeting
signal organization (Table 3). Viral signal peptides
exhibit the same arginine content as standard short
signal peptides.

It has been shown that the interchange of signal
peptides between eukaryotic and prokaryotic cells is
sometimes possible [56]. It would be most interesting
to see whether the exchange of long signal peptides
follows the same or other rules. If they exhibit post-
targeting function, does this interfere with their

<table>
<thead>
<tr>
<th>SRS accession</th>
<th>SignalP 3.0</th>
<th>Phobius</th>
<th>Philius</th>
<th>PrediSi</th>
<th>Signal-CF</th>
<th>Signal-3L</th>
<th>SPOCTOPUS</th>
<th>Annotated SP length</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADA25_MOUSE</td>
<td>50 (no)</td>
<td>(no)</td>
<td>(no)</td>
<td>50 (no)</td>
<td>50</td>
<td>50</td>
<td>47</td>
<td>50</td>
</tr>
<tr>
<td>AJAP1_HUMAN</td>
<td>43 (SP)</td>
<td>(no)</td>
<td>(no)</td>
<td>43 (yes)</td>
<td>43</td>
<td>43</td>
<td>(no)</td>
<td>43 (*)</td>
</tr>
<tr>
<td>ATS4_HUMAN</td>
<td>52 (SP)</td>
<td>24–46 (TMS)</td>
<td>52 (SP)</td>
<td>51 (yes)</td>
<td>51</td>
<td>43</td>
<td>(no)</td>
<td>51</td>
</tr>
<tr>
<td>CGB2_HUMAN</td>
<td>50 (SP)</td>
<td>33–50 (TMS)</td>
<td>31–51 (TMS)</td>
<td>50 (yes)</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>IGF2B_XENLA</td>
<td>56 (SA)</td>
<td>35–52 (TMS)</td>
<td>(no)</td>
<td>56 (yes)</td>
<td>56</td>
<td>56</td>
<td>56</td>
<td>56</td>
</tr>
<tr>
<td>NCKX2_CHICK</td>
<td>55 (SA)</td>
<td>36–60 (TMS)</td>
<td>36–59 (TMS)</td>
<td>57 (yes)</td>
<td>48</td>
<td>(no)</td>
<td>(no)</td>
<td>58</td>
</tr>
<tr>
<td>DCBD2_HUMAN</td>
<td>57 (SP/SA)</td>
<td>1–35 (SP); 45–61 TMS</td>
<td>67 (SP)</td>
<td>66 (yes)</td>
<td>66</td>
<td>(no)</td>
<td>(no)</td>
<td>66</td>
</tr>
<tr>
<td>GUC2D_BOVIN</td>
<td>57 (SP/SA)</td>
<td>44–68 (TMS)</td>
<td>44–68 (TMS)</td>
<td>56 (yes)</td>
<td>56</td>
<td>56</td>
<td>56</td>
<td>56</td>
</tr>
<tr>
<td>Q6Y6279_VIRU</td>
<td>34 (no)</td>
<td>20–53 (TMS)</td>
<td>29–33 (TMS)</td>
<td>58 (yes)</td>
<td>(no)</td>
<td>30</td>
<td>(no)</td>
<td>58</td>
</tr>
</tbody>
</table>


Table 2: Prediction of long signal peptides
primary targeting function? Can long signal peptides be exchanged? First hints are given by the observation that long signal peptides exhibit their proper targeting and post-targeting function in cell-free translation/translocation assays [34]. This argues for functions that are locally encoded in the signal peptide since these additional functions can be found and assessed in chimeric proteins. On the other hand, it has been shown that the combination of signal peptide and transmembrane sequence influences the membrane topology of the native prion protein [57]. The prion protein is found in three forms, with (i) the N-terminus intracellular or (ii) extracellular and (iii) as secreted protein. The signal sequence directly influences the ratio between these three forms [57, 58].

The location of the predicted turn-forming sequence elements deserves particular attention. Turn structures, in particular as part of beta-hairpins, are well known to mediate protein–protein interaction [59, 60]. A prominent example involving an essential β-turn is the virus receptor interaction between HIV type I and chemokine receptors [61]. Thus, one might consider the central ‘transition area’ of long signal peptides as important for their overall structural topology. An example of signal sequence-dependent topology is cytochrome b5, found primarily in the ER membrane where it is inserted by 20 uncharged amino acids at the C-terminus and the N-terminus facing the cytoplasm. By creating a chimeric protein, which additionally contains a prokaryotic signal sequence, Kaderbhai and co-workers [62] demonstrated an inversion of the topology. Analog observations were made for the maltose-binding protein and ribose-binding protein from *Escherichia coli* where the leader peptide retards the folding process [63, 64].

Taken together, these findings support an essential interaction between the long signal peptide and its native protein. The β-turns required to identify additional functions coded in long signal peptides may be required for the additional function itself, such as protein–protein interaction. Such an interaction requires thorough investigation by systematic exchange of long signal sequences with other long and short signals. Further research will have to scrutinize this hypothesis and clarify why not all signal sequences share this feature. Irrespective of the outcome of these studies, it will be most worthwhile to carefully re-assess sequence database annotations using appropriate prediction software and explicitly look for long signal peptides. They may be more abundant than currently known and systematic functional investigations will undoubtedly provide new insights in the role of long signals in cell–cell recognition and membrane protein interaction.

### Key Points
- Current prediction software for the detection of signal peptides inaccurately recognizes exceptionally long signals.
- Long signal peptides can have multiple biological functions.
- We suggest a structure-based domain organization of long signal peptides for improved prediction and function annotation.

### Acknowledgement
We are most grateful to Bernhard Dobberstein and Katja Kapp for fruitful discussion.

### FUNDING
The Beilstein-Institut zur Förderung der Chemischen Wissenschaften and the Centre for Membrane Proteomics at Goethe-University Frankfurt am Main.

| Amino acids | A | C | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W |
| SP < 40     | 13 ± 3 | 4 | 1 | 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 |
| SP ≥ 40     | 12 ± 3 | 3 | 1 | 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 |
| SP > 40     | 15 ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 |
| mTP         | 12 ± 2 | ± 3 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 |
| cTP         | 11 ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 |
| vSP         | 9 ± 2  | ± 3 | ± 4 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 | ± 2 |

**Table 3:** Frequency of amino acids in signal peptides (in percent ± standard deviation)

SP < 40: Eukaryotic signal peptides with <40 amino acids (N = 7635).
SP ≥ 40: Eukaryotic signal peptides with ≥40 amino acids (N = 145).
SP > 40: Eukaryotic signal peptides with >40 amino acids and N-domains predicted to act as mTP (N = 32).
SP: Chloroplast transit peptide (N = 63). cTP: Chloroplast transit peptide (N = 64). vSP: Viral signal peptides (N = 213). All sequences were taken from UniProtKB (Version 14.0). Proteins annotated as putative were excluded from the analysis.
References


46. Hiss and Schneider