Published in final edited form as:

J Theor Biol. 2016 November 21; 409: 11–17. doi:10.1016/j.jtbi.2016.08.036.

Skeletal Muscle Signal Peptide Optimization for Enhancing Propeptide or Cytokine Secretion

Manoel Figueiredo Neto and Marxa L. Figueiredo*

Department of Basic Medical Sciences, College of Veterinary Medicine, Purdue University, 625 Harrison St, West Lafayette, IN 47904

Abstract

We have utilized hidden Markov models using HMMER software to predict and generate putative strong secretory signal peptide sequences for directing efficient secretion of cytokines from skeletal muscle for therapeutic applications. The results show that this approach can analyze signal sequences of a skeletal muscle secretome dataset and classify them, emitting new sequences that are strong candidate skeletal muscle-enriched signal peptides. The emitted signal peptides also were analyzed for their hydropathy and secondary structure profiles as compared to native signal peptides. The emitted signal peptides had a higher degree of hydropathy and helical composition relative to native sequences, which may suggest that these new sequences may hold promise for promoting enhanced secretion of proteins including cytokines or propeptides from skeletal muscle.

Keywords

signal peptide prediction; cytokine secretion; Hidden markov model; neural networks

INTRODUCTION

In eukaryotes, the secretion of proteins to the outside of the cell is a complex process that begins with the insertion of proteins from the ribosome into the endoplasmic reticulum (ER). This targeting of the protein to the secretory pathway occurs through recognition of the N-terminal signal peptide (SP) by the signal recognition particle (SRP). SP are comprised of three domains, the first of which is a short chain of positively charged amino acid residues, the N-terminal domain. These residues are usually arginines or lysines, which give this region a hydrophilic net positive charge. The SP core is known as the H region since it comprises a hydrophobic stretch of non-polar alanine and leucine residues, which form an alpha-helical conformation. The C-terminus is the last domain of the SP and is called the C-region, typically beginning with either proline or glycine; the last three amino acids are commonly referred to as a signal-peptidase processing site. As an N-terminus protein region

*Corresponding author.: Department of Basic Medical Sciences, College of Veterinary Medicine, Purdue University, 625 Harrison St, LYNN 2177, West Lafayette, IN 47904. Ph: 765-494-5790. Fax: 765-494-0781. mlfiguei@purdue.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

leaves the ribosome, SRP halts further translation and interacts with this short sequence of amino acids, consequently binding to a protein complex known as the SRP receptor, which is located on the surface of the ER membrane. Signal peptidases recognize the end of the SP and specifically cleave a stretch of amino acids, followed by co-translational proteolytic processing. The result is a free signal peptide and a mature protein.

Skeletal muscle is a metabolically active tissue that secretes a wide range of proteins that affect muscle physiology locally and also can have a significant impact on the whole organism. Recently, through an integrated genomic and proteomic approach, Hartwig *et al.* have defined the secretome of human skeletal muscle cells using mass spectrometry and multiplex immunoassay methods [1]. The proteins detected were validated using quantitative transcriptomics. In the present study, we have built on the knowledge generated from the referenced secretome study to generate novel optimized SP sequences with the goal of promoting more efficient secretion of proteins from skeletal muscle. The application of these strong muscle-enriched SP will be to enhance expression of cytokines or other secreted therapeutics (growth factors, propeptides) from skeletal muscle following gene delivery [2]. Cytokines can be expressed in skeletal muscle cells following gene delivery to treat a wide variety of diseases; yet one critical aspect of therapeutic efficacy is that is difficult to achieve therapeutic levels of cytokines in serum. For example, it has been reported that for Interleukin-27 (IL-27), levels of ~100-120 pg/ml are inefficient in tumor therapy, while levels of ~500-600 pg/ml can dramatically induce tumor regression in colon cancer xenograft models (CT29) without inducing toxicity [3]. Although promising, it is difficult to predict the levels of cytokine expression that might be sufficient for eliminating tumors and this may vary widely depending on tumor type. For example, levels that can easily eliminate CT29 tumors are insufficient to treat aggressive 4T1 breast tumors [3]. Therefore, we anticipate that certain tumors or diseases may necessitate a higher-level of cytokine expression and/or secretion into serum for reaching full therapeutic efficacy. Using mathematical models [4], the serum levels needed for IL-27 efficacy could be significantly higher for more aggressive tumors, such as >1300 pg/ml. Therefore, it is critical to develop strategies to enhance secretion of cytokines from muscle so therapeutic levels can be achieved in serum for eliminating aggressive or difficult to treat tumors.

Designing stronger and muscle-enriched SP is possible with current methods that include machine-learning approaches such as neural networks and Hidden Markov Models (HMM) [5]. The use of HMM has grown recently for profiling protein domains based on a training dataset in order to model and engineer optimized biological sequences. In this study, we utilized HMM analyses based on a secretome dataset to generate optimized SP, which can be muscle-enriched and would possess optimal features for secretory processing by muscle cells. The key concept in the HMM approach is to use a finite model representing probability distributions over infinite numbers of possible sequences. Therefore, an HMM model constructed for given protein sequences consists of a set of states, each with a characteristic amino acid distribution. The states are then connected by transition probabilities, which specify several possible state orders. A model is constructed on protein sequences of variable length, then subsequently used to predict the most probable way for generating a strong consensus sequence. Any sequence insertions and deletions are implemented into the model in the form of regular transition states. Profile HMM is the most

commonly used form in computational biology [6, 7], generating and emitting final structures based on a prior built HMM profile [8]. HMM has been utilized successfully to predict and model SP by several groups [9-12]. Using a SP dataset, the model can be trained to score an input sequence to predict and generate a consensus signal sequence, then emit artificial sequences at given bit scores. The goal is to generate stronger, muscle-optimized SP, which will likely enhance cytokine secretion. Since experimental approaches to characterize amino acid changes and optimization in SP are time consuming and expensive, computationally predicting the effects of such changes in order to promote SP optimization could be an important approach to assess the impact of novel changes and prioritize them for experimental validation.

Materials And Methods

Defining a dataset enriched for strong skeletal muscle signal peptides

We used a dataset described for the human skeletal muscle cell (hSkMC) secretome, in which 548 proteins were detected in conditioned media of hSkMC derived from 22 adult donors (balanced between male and female) [1], and used to construct a model for the secretome. Quantitative transcriptomics then were used to create a “virtual secretome” with the intent of validating the proteomics data. The virtual secretome has significant overlap with the secretome in that significant mRNA expression was demonstrated for 501 (91.4%) out of the 548 proteins detected by proteomics analyses. The 501 validated gene products were identified as strong candidates to be secreted, and subjected to consecutive filtering using SignalP, SecretomeP and ER_retention signal databases. This analysis resulted in a list of 169 gene products that encoded proteins with a predicted or validated signal peptide (SignalP positive or SP+), while 136 gene products followed non-classical SP-independent mechanisms (SP- or SecretomeP score >0.5), and 196 gene products did not comply with the computational filtering applied. From these 169 proteins, we applied a selective cutoff to classify expressed transcripts as candidates for protein secretion if expression levels were equal to or exceeded 2,500 (2,500-8,999), based on the expression levels of housekeeping gene B2M (2,614), yielding 67 samples. From the 67 samples, a more stringent cutoff was applied relative to the expression of housekeeping gene ATP5B for classifying highly expressing transcripts as strong candidates for protein secretion. These levels were set to yield samples with the highest expression levels (i.e., 8,538), and 14 samples met this cutoff. We then used a signal peptide database (<http://www.signalpeptide.de/>) to obtain the SP sequences for these 14 proteins validated by the quantitative transcriptomic data for use in the following model generation and analyses.

A HMM model for emitting novel signal peptides from the skeletal muscle dataset

In order to build the model based on HMM, we used analyses based on Razmara *et al* [13] for using HMMER for analyzing biological sequences using profile hidden Markov models [6, 7]. We used the set of programs accessible at Mobyle1.5, a bioinformatics framework platform [14]. We built the model using the alignment obtained with *hmmalign* to train the model and build a profile (*hmmbuild*) and *hmmemit* to emit sequences utilizing the model. Therefore, we used HMMER to model and analyze signal sequences from the dataset and then to classify them based on a similarity score. From *hmmemit*, the emitted consensus

sequences were analyzed further to confirm the presence of a SP, as well as the predicted strength by using SignalP 4.1 software. Signal P4.1 [10] predicts traditional N-terminal SP in both prokaryotic and eukaryotic proteins. The emitting process yielded consensus sequences, of which the top consensus sequences are shown in Table 1. These emitted SP were validated as signal peptides and were scored by Signal P4.1 and TargetP1.1. The multi alignment file from HMMER was used to search against a sequence database in Signal-BLAST (Uniprot-sprot) for determining extent of identity with native SPs. Signal-BLAST also was utilized as an independent means to validate the consensus sequences emitted for the presence of a detectable SP. SP were identified by Signal-BLAST and a bit score was obtained which reflected the extent of similarity of the consensus-emitted SP to native SP in the database.

Predicting signal peptide secondary structure

The software I-TASSER [15] was used to structurally model the top two HMM emitted SP. I-TASSER approaches protein function prediction hierarchically, yielding structures with similarity to signal peptide topology including alpha helical composition. Using ProtScale [16], we also calculated hydropathy index profiles for the top two emitted SP as well as for the two native SP with the highest degree of homology to the emitted sequences as determined by Signal-BLAST.

RESULTS AND DISCUSSION

Training a computational model on a skeletal muscle-enriched dataset in order to develop strong signal peptide sequences

In order to generate a training dataset, we used data that correlated abundance of myokines or proteins secreted from human skeletal muscle cells (hSkMC) with signal peptide strength profiles obtained with SignalP4.1, as described in materials and methods. Briefly, we first determined the correlation between predicted signal strength to levels of gene expression from hSkMC array data to define a virtual secretome, a term first coined by LeBihan *et al* [17]. Fourteen candidate genes were identified in our pre-screen (Figure 1) and the majority of these virtual secretome candidates including COL1A2 and MMP2 were identified also in the conditioned media in a study that utilized neonatal primary human muscle cells [17]. Our 14 top secreted candidates were subsequently utilized to build a consensus HMM model using the software module *hmmbuild*. The output multiple alignment files were used to create an HMM model based on maximum likelihood parameter estimation. An example of a multiple alignment is shown in Figure 2. Next, we proceeded to use *hmmemit* to emit novel skeletal muscle-enriched SP, which are candidates for promoting strong expression and secretion of proteins such as cytokines or peptides from skeletal muscle. The emitting process yielded consensus sequences, of which the top six are shown in Table 1. These emitted SP were validated as signal peptides and were scored by Signal P4.1 and TargetP1.1.

The rationale for SP optimization using a skeletal muscle-trained dataset is that this strategy can produce consensus sequences of greater strength compared to native SP. This is important when considering protein production from skeletal muscle, since simply substituting a SP for a cytokine's native sequence may not necessarily enhance its muscle

secretion. There are several contextual elements that may augment or reduce SP strength when it is engineered onto a cytokine, thus a consensus sequence might be more effective in promoting higher cytokine levels. Also, typically, cytokines are expressed far below the abundant secreted protein threshold and have SP that can benefit from further optimization (Figure 1, yellow diamonds). We also considered addition of highly expressed genes Fibronectin 1 (FN1) and beta-actin (ACTB) (Figure 1) to our analyses, especially since FN1 also was detected in the LeBihan *et al*/dataset. However, our rationale was that FN1 was too large (~260 kD) compared to typical cytokines (~20-100 kD) to use for our application of promoting optimal cytokine secretion. The other highly expressed gene, ACTB, does not contain a SP, and we did not include SP- proteins in the analyses. At any rate, despite limitations to using a virtual secretome analysis that is confirmed by direct identification of proteins via MS techniques, we still envision a significant benefit for strategies that aim to develop or predict a strong skeletal muscle-enriched consensus SP sequence for augmenting secretion of a cytokine. The engineered cytokines might be delivered following viral or non-viral gene delivery methods and the applications could be manifold, including enhanced cytokine expression for treating cancer or genetically related pathologies, for example.

Although we used a dataset that had qualitatively identified 538 proteins by MS and was validated by quantitative transcriptomics for 501 of the gene products to create a virtual secretome, our analyses have the limitation that the proteomics dataset was not quantitative. Other groups have reported the challenge in quantifying proteins in the secretome of primary human skeletal muscle cells [17]. To the best of our knowledge, there are two studies that examine primary human skeletal muscle cells without studying additional confounding variables such as exercise or obesity effects [1, 17], and there is great concordance between these two studies. Interestingly, LeBihan *et al* also detected many of our virtual secretome hits in conditioned media of neonatal muscle cells [17]. Additional support for using the virtual secretome approach for analysis is based on two other studies by Yi *et al* [18] and Drexler *et al* [19], which examined transcripts and proteins from whole muscle fibers. Drexler *et al* reported a good Pearson's correlation (0.81) between mRNA and protein levels, while Yi *et al* showed that although detected proteins were underrepresented, the majority (88%) of the protein groups detected had corresponding transcripts detected by microarray. Finally, a mouse muscle secretome study may support the combination virtual secretome/proteomics models in ours and LeBihan reports [20], since they also captured the main groups of the gene products from our analyses (ECM, proteases including MMPs, fibrillar matrix, and basement membrane).

Confirmation of SP by Signal-BLAST analyses and Signal P 4.1 with TargetP1.1 validation

The emitted SP were examined using TargetP1.1 in order to validate the sequences for their likelihood of targeting proteins to the secretory pathway [21, 22]. We found that TargetP1.1 was able to validate SignalP4.1-defined SP with closely related scores (Table 1). Signal-BLAST was utilized next as an independent means to validate the consensus sequences emitted for the presence of a detectable SP as well as for searching against a sequence database for determining extent of identity with native SPs. SP were identified and a bit score was obtained from Signal-BLAST which reflected the extent of similarity of the consensus-emitted SP relative to native SPs in the database [23] (Table 1). The emitted

sequences SP1 and SP2 had the highest degree of homology with COL1A2 (70% identity, bit score 30.0) and Mmp2 (69% identity, bit score 40.4) sequences, respectively. SP1 and SP2 therefore were selected as the top two emitted peptides for proceeding onto further structural analyses. Prior to structural analyses, we also examined the sequences for putative N-glycosylation sites using NetNGlyc1.0, as previously described [24]. Our analyses showed an absence of glycosylation sites in both the emitted SP and native SP, which would be important to avoid interference with Signal Peptidase I processing events.

Structural comparison between native and top two emitted SP.

Using information from Signal-BLAST that identified the closest native SP homologs for SP1 and for SP2 (Col1a2 and Mmp2, respectively), we compared the emitted and native SP using Protscale (Figure 3). The analyses indicated that the emitted signal peptide SP1 had a greater hydrophobicity extent between aa 5-13 compared to the native SP examined, whereas the emitted SP2 had a greater hydrophobicity between aa 8-17 and 21-25 compared to the native SP examined (Figure 3). As the typical SP structure has inherent flexibility in its length but requires a hydrophobic H-region that is preferably alpha-helical, increases in the hydropathy index can be viewed as positive in that the total hydrophobicity may be a more important factor for SP recognition by the SRP than the length of the H-region [25]. Bird *et al*/has suggested that the efficiency of protein translocation is related to the degree of hydrophobicity of the SP [26], while Zhang *et al*/have observed up to a 3.5-fold enhancement in the secretion of interleukins when the hydrophobicity of the H-region or the basicity of the N-region of the SP are increased [27].

Next, sequences were assessed for secondary structure using I-TASSER. The goal was to obtain a model of the emitted peptides and their potential alpha-helical composition for comparison with native SP. First, the general structure of eukaryotic SP is depicted in Figure 4a. There is a critical importance in the hydrophobic H-region and in that it forms an alpha-helical conformation. Using Signal4.1 analyses of the SP strength, as well as secondary structure prediction by I-TASSER, we observed an increase in several scores for both SP1 and SP2 relative to native SP (Figure 4b, plots). The output from SignalP4.1 indicated that the emitted SPs (SP1 or SP2) were stronger than the native SP used in the analysis. Fig. 4b shows that SP1 exceeded the native COL1A2 SP by 68% in the C- score, by 29% in the Y-score and by 10% in the D-score. The C-score is the raw cleavage site score, reflecting the output from the CS networks, which are trained to distinguish signal peptide cleavage sites from everything else. The Y-score is the combined cleavage site score, which is an average of the C-score and the slope of the S-score, which may result in a better cleavage site prediction than the raw C-score alone. The SP2 also exceeded the native Mmp2 SP by 6.3% in the maximum S-score, 14% in the mean S-score, and 8% in the mean D-score (Fig. 4b). The S-score is the signal peptide score, which distinguishes positions within SP from positions in the mature part of the proteins and from proteins without SP. The mean S is the average S-score of the possible signal peptide (from position 1 to the position immediately before the maximal Y-score). Finally, the D-score is a discrimination score and is a weighted average of the mean S and the max. Y scores. This is the score that is used to discriminate signal peptides from non-signal peptides.

Besides having stronger signal peptide scores relative to the native SP, the emitted SP also appeared to have a higher degree of helical composition as determined by the structural analyses and modeling. Depicted in Figure 4b are the structural models generated, which suggest a higher degree of helical composition of the emitted (SP1 and SP2) relative to native SP. Therefore, our combined analyses lead us to postulate that the consensus emitted novel SP1 and SP2 are examples of putative strong peptides that can be used for enhancing secretion of heterologous proteins from skeletal muscle relative to the native SP examined. The targeting and insertion of the SP into the ER membrane is governed in part by the information in the hydrophobic regions of the SP and influences timing and efficiency of maturational events, especially signal peptide cleavage and glycosylation [28]. For a SP to be recognized by Signal Recognition Particle (SRP), total hydrophobicity is an important factor. While some SP seem to be general in their ability to promote protein secretion of a variety of proteins, others are more protein specific. In this context, it is very interesting that the amino acid sequence downstream of a signal peptide can affect its efficiency [29], and also that particular SP have evolved to achieve optimum translocation of particular passenger proteins. Importantly, although the sequences were used here for modeling folding, the novel SP would be utilized as to replace a cytokine's native SP. For example, an IL-27 cytokine would have its native SP replaced with a putative skeletal muscle-enriched SP, and ideally in frame with the protein's normal site of cleavage. This means that only the first ~22-25 aa of the novel SP would be used to enhance secretion of IL-27 and the mature cytokine itself would not be modified following cleavage by signal peptidase I.

And although these emitted SP would have to be experimentally tested in subsequent studies, previous work has shown that simple changes in scores between wild type and mutated sequences can already yield enhanced functional effects for certain SP [30, 31]. An HMM consensus or emitted SP based on a profile from a tissue secretory pattern, such as described here for skeletal muscle, could be useful in reducing heterologous protein secretion from a certain cell type since the context of a successful SP can be maintained regardless of its fusion with a heterologous protein. Since SP have a remarkable diversity in regards to length, hydrophobicity and net charge, a machine-learning design that can incorporate length variability, hydrophobicity, and size and peak of the hydropathy index into consideration when targeting expression of a heterologous protein to skeletal muscle may be of considerable interest for therapeutic applications.

CONCLUSIONS

Computational methods provide the opportunity of rapid prediction of potential optimized secretory SPs and their potential cleavage sites. In this study, we have developed a signal peptide model based on HMM for modeling and optimizing protein secretion by human skeletal muscle cells. A model was built on SP sequences of a set of proteins highly expressed in skeletal muscle and also predicted to be secreted. We used the model to emit the top artificial strong SPs. In conclusion, the study suggests a novel idea to improve the strength of SP for enhancing secretion of cytokines, in particular in skeletal muscle. Replacement of a native cytokine SP (e.g.IL-27) with a stronger SP designed computationally is hypothesized to promote increased secretion of the cytokine from skeletal muscle and enhance its therapeutic effect in a systemic manner.

Acknowledgements

We would like to acknowledge support from NIH grants R21CA179699 and R01CA196947-01A1 (MLF) and support from the College of Veterinary Medicine and Basic Medical Sciences at Purdue University.

References

1. Hartwig S, et al. Secretome profiling of primary human skeletal muscle cells. *Biochim Biophys Acta*. 2014; 1844(5):1011–7. [PubMed: 23994228]
2. Neto MF, et al. Sonodelivery Facilitates Sustained Luciferase Expression from an Episomal Vector in Skeletal Muscle. *Materials (Basel)*. 2015; 8(7):4608–4617. [PubMed: 26834950]
3. Zhu S, Lee DA, Li S. IL-12 and IL-27 sequential gene therapy via intramuscular electroporation delivery for eliminating distal aggressive tumors. *J Immunol*. 2010; 184(5):2348–54. [PubMed: 20139275]
4. Liao KL, Bai XF, Friedman A. Mathematical modeling of interleukin-27 induction of anti-tumor T cells response. *PLoS One*. 2014; 9(3):e91844. [PubMed: 24633175]
5. Menne KM, Hermjakob H, Apweiler R. A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics*. 2000; 16(8):741–2. [PubMed: 11099261]
6. Eddy SR. Hidden Markov models. *Curr Opin Struct Biol*. 1996; 6(3):361–5. [PubMed: 8804822]
7. Krogh A, et al. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol*. 1994; 235(5):1501–31. [PubMed: 8107089]
8. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A*. 1987; 84(13):4355–8. [PubMed: 3474607]
9. Bagos PG, et al. Combined prediction of Tat and Sec signal peptides with hidden Markov models. *Bioinformatics*. 2010; 26(22):2811–7. [PubMed: 20847219]
10. Bendtsen JD, et al. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*. 2004; 340(4):783–95. [PubMed: 15223320]
11. Kall L, Krogh A, Sonnhammer EL. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res*. 2007; 35:W429–32. Web Server issue. [PubMed: 17483518]
12. Oliver TF, et al. High speed biological sequence analysis with hidden Markov models on reconfigurable platforms. *IEEE Trans Inf Technol Biomed*. 2009; 13(5):740–6. [PubMed: 19273034]
13. Razmara J, et al. Artificial signal peptide prediction by a hidden markov model to improve protein secretion via *Lactococcus lactis* bacteria. *Bioinformation*. 2013; 9(7):345–8. [PubMed: 23750078]
14. Neron B, et al. Mobyle: a new full web bioinformatics framework. *Bioinformatics*. 2009; 25(22):3005–11. [PubMed: 19689959]
15. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc*. 2010; 5(4):725–38. [PubMed: 20360767]
16. Gasteiger, E.; Gattiker, A.; Duvaud, S.; Wilkins, MR.; Appel, RD.; Bairoch, A. Protein Identification and Analysis Tools on the ExPASy Server. In: Walker, JM., editor. *The Proteomics Protocols Handbook*. Humana Press; 2005. H.C.
17. Le Bihan MC, et al. In-depth analysis of the secretome identifies three major independent secretory pathways in differentiating human myoblasts. *J Proteomics*. 2012; 77:344–56. [PubMed: 23000592]
18. Yi Z, et al. Global relationship between the proteome and transcriptome of human skeletal muscle. *J Proteome Res*. 2008; 7(8):3230–41. [PubMed: 18613714]
19. Drexler HC, et al. On marathons and Sprints: an integrated quantitative proteomics and transcriptomics analysis of differences between slow and fast muscle fibers. *Mol Cell Proteomics*. 2012; 11(6):M111. 010801. [PubMed: 22210690]
20. Ojima K, Oe M, Nakajima I, Shibata M, Chikuni K, Muroya S, Nishimura T. Proteomic analysis of secreted proteins from skeletal muscle cells during differentiation. *EuPA Open Proteomics*. 2014; 5:1–9.

21. Emanuelsson O, et al. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc.* 2007; 2(4):953–71. [PubMed: 17446895]
22. Emanuelsson O, et al. Predicting subcellular localization of proteins based on their N- terminal amino acid sequence. *J Mol Biol.* 2000; 300(4):1005–16. [PubMed: 10891285]
23. Frank K, Sippl MJ. High-performance signal peptide prediction based on sequence alignment techniques. *Bioinformatics.* 2008; 24(19):2172–6. [PubMed: 18697773]
24. Massahi A, Calik P. In-silico determination of *Pichia pastoris* signal peptides for extracellular recombinant protein production. *J Theor Biol.* 2015; 364:179–88. [PubMed: 25218497]
25. Hatsuzawa K, Tagaya M, Mizushima S. The hydrophobic region of signal peptides is a determinant for SRP recognition and protein translocation across the ER membrane. *J Biochem.* 1997; 121(2): 270–7. [PubMed: 9089400]
26. Bird P, Gething MJ, Sambrook J. The functional efficiency of a mammalian signal peptide is directly related to its hydrophobicity. *J Biol Chem.* 1990; 265(15):8420–5. [PubMed: 2187865]
27. Zhang L, Leng Q, Mixson AJ. Alteration in the IL-2 signal peptide affects secretion of proteins in vitro and in vivo. *J Gene Med.* 2005; 7(3):354–65. [PubMed: 15619290]
28. Rutkowski DT, et al. Signal sequences initiate the pathway of maturation in the endoplasmic reticulum lumen. *J Biol Chem.* 2003; 278(32):30365–72. [PubMed: 12771148]
29. Andrews DW, et al. Sequences beyond the cleavage site influence signal peptide function. *J Biol Chem.* 1988; 263(30):15791–8. [PubMed: 3170612]
30. Jarjanazi H, et al. Biological implications of SNPs in signal peptide domains of human proteins. *Proteins.* 2008; 70(2):394–403. [PubMed: 17680692]
31. Pidasheva S, et al. Impaired cotranslational processing of the calcium-sensing receptor due to signal peptide missense mutations in familial hypocalciuric hypercalcemia. *Hum Mol Genet.* 2005; 14(12):1679–90. [PubMed: 15879434]

Highlights

- Hidden Markov modeling (HMM) generates signal peptides (SP) for cytokine muscle secretion
- HMM can analyze signal sequences of a skeletal muscle virtual secretome dataset and emit SP
- Emitted SP have a higher degree of hydropathy and helical composition relative to native SP
- Emitted SP may promote enhanced secretion of cytokines or propeptides from skeletal muscle

Figure 1. Graphical representation of signal peptide strength and expression level of several native interleukins and myokines of skeletal muscle

Several proteins enriched in a skeletal muscle virtual secretome from differentiated primary human skeletal muscle cells (hSkMC) are depicted based on validation by transcriptomics (expression level) and SignalP4.1-predicted signal peptide strength. The whole virtual secretome with high-level expressed myokines is shown with expression exceeding cutoff of that of beta2 microglobulin (B2M) and ATP Synthase, H⁺ Transporting, Mitochondrial F1 Complex, Beta Polypeptide (ATP5B) housekeeping genes. Strength of SP > 0.47 considered positive for SP as determined by SignalP 4.1. The top 14 SP of the myokines ranked the highest with this approach were chosen for further HMM analysis based on levels exceeding those of ATP5B control. Blue diamonds represent all myokine datapoints; yellow diamonds represent native interleukins within the dataset. ACTB, beta actin, FN1, fibronectin 1.

```

P09486|SPRC_HUMAN      -----MRA-WIFFLLCLAGRALAAPQ-----QEALPDETEVV-EETVAEVEV
P01033|TIMP1_HUMAN     -----MAPFEPLASGILL-----LLWLIAPSRA-----CT-----CVP-PHPQT-----
P07996|TSP1_HUMAN      -----MG-----LAWGLGVLFMHVCGTNRIPESGGD-----NSVDFIFELTGAAR
Q15063|POSTN_HUMAN     -----MIPFLPMFSLLLL-----LI--VNPI-----NANNHYDK-----
Q15582|BGH3_HUMAN      -----MALFVRLALALA-----LA--LGPAAT-----LA-----GPAKSPYQL-----
P02452|CO1A1_HUMAN     -----MFSFVDLRLLLLAATALLTHGQEE-----
P07237|PDIA1_HUMAN     -----MLRRALLCLAAVALVR-----ADAPEEEDHVLVLRKSNFAEA
P21810|PGS1_HUMAN      -----MWPLWRLVSLALSQALPFEQRFWDFTLDDG-PFMMNDEEASGADT
P14625|ENPL_HUMAN      -----MRALWVLGLCCVLLTFGSVRADD--EVDVDGTVEE--D--LGKSR
P50454|SERPH_HUMAN     -----MRSLLLLSAFCLLEAALAAEVKKPAAAAAPGTAEK-LSPKAATLAER
P08123|CO1A2_HUMAN     -----MLSFVDTRLLLLAVTLCCLATCQSL-----
P08253|MMP2_HUMAN      MEALMARGALTGFLRALCLLGCLLSHAAAAPSPPIKFPGDVAPKT-----DK
P05067|A4_HUMAN        -----MLPGLALLLL-----AAWTARALEVPTDGNA-----GLLAEPQIA-----
Q12841|FSTL1_HUMAN     -----MWKRWLALALVA-----VAWVRAEEELRSKS-----KICAN-----

```

Figure 2. A sample alignment prior to clustering analyses and emitting optimized consensus sequences

Fourteen candidate genes identified in a pre-screen (Figure 1) were used to build a consensus HMM model with *hmmbuild*. The output multiple alignment files were used to create an HMM model based on maximum likelihood parameter estimation.

Figure 3. Hydropathy plots of emitted sequences SP1 or SP2 compared to Native SP

Using information from Signal-BLAST that identified the closest in homology native SP to be COL1A2 (for SP1) and Mmp2 (for SP2), we compared SP1 and SP2 to these native SP using ProtScale. The emitted signal peptide SP1 had a greater hydrophobicity extent between aa 5-13 compared to a native sequence, and the emitted SP2 had a greater hydrophobicity between aa 8-17 and 21-25 compared to the native SP.

A

Schematic representation of the usual composition of a signal peptide. Average length: 22 to 30 AAs

NH₃⁺ Positive or Polar AAs Hydrophobic AAs -3 -1
N-region H-region C-region Cleavage Site

α -Helix β strand

Recognized by signal peptidase I

Native CoIa2

Sequence MSLFVQETILLIAVLTLGLATCGQLQEKTVRKNGPAGRGPRGR
Prediction CCCCHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCC
Conf. Score 9015655368986665200157688899898989898988

Helix; SStrand; CCoil

SP score: 0.804

SP1

Sequence MLALFDLEILLLIALAALAATAQLSFPFEERTLQLAEKNATR
Prediction CCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHCCCCC
Conf. Score 936999999999999999999987402866579999998716960

Helix; SStrand; CCoil

SP score: 0.932

Native MMP2

Sequence HEALKARGALTPLRALCLGCTLLSHAAASPPIKFQGVAPPT
Prediction CCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHCCCCC
Conf. Score 98430457887659999999999810259987054789989864

Helix; SStrand; CCoil

SP score: 0.845

SP2

Sequence HEALAKANASPPRLALLIGLVLLLAAPSEIFPGAVVEIEE
Prediction CCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHCCCCC
Conf. Score 95699998972431899999999999603833898666543887

Helix; SStrand; CCoil

SP score: 0.961

Table 1

Top six HMM peptide sequences emitted and validated by analyses using signal peptide identification and strength prediction software.

HMM peptide sequence (predicted cleavage site underlined)	Signal Peptide <i>S</i> -scores		Bit score (Signal-BLAST)
	SignalP 4.1	TargetP1.1	
SP1 MLALFVDLRLLLLLAAALAAALATAQSLPEEEEEETLQLAEKSAGTKVIQVCK	0.932	0.963	30.0
SP2 MEALMARMASFVPLRALLLLGVLLLLATAAPSEIPKGAEVVEIEEAVKDPD	0.961	0.966	40.4
SP3 MWSFLLQIFLLAVSPSERLNPESPGFITEVAETQPEEKKAPTKLRQVKSQ	0.826	0.892	25.4
SP4 MAKKWGPVGACLLLLLAAVALSEGKQLVNRVPEYDATREGVTGKNDRM	0.923	0.959	28.5
SP5 MAAGIRAVAQLLLLQVARVSAVSATKRTQKQTIKNHNTQESTIEIGSYSTV	0.846	0.651	25.8
SP6 MRLFYDLVSLIFLVVAIPLCGILQRVANETSPKVFMLFDPDEGAEVFAVTA	0.935	0.958	18.1