Biochemical and Biophysical Research Communications 483 (2017) 972–977

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Signal peptidase I processed secretory signal sequences: Selection for and against specific amino acids at the second position of mature protein

Yaramah M. Zalucki, Michael P. Jennings*

Institute for Glycomics, Griffith University, Gold Coast Campus, Gold Coast, Qld 4222, Australia

ARTICLE INFO

Article history:
Received 17 December 2016
Accepted 10 January 2017
Available online 11 January 2017

Keywords:
Signal peptidase I
Signal peptide
Cleavage site
Secretion
Amino acid bias

ABSTRACT

Signal peptides direct proteins from the cytoplasm to the periplasm. These N-terminal peptides are cleaved upon entry to the periplasm by either signal peptidase I, or signal peptidase II for lipoproteins. Signal peptidase I is a serine protease that has either a serine-lysine or serine-histidine catalytic dyad present in the active site. The recognition site for signal peptide cleavage by signal peptidase I has been defined primarily by an Ala-X-Ala motif at the C-terminal end of the signal peptide, one amino acid away from the cleavage site. We used a verified set of signal peptidase I cleaved proteins from *E. coli* to look for novel conserved features, focusing on the N-terminus of the mature protein. We observed a striking bias for the presence of acidic residues at second position of the mature protein (P2′), and a complete absence of aromatic amino acids at the same position. Whole genome analysis of the predicted set of all *E. coli* and *B. subtilis* secreted proteins confirmed the same strong bias for acidic residues at P2′ of the mature protein, and against aromatic amino acids at the same position. When these studies were extended to archaeal genomes (*M. voltae* and *S. tokodaii*) and the yeast genome from *S. cerevisiae*, this bias was not observed. *E. coli* and *B. subtilis* primarily express a signal peptidase I contains a serine-lysine catalytic dyad, whilst those of archaeal and eukaryotic origin generally have a serine-histidine catalytic dyad. This difference may explain the differential bias for acidic residues and against aromatic residues at P2′. These observations suggest additional key residues that may favor or prevent signal sequence recognition or cleavage by signal peptidase I, and thereby facilitate more accurate *in silico* prediction of signal peptidase I cleavage sites.

1. Introduction

In bacterial species, about 20% of all proteins are exported outside the cytoplasm. Proteins destined for the periplasm via the Sec pathway are synthesized with an N-terminal signal peptide that is cleaved upon entry to the periplasm by signal peptidase I or signal peptidase II for lipoproteins [1]. Signal peptidase I is also responsible for the cleavage signal peptides of proteins exported by the twin-arginine translocation system (Tat) [2]. The signal peptide contains three distinct regions: a positively charged N-terminus, a hydrophobic core and signal peptide cleavage site (Fig. 1, [3]). The signal peptide cleavage site has been defined primarily by a canonical Ala-X-Ala motif at positions −3 to −1 relative to the cleavage site [1].

In general, residues that comprise the N-terminal region of the mature protein are either electroneutral or electronegative [4–6] (Fig. 1). Analysis of the start of mature proteins reveal a high prevalence of acidic residues, particularly in bacterial species [7]. Previous studies have shown that the addition of the signal peptide to the N-terminus of a cytoplasmic proteins is not sufficient to allow export to the periplasm in gram-negative bacteria [8–10]. Experimental evidence has shown that long hydrophobic stretches after the signal peptide cleavage site prevent export to the periplasm of pIII protein [11]. Hence, these studies indicate that N-terminus of the mature protein, adjacent to the signal peptide may contain amino acids that are permissive for, or prevent, efficient recognition and/or cleavage by signal peptidase 1.

Whilst mutations in the signal peptide can lead to an abrogation of protein export or signal peptide processing [12,13], very few such mutations have been found on the mature side of the signal sequence cleavage site. One such mutation was found for maltose

* Corresponding author.
E-mail address: m.jennings@griffith.edu.au (M.P. Jennings).

http://dx.doi.org/10.1016/j.bbrc.2017.01.044

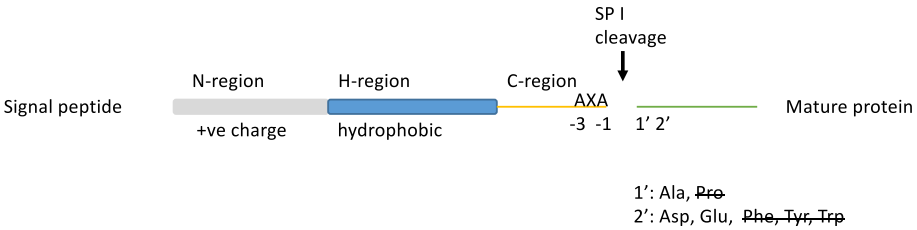

Fig. 1. Schematic of Sec dependent signal peptide and cleavage site as recognized by signal peptidase I. The biased residues at P1′ are based on [30], while the P2′ biased residues are based on results in this study.

binding protein, where a proline at the first position in the mature protein (P1′) prevented signal peptide processing by signal peptidase I [14]. Proline at the P1′ position can also block non-Sec dependent proteins cleaved by signal peptidase I [15]. Other studies have shown addition of basic amino acids either immediately after the signal peptide processing site [16,17] or within the first 14 residues of the mature protein [18], can block export to the periplasm. Despite these findings, amino acids important for signal peptide cleavage are poorly defined on the mature side of the cleavage site relative to those in the C-region of the signal peptide.

In our previous work, we have analysed codon [19,20] and amino acid [20] biases in signal peptides. In this current study, we utilize the same approaches to examine amino acid biases by position at the N-terminus of the mature proteins from species from all kingdoms of life (eubacterial or prokaryotic, archaeal and eukaryotic).

2. Materials and methods

2.1. Open reading frame sequences

All open reading frames and comprising the data-sets used in this study were downloaded from various databases. The *E. coli* protein sequences (without pseudogenes) dataset was downloaded from www.ecogene.org. The *Bacillus subtilis* str168 protein dataset was downloaded from http://genolist.pasteur.fr/SubtiList/. The archaea protein data-sets used in this study (*Methanococcus voltae* A3 and *Sulfobolus tokodaii* str_7) were downloaded from http://bacteria.ensembl.org/ensembl. While the *Saccharomyces cerevisiae* strain S288C protein dataset were downloaded from www.yeastgenome.org.

2.2. Identifying Sec dependent proteins

For *E. coli, B. subtilis* and *S. cerevisiae*, the first 70 amino acids were entered into the SignalP server (cbs.dtu.dk/services/SignalP/) and run using version 4.0 [21]. For the archaea protein sets, the web-tool http://www.compgen.org/tools/PRED-SIGNAL was used to identify Sec signal containing proteins [22]. All proteins that scored positive for a signal peptide by either method were aligned by signal peptide length for further analysis in Microsoft excel. All predicted signal peptide proteins and their cleavage sites are listed in Supplemental data.

2.3. Ecogene verified SP I dataset

The verified signal peptidase I (SP I) cleaved proteins (topic ID 560) was downloaded from ecogene database [23]. For all 143 proteins in this dataset, the signal peptidase I cleavage site was confirmed by checking the record of each protein on the website. These were entered manually and all proteins were aligned by signal peptide cleavage site for further analysis.

2.4. Statistical analysis

All amino acids at the first five positions after the signal peptide cleavage site were analysed for biases. One such analysis was done using Chi-Square test: ((observed − expected)^2/expected), with the expected value for an amino acid being the average use from positions P2-200 of all genes in the genome. P1 was not included in the calculation to avoid over representing methionine in the predicted values. To calculate the p-value for any amino acid biases, 19° of freedom was used. Biases were also tested amino acid grouped by chemical properties. The categories of chemical properties were hydrophobic (alanine, leucine, isoleucine, glycine and valine), aromatic (phenylalanine, tyrosine and tryptophan), polar non-charged (methionine, threonine, asparagine, glutamine and serine), basic (lysine, arginine and histidine) and acidic (glutamic acid and aspartic acid). Proline and cysteine were kept separate due to their unique chemical properties. Any biases amongst these amino acid categories was tested using difference of proportion test. The population proportion was the proportional use of each category amongst the P2-200 dataset of all proteins.

2.5. Calculation of average net-charge

The net charge by position (i) was calculated according to the following formula:

$$((\text{lysine} + \text{arginine}) - (\text{aspartic acid} + \text{glutamic acid}))_{(i)} \Big/ n_{(i)}:$$

where n is the number of proteins at position i. Basically it is the sum of all basic residues (arginine and lysine) minus the sum of all acidic residues (aspartic acid and glutamic acid) at position (i) divided by the number of proteins in the data-set at position (i).

3. Results

3.1. Analysis of an E. coli dataset of verified signal peptidase I cleaved secreted proteins for residues in the mature N-terminus that may impact recognition and cleavage

In our previous work, we analyzed the amino acid usage at the N-terminus of signal peptides to determine if there were any positional codon or amino acid biases [20]. Here we extended that analysis to the N-terminal region of the mature protein. A summary of the canonical signal peptide and conserved cleavage site residues are shown in Fig. 1. To determine whether there are any biases in amino acid usage on the mature side of the signal sequence cleavage site, we first analyzed proteins from *E. coli* that are known to be cleaved signal peptidase I. These include Sec-dependent signal peptides and twin-arginine exported signal peptides, as both are cleaved by signal peptidase I [1]. One such experimentally confirmed group exists on the ecogene database (topic ID 560) [23].