Analysing the sequence diversity of secreted signal peptides in the *Bacillus* genus through novel machine learning techniques

Pranav Pativada

Supervisor: Dr Joseph Brock

ABSTRACT

Optimising recombinant protein secretion has been vital for a myriad of biotechnological applications. Protein secretion is primarily involved with the signal peptide and it's interaction with the Sec pathway. Current methods for recombinant protein secretion are bottlenecked as they require excessive monitoring, are time and labour intensive, and lack of a generalised, non-protein specific secretion approach, resulting in low protein throughputs. Combinatorial methods to counteract these consequences include engineering signal peptides that are protein-generic and can be appended to a BEV and used under the MoClo Golden Gate Cloning toolbox, allowing for a wide range of combinations for optimisation purposes. This project focuses on presenting a signal peptide library that optimises signal peptide interactions in the Sec pathway under the MoClo Golden Gate Cloning toolbox. To do so, we use the common t-SNE machine learning model to extract a set of 17 maximally diverse signal peptides can be utilised under the MoClo toolbox from the *Bacillus* genus.

INTRODUCTION

The control and expression of recombinant protein secretion has enabled an array of emerging opportunities in the fields of synthetic biology and biotechnology (Freudl 2018). Optimisation and maneuverability of these procedures can serve as a compelling application in large-scale production of heterologous proteins (Burdette et al. 2018). Metabolic engineering with model organisms such as *Saccharomyces cerevisiae* and *Bacillus subtilis*, combined with these potential optimisations, can result in the fast processing of pharmaceuticals and high throughputs of therapeutic development (Freudl 2018, Moore et al. 2016). These could provide great value to communities involved in research and academia and can further accelerate the scope of synthetic biology.

Bacterial Expression Systems

Expression systems for recombinant protein production exist in both eukaryotes and prokaryotes and are currently in use for the aforementioned biotechnological applications. Current bacterial expression methods include engineering a bacterial expression vector (BEV) which is transfected into the model organism (Cantoia et al. 2021). Figure 1 shows the common elements of a BEV (Old & Primrose 2001).

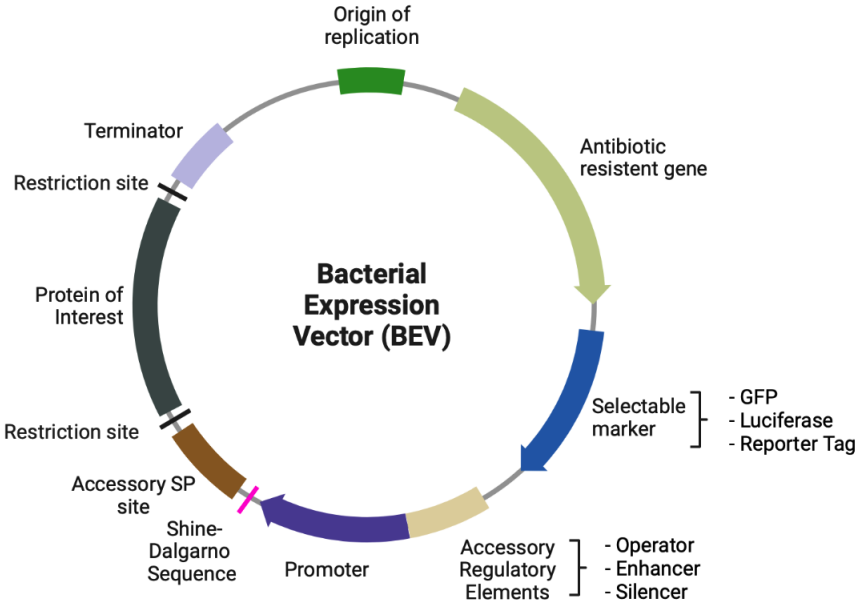


Figure 1: Elements found within a BEV used for metabolic engineering of protein production (Cantoia et al. 2021, Old & Primrose 2001). Created with BioRender.

Once a BEV is transfected into the model bacterial organism, the transcriptional cell apparatus accesses it to undergo protein synthesis. The combination of all these BEV elements (Figure 1) allows for fine-tuning and close monitoring of recombinant protein production. Through the exchange of these elements, control of protein yield can be achieved. This is essential for the variety of applications mentioned. A key part of the BEV is the accessory signal peptide (SP) site, which contains a SP. Found just after the promoter and before the protein of interest, a SP is a stretch of amino acids (20-30) that is present at the N-terminus of a newly synthesised and unfolded protein (referred to as

a precursor protein) (Anné et al. 2016, Zimmermann 2009). The SP's primary role is to transport the precursor protein to a destination (Freudl 2018, Prabudiansyah & Driessen 2016). In the case of bacteria, this is usually out of the cytoplasmic membrane and to the extracellular space, guaranteeing protein translocation and moreover - recombinant protein secretion (Prabudiansyah & Driessen 2016). This is important as it forgoes the use of lysis to break the cell to extract the proteins, making it faster to produce and access recombinant proteins.

However, current methods are bottlenecked as they are time and labour intensive, and also not generalised. To effectively secrete a protein of interest, it is the case that the engineered expression vectors must (i) secrete the protein to the required destination, (ii) activate the protein through post-translational modifications, and (iii) control protein levels and monitor transcription in the host organism (Burdette et al. 2018). For this purpose, we look at generalised SP's to add to the BEV model that works for any protein of interest to counteract the consequences of current bottlenecked methods.

Gram-positive Sec Pathway

Gram-positive bacteria are especially efficient hosts as they are easy to handle and uncomplicated (Freudl 2018). They only possess a singular membrane (the cytoplasmic membrane) and thus the export of a target protein across this one barrier can result in the direct release to the extracellular space (Burdette et al. 2018, Freudl 2018). This makes them particularly useful for recombinant protein secretion. The export of proteins in hosts of this class out of their singular cytoplasmic membrane is likely to be utilised using the general secretion (Sec) pathway (Freudl 2013, Prabudiansyah & Driessen 2016, Anné et al. 2016). Protein translocation via the Sec pathway occurs post-translationally or co-translationally (Freudl 2013).

Post-translational secretion initially involves complete translation of the protein by the ribosome before the SP and precursor engages with the Sec pathway. Once translated, a post-translationally interacting protein (PIP), also called a chaperone protein, stabilises the protein in an translocation-competent state, allowing it to remain unfolded (Freudl 2013). It then directs the protein to the translocation site SecA (Prabudiansyah & Driessen 2016). In gram-negative bacteria, this chaperone function is done by usually by a protein called Sec B. Gram-positive bacteria lack this Sec B homologue (Freudl 2013, Anné et al. 2016). The CsaA protein in *Bacillus subtillus* has been identified to the bidding of SecB instead as it has shown binding affinity to SecA, but there exists no clear protein substitute for SecB (Freudl 2013, Anné et al. 2016).

Co-translational secretion, while dealing more with protein insertion rather than secretion, involves a signal recognition particle (SRP) (Green & Mecsas 2016). Precursor proteins are bound to the ribosome and are targeted by the SRP (Prabudiansyah & Driessen 2016). The SRP binds to hydrophobic N-terminus signal sequences, which is usually the signal peptide, or transmembrane segments as they emerge from the ribosome, forming a SRP/RNC (ribosome nascent chain) complex (Green & Mecsas 2016, Anné et al. 2016). This complex interacts with the SRP receptor called the FtsY and the RNC is subsequently released to the SecYEG translocation site through the heterodimerization of SRP and FtsY as well as GTP hydrolysis coming from the elongation of translation (Freudl 2013, Prabudiansyah & Driessen 2016). The precursor protein is

followed separately to SecYEG with continuing translation and is facilitated by SecA through ATP hydrolysis (Freudl 2013, Prabudiansyah & Driessen 2016). The key distinction between these transports methods apart from their different interacting proteins is that the post-translational secretion translocates the entire protein (including the SP) to the SecA site. Co-translational secretion involves translocating only the precursor to the SecA site, and the RNC is translocated to the SecYEG site.

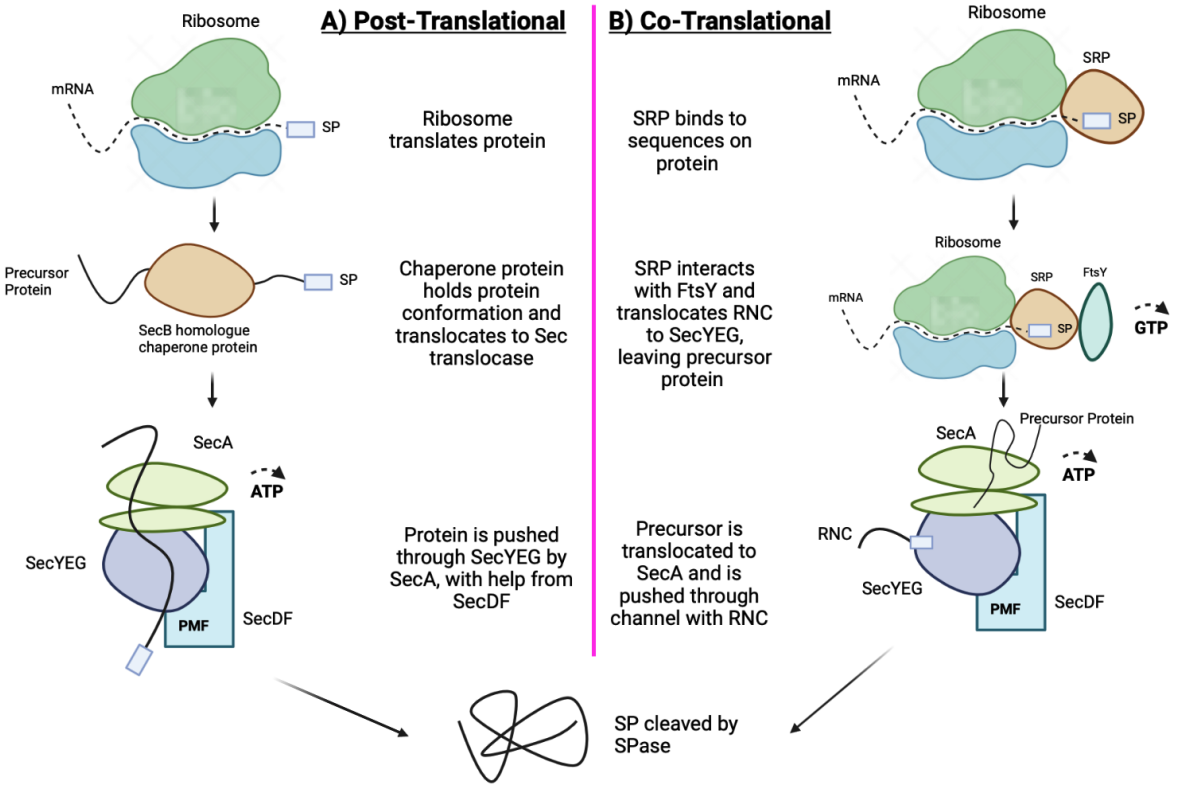


Figure 2: Summary of post-translational (A) and co-translational (B) systems in the Sec pathway and their key differences. After processing through the Sec translocase, SP's are cleaved by signal peptidase (SPase), leaving the precursor protein translocated and folded (Freudl 2018, Prabudiansyah & Driessen 2016). Created with BioRender.

The Sec translocase, consisting of SecA, SecYEG, and SecDF, is the primary machine that transports proteins (Freudl 2013). SecA is a homodimeric translocation motor protein that facilitates protein translocation across the central SecYEG pore (Prabudiansyah & Driessen 2016). It drives the translocation step through repeated cycles of ATP-hydrolysis, pushing the precursor protein through the channel. The SecA consists of the DEAD and C domains (Freudl 2013, Anné et al. 2016).

The DEAD domain consists of nucleotide-binding folds NFB1 and NFB2 and a preprotein cross-linking domain PPXD (Prabudiansyah & Driessen 2016). NFB1 and NFB2 allow for ATP binding and hydrolysis for translocation through the channel, whereas PPXD controls the binding of the protein so that it can be utilised by NFB1 and NFB2. The C domain comprises of the helical scaffold domain HSD, which controls the opening and closing of the DEAD motor (Prabudiansyah & Driessen 2016). Regulation of ATP hydrolysis is controlled by IRA1 which inhibits it, and the C-terminus linker CTL, while it's function is unknown, is important in the interaction with the chaperone protein in

post-translational secretion (Prabudiansyah & Driessen 2016). In the case that ATP hydrolysis does not provide enough energy, a proton-motive force (pmf) to further assist with protein translocation is provided by SecDF (Freudl 2018, 2013). The force exerted pulls the protein through the channel and closer to the extracellular space (Freudl 2018).

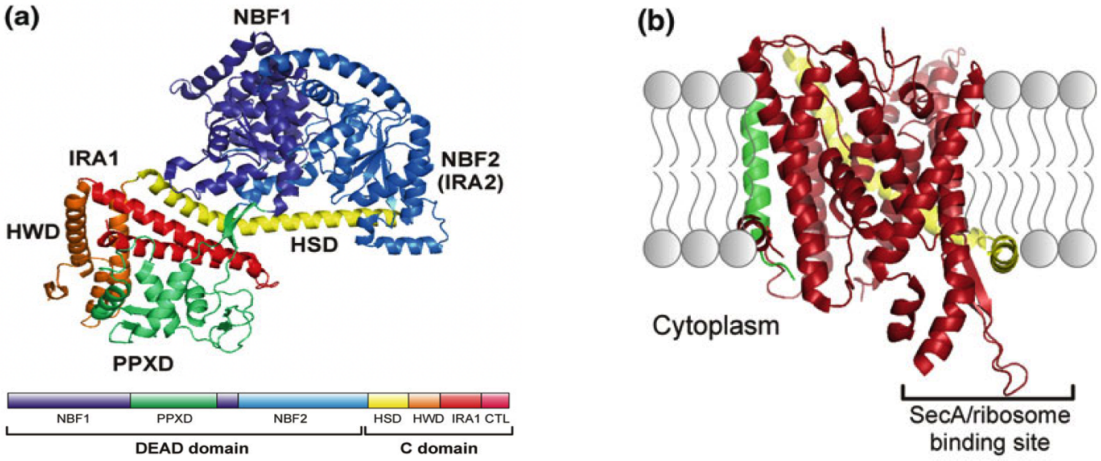


Figure 3: X-ray crystallography structures of SecA (a) and SecYEG (b). SecA diagram shows structure with the DEAD and C domain. The SecYEG diagram is a lateral structure that is viewed from the side, linking to the cytoplasm. For SecYEG, green is SecG, red is SecY and yellow is SecE.

SecY, SecE and SecG together form SecYEG, the protein conducing channel that is homologous to that of Sec61 in eukaryotes (Prabudiansyah & Driessen 2016). The channel is made up by SecY, which contains 10 α-helical transmembrane segments (TMS), organised as two domains as TMSs 1-5 and TMSs 6-10 (Freudl 2013, Prabudiansyah & Driessen 2016). These domains are connected together and form a clamshell. This is what the protein channels through to get translocated. SecE contains of one TMS and an amphiphatic helix in gram-positive bacteria, though usually there are 3 TMSs (Prabudiansyah & Driessen 2016). This one TMS and helix reinforce the clamshell structure and allow for flexibility of SecY depending on the incoming protein. The final part, SecG, carries the same function as the eukaryotic Sec61β. It shows limited contact with SecY and is not necessary, but improves the efficiency of translocation and assosciates with SecA (Prabudiansyah & Driessen 2016). The combination of these elements form the SecYEG complex. This hourglass structure has a hydrophobic ring that allows proteins to be undisturbed by water and other ions, which ultimately results in efficient protein translocation (Prabudiansyah & Driessen 2016). Once the protein is translocated and the SP is cleaved, the protein folds into it's required orientation (Freudl 2018).

Signal Peptide Interactions

Secretion is also dependent on the SP structure, which consists of the n-domain, h-domain and c-domain (Freudl 2013, Heijne & Gavel 1988). The n-domain, being positively charged, interacts with the cytoplasmic membrane which is negatively charged, and aids the translocation process (Duffaud et al. 1985). It consists mainly of proline, glycine, or serine residues and connects to the h-domain (Duffaud et al. 1985).

The h-domain is the central part of the SP. It is a long, hydrophobic stretch of amino acids that is usually where the SRP binds to for co-translational processing (Anné et al. 2016, Heijne & Gavel 1988). Binding affinity to the SRP is influenced by the strength of this hydrophobic sequence. This then impacts the FtsY receptor and subsequent translocation to the SecYEG site. The h-domain consists of a random distribution of 5 amino acids as seen in Table 1 (Duffaud et al. 1985). Though the reason for these specific amino acids is unknown, it is likely to help the SP be in a specific conformation and carry out it's function (Duffaud et al. 1985). The c-domain is the uncharged end of the SP. It contains the cleavage site that usually has either an Alanine or Glycine residue, which is targeted by the signal peptidase (Duffaud et al. 1985).

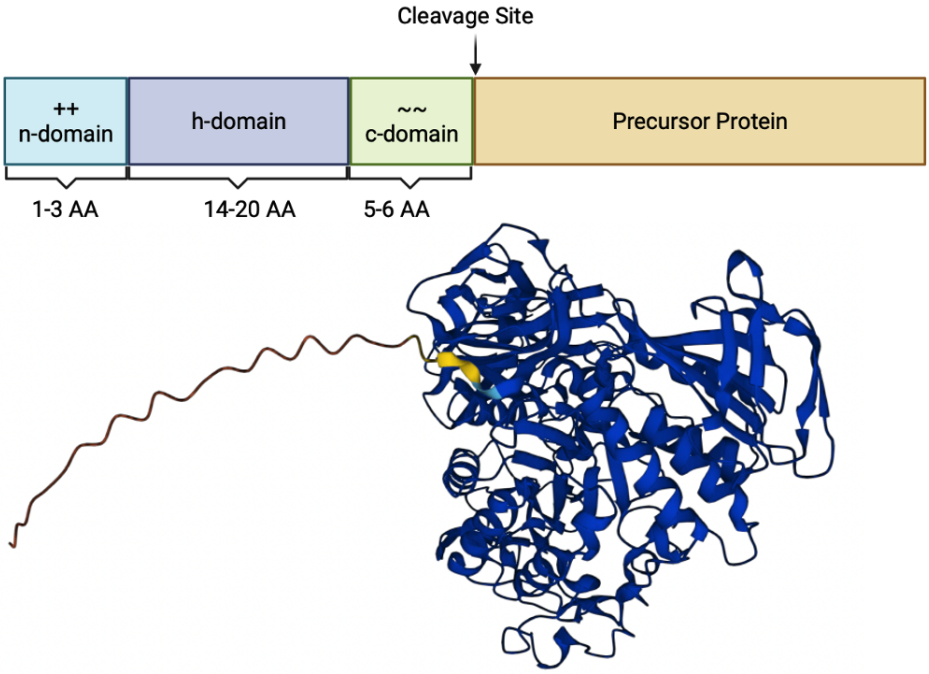


Figure 4: Diagram of domains found in a SP alongside the structure of Cyclomaltodextrin glucanotransferase (cgt) protein in *Bacillus subtillis* connected to a SP. Visualised using Alphafold and created with BioRender (AlphaFold 2021).

Domain	**AA Frequencies**
n-domain (positively charged)	Pro, Gly, Ser
h-domain	Ala, Val, Leu, Tyr, Met
c-domain (uncharged)	Ala, Gly at C-terminus

Table 1: Regularly occurring amino acids within the domains of a SP (Duffaud et al. 1985)

Analysis of many SP's have lead to the above amino acid frequencies. This could be due to mutations that have accumulated through evolution. Due to this, there is compelling evidence that interactions with the SRP are to do with the properties of the domains rather than the interactions with the residues themselves (Janda et al. 2010). This allows for a diverse SP to be utilised that contain the same chemical properties, but different residues in their sequences (Nilsson et al. 2015). A diverse set of SP can thus lead to optimizing interactions with the SRP and tuning the Sec secretion system, and ultimately better control of recombinant protein production.

Aim: Combinatorial Cloning Control

Current methods for recombinant protein secretion are bottlenecked as they are time and labour intensive. Recent literature also argues that there is no universal SP that optimises and promotes secretory protein production for any heterologous protein in any bacterial host (Freudl 2018). This means that expression of recombinant proteins are specific to the protein of interest as well as the host organism. However, SP structure and amino acid frequencies show that interactions with the Sec pathway are influenced by the chemical properties of the SP domains, and not by the residues in their sequences (Janda et al. 2010, Nilsson et al. 2015). This allows us to have a diverse set of SP that have different sequences within their domains that can influence interactions with Sec pathway, ultimately influencing recombinant protein secretion.

MoClo Golden Gate Cloning is a popular expression system for recombinant proteins that utilises modular vector assembly. BEV's are designed with a series of discrete, linked modules that attach together at different positions (Geddes et al. 2019). It allows for the assembly of multiple DNA fragments in a specific, linearised order (Geddes et al. 2019, Weber et al. 2011). Given a toolbox under MoClo Golden Gate terms that includes metabolically engineered BEV expression elements, we achieve combinatorial control for expressing recombinant proteins (Moore et al. 2016). By increasing the number of SP's available that can be used under MoClo, we increase the number of combinations available that can be tested in relation to specific gram-positive hosts (Geddes et al. 2019, Freudl 2018). This greatly allows for the fine-tuning of recombinant protein secretion due to the vast number of combinations.

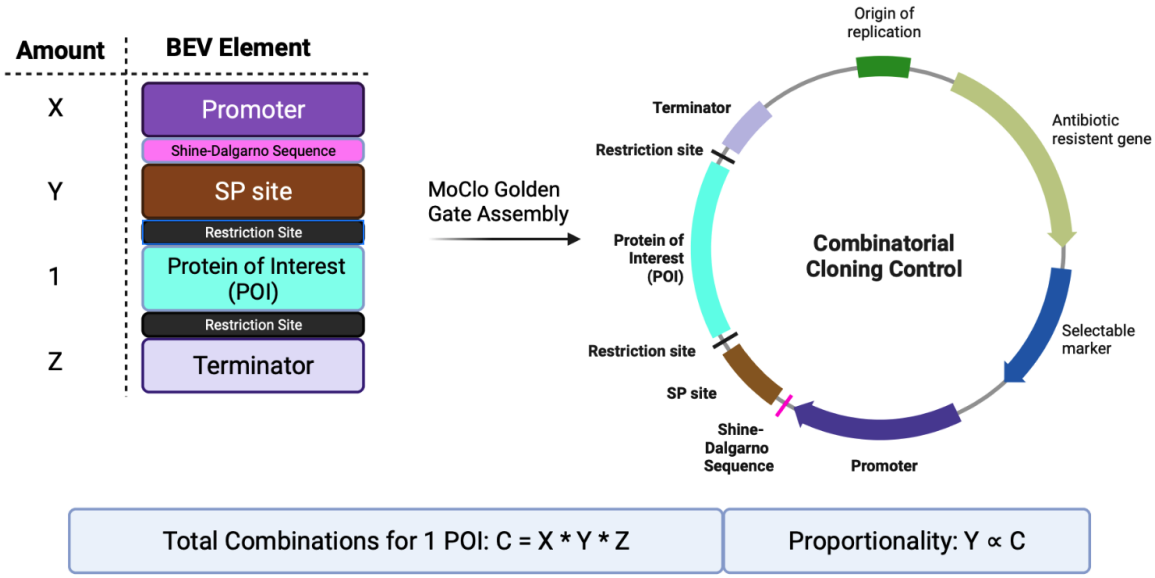


Figure 5: Combinatorial cloning control through MoClo Golden Gate Assembly given a toolbox containing a set of promoters, SPs and terminators. Maximising SP diversity and thus the amount of SP available (Y) is proportional to the combinations (C).

We screen a diverse range of SPs and create a library that can be used under MoClo cloning, allowing high throughput screening methods that can lead to fast and reliable monitoring of recombinant protein secretion. This study focuses on creating an SP library from the *Bacillus* genus. We use novel machine learning techniques to extract visually

diverse SPs that can be used for this purpose.

METHODS

Bio-informatic analysis for sequence diversity was done through 3 stages. More information for the following procedures can be found at: `https://github.com/pranav-pativada/Sequence-Diversity-of-Signal-Peptides/blob/main/Lab-Book.ipynb`

Characterisation and Processing of the Bacillus genus

138 reviewed amino acid sequences pertaining to secreted signal peptides amongst the *Bacillus* genus were collected. This was done using an advanced search on the UniprotKB database with the following as input: `locations:(location:"Secreted [SL-0243]") annotation:(type:signal) taxonomy:"Bacillus [1386] AND reviewed:yes"` (Uniprot 2022).

These sequences were downloaded from UniprotKB as a compressed `FASTA (canonical)` file. They were then input into SignalP 5.0 for the prediction of their cleavage sites with the following as input: `organism: Gram-positive, output: long format`. After being processed, results from SignalP 5.0 were downloaded as processed `gff3` entries, `JSON` summary, processed `FASTA` entry and a prediction summary.

Processed `gff3` entries containing the cleavage site predictions of the SPs were converted to a +5 post-cleavage site prediction. This is because the sequences after the cleavage site have effects on overall SP diversity. `gff3` conversion to a +5 post-cleavage site was done using `awk` commands. Converted `gff3` files were then used for `FASTA` extraction to obtain the required sequence sets needed for the different SPs. `FASTA` extraction with `gff3` entries as input was processed using the `BEDTools2` Suite (`bedtools2 v2.30.0`).

Pairwise Distance Matrix Calculations

Extracted `FASTA` files were put into MAFFT (Multiple Alignment using Fast Fourier Transform) to align the sequences. This was done with the MAFFT online service. The resulting `FASTA` file from MAFFT sequence alignment was then used as input for a pairwise distance matrix propagation using the `MEGA11` GUI software. Notably, the Dayhoff model was used for the calculation of the distance matrix, as this decreased the sparsity of the matrix, making it effective for t-SNE analysis later.

Sequence Diversity Visualisation

Singular value decomposition (SVD) followed by a t-distributed stochastic neighbor embedding (t-SNE) was applied to the resulting distance matrix calculated from `MEGA11` for visualisation. A dimensionality of 3, alongside a `precomputed` state as the metric and a `random_state` of 0 was chosen for both SVD and t-SNE. A `precomputed` metric assumes that the input given is the form of a distance matrix and the 0 `random_state` was chosen to ensure consistency. SVD was implemented with 12 components, 13 iterations, whereas t-SNE was implemented with 3 components (the dimensionality) and 250 iterations. A `perplixity=30` was chosen to standardise the clusters around the origin point.

A 3D scatter plot was constructed as the output for visualisation using the `plotly` package. `matplotlib` was used to separate the t-SNE dimensions into x-y, x-z, and y-z respectively. A neighbourhood (`n_hood`) under the Euclidean metric space was initialised with a value of 30 (`n_hood=30`) to obtain the most visually diverse SP's. Relevant packages used for for the SVD and t-SNE implementations are `scikit-learn, pandas, numpy, scipy` and `functools`.

RESULTS AND DISCUSSION

Optimising recombinant protein secretion with the BEV model under MoClo Cloning requires a library of maximally diverse SPs. This is because SPs with varying diversity can be fitted into a BEV such that their interactions with the SRP and greater Sec pathway can be carefully studied, allowing for protein control and expression. We suppose a generalised approach of these diverse SP that are not protein specific, which is complemented by the MoClo method. The following results show the significance of the dataset used and the extracted SPs from the dataset that make up the library that we have created.

Significance of UniprotKB Database

Subset of SignalP 5.0 Processing of Characterised SPs from UniprotKB		
SP name	**SignalP 5.0 AA Cleavage Site Prediction**	**SignalP 5.0 Likelihood Score**
sp_B1B6T1_PTLY_BACSP	ASALNSGKVNPLADFSLKGFAALNG GTTGGEG	0.9942
sp_O31803_YNCM_BACSU	**QVAKAASELPNGIGGRVYLNSTGA VFTAKIVLPETVKNNDSVSTPYI**	**0.7453**
sp_D4G3R4_WAPA_BACNB	KTTEEENGNRIVADDPEETLQKEQTE EAVPFDPKDINKEGEITSERTENT	0.8630
sp_O34344_SDPC_BACSU	**KENHTFSGEDYFRGLLFGQGEVGK LISNDLDPKLVKE**	**0.7189**
sp_P36550_CWLL_BACLI	**N/A**	**0.2297**
sp_P40949_TAPA_BACSU	**AFHDIETFDVSLQTCKDFQHTDKN CHYDKRWDQSDLHISDQTDTKGTV**	**0.7525**
sp_G4NYJ6_WAPA_BACS4	KTTEEEAGNRIVSDDPEETPRNEQTEE AVPFPSKDIN	0.9749
sp_P68569_BDBA_BACSU	**EKPFYNDINLTQYQKEVDSKKPKFIY VYETS**	**0.7595**
sp_Q45071_XYND_BACSU	**ATSTTIAKHIGNSNPLIDHHLGADPVA LTYN**	**0.7722**

Table 2: A sized 10 subset of the 138 sequences from UniprotKB processed through SignalP 5.0 for their cleavage site and SP likelihood.

We first analyse the 138 sequences gathered from the UniprotKB database to ensure their correctness and characterisation as a SP through SignalP 5.0 processing. This shows a prediction of the cleavage site and SP likelihood. We select a confidence likelihood score of

0.8 and search the sequences processed by SignalP 5.0. Table 2 shows a 10 sized sample subset of the 138 UniprotKB sequences processed. Seen in bold are the 6 classified sequences from UniprotKB that have a SP likelihood score of less than 0.8.

In specific, the sp_P36550_CWLL_BACLI identified by SignalP 5.0 has no cleavage site prediction, but was still included in the database. This raises concerns to the reliability of the database, although all sequences are characterised as reviewed. Thus, experimental validity of these sequences is necessary to test whether or not they are indeed SPs. For the purpose of reliability, we do not include the SPs with that do not meet the likelihood criteria of having a score above 0.8.

We also compare the SPs taken from UniprotKB in reference to other databases. *Bacillus subtilis* SP libraries have been identified to contain 148 experimentally validated SPs for cultinase secretion (Hemmerich et al. 2016). This shows that there are missing validated SPs pertaining to *Bacillus subtilis* in the characterisation used. It is possible that these missing SPs are part of the unreviewed sequences and have not been incorporated. Further investigation into validating these missing sequences to the database would be necessary in the case that these sequences are maximally diverse SPs. Thus, we use these experimentally validated SPs in our analysis as well to generate distinct clusters and extract maximally diverse SPs to compensate.

Sequence Diversity Visualisation

We use the data analysis and common machine learning pipeline, t-SNE, to generate clusters of SPs to visual sequence diversity. Suitable for small datasets and for maximising localisation relationships, the t-SNE approach applied can reveal the diversity between SPs, ultimately being useful for optimising protein secretion. We use a PCA-like truncated SVD with numerous iterations to reduce the dimensionality of the dataset to 3 and hence allow t-SNE to work in a localised, 3D space under a Euclidean metric space.

We implement t-SNE to be based on a standard substitution BLOSUM62 matrix generated from the MAFFT sequence alignment. We then calculate the pairwise distance matrix under the Dayhoff PAM model. This serves to reduce matrix sparsity and thus benefits the t-SNE model that will be applied to it. Repeated iterations of t-SNE with perplexities of 30 and 75 being used for model consistency were applied to the distance matrix. This yielded clusters around the origin point. We observe in Figures 6 and 7 the result of the model under a neighbourhood of radius 30 in 2D.

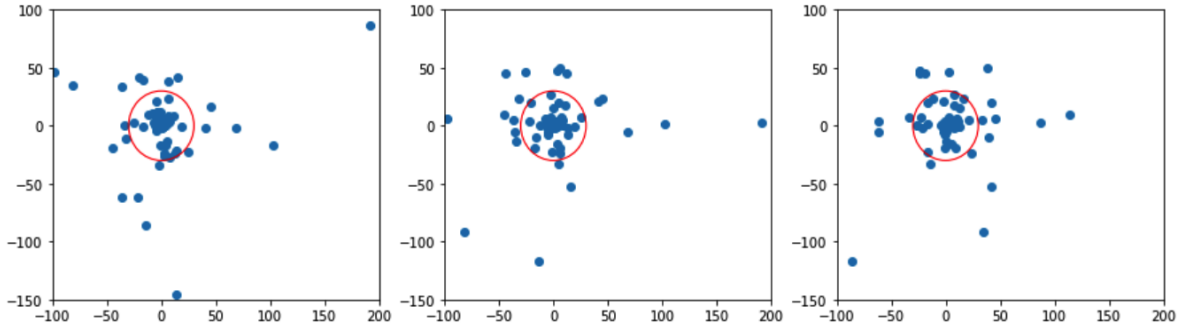


Figure 6: 2D cross-sections of t-SNE visualisation with a perplexity of 30. Left to right is in order of x-y, x-z, and y-z.

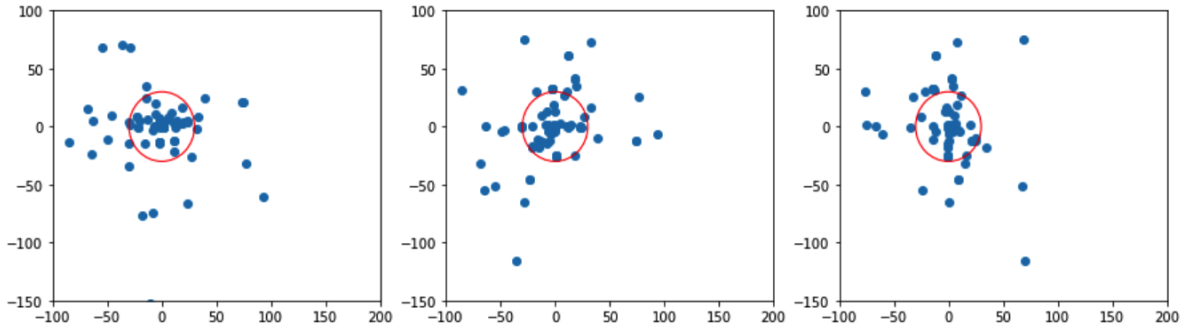


Figure 7: 2D cross-sections of t-SNE visualisation with a perplexity of 75. Order of graphs is the same as Figure 6.

17 and 22 maximally diverse SPs were found outside of the neighbourhood range for Figures 6 and 7. This shows consistency of the model, as under two different cluster generating environments, the number of diverse SPs are similar. The key difference between the two models is the model with perplexity 75 was more scattered than the its counterpart. Table 3 shows these same sequences found in both models with perplexities 30 and 75. This shows that though the models have roughly the same amount of maximally diverse SPs outside the neighbourhood range, there is little similarity between them, as Table 3 only shows that 6 occur in both. We can thus use these two models that contain different, diverse SPs to experimentally determine their effects for secretion.

SP sequence similarities in different perplexities	
SP name	**SP sequence**
sp_P00648_RNBR_BACAM	KTETSSHKAHTEAQVINTFDGVADYLQT YHKLPDNYITKS
sp_P39899_NPRB_BACSU	EESIEYDHTYQTPSYIIEKSPQKPVQNTTQ KES
sp_Q9RMZ0_Y6545_BACAN	EKKTFTDVPNWAQQSVNYLMKKALDGKP DGTFS
sp_P05655_SACB_BACSU	KETNQKPYKETYGISHITRHDMLQIPEQQK NEKY
sp_Q6YK37_XYNC_BACIU	ASDAKVNISADRQVIRGFGGMNHPAWIGD LTAAQRETA
sp_Q03091_BSN1_BACAM	GAPADTNLYSRLAVSTAGGTTLFPQTSSAVI

Table 3: The 6 same sequences that occur with the t-SNE model with perplexities 30 and 75.

Confirmation of sequence diversity

The t-SNE model shows the local relationships and similarities between the SPs in our dataset, but validation of this is not guaranteed. To confirm that perplexities of both 30 and 75 yielded maximally diverse SPs that are different to each other (apart from the re-occuring 6 in Table 3), we perform a color coded sequence alignment between them. We use ClustalOmega and sort by sequence similarity to see the diversity between the two, as seen in Figure 8 and Figure 9.

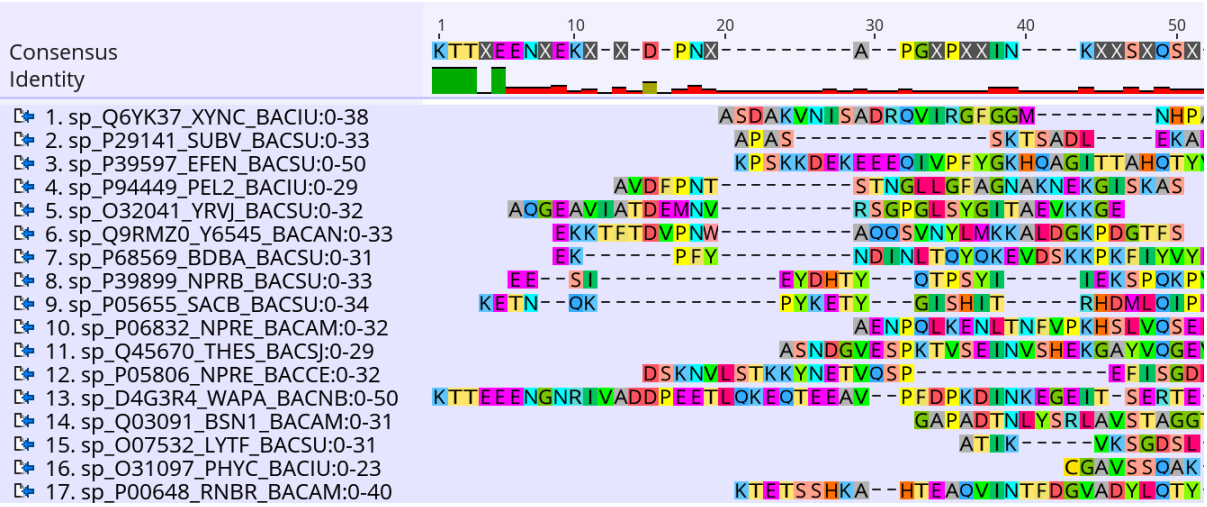


Figure 8: Sequence alignment of maximally diverse sequences gathered from the t-SNE perplexity 30 model containing 17 SPs

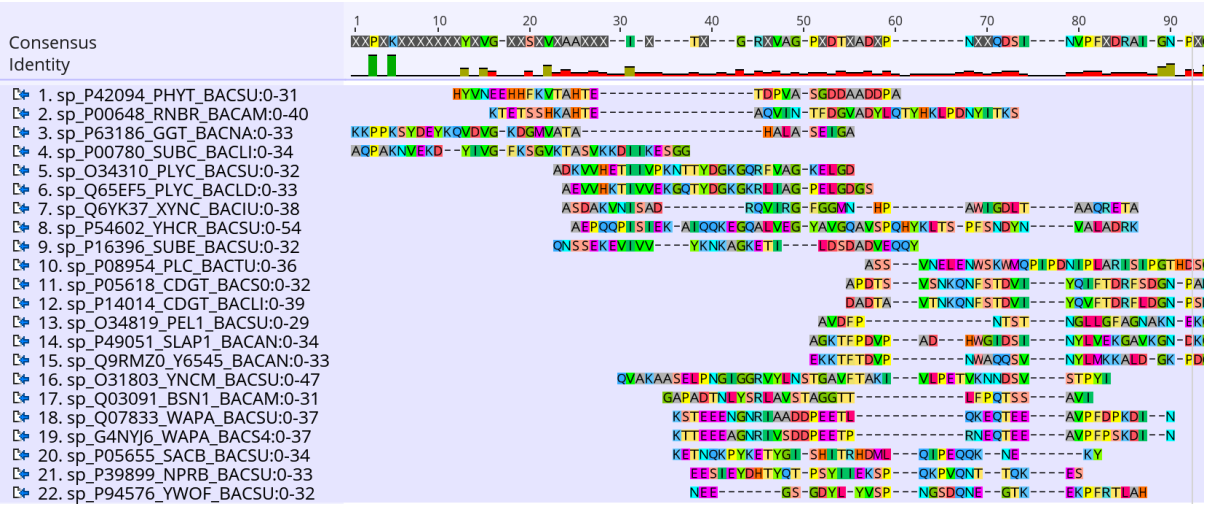
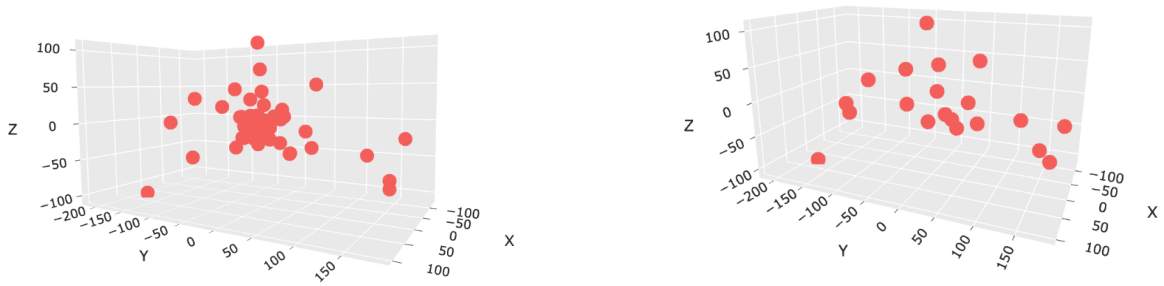


Figure 9: Sequence alignment of maximally diverse sequences gathered from the t-SNE perplexity 75 model containing 22 SPs

The ClustalOmega sequence alignment for the 17 SPs seen in Figure 8 shows scatterdness in the positions of 1-20, but then strong alignment for those after especially in the positions of 20-40. This likely means there is strong diversity in the hydrophobic region of the SPs. This is different to the set of 22 SPs seen in Figure 9, which is much more scattered and the sequence alignment prolongs to much past the length of an average SP. Alignment diversity between the 22 SPs is found in different areas. We see this in AA positions 20-50 in the first half, then in positions 60+ for the lower half of SPs. Although sequenced by diversity and not input order, this shows that the set of SPs computed by the model are inconsistent with the general due to much wider scattering. Additionally, we see that between these sets, there is much more diversity of AA residues compared to the literature specific frequent AAs as introduced in Table 1. Given there is strong diversity in the hydrophobic region, combined with the added diversity of more AAs, this could be ideal for optimising the different interactions with the SRP. For these purpose, we create our library with the 17 SPs identified in Figure 8.

Validation and Testing with 148 SP set

We finally test our results against the validated 148 SP sequence set obtained as mentioned before in a study done to measure cultinase secretion (Hemmerich et al. 2016). We scrape these SPs and then follow the procedure and apply our t-SNE model under a neighbourhood of radius 30. We also use a constant perplexity of 30 as per the more diverse and better alignment found with our own dataset for this purpose.



(a) All SPs

(b) Maximally diverse SPs outside of neighbourhood range

Figure 10: 3D representation of t-SNE model applied to 148 SP sequence set. (a) represents all 148 sequences in 3D space, whereas (b) represents only the SPs outside of the neighbourhood distance of 30.

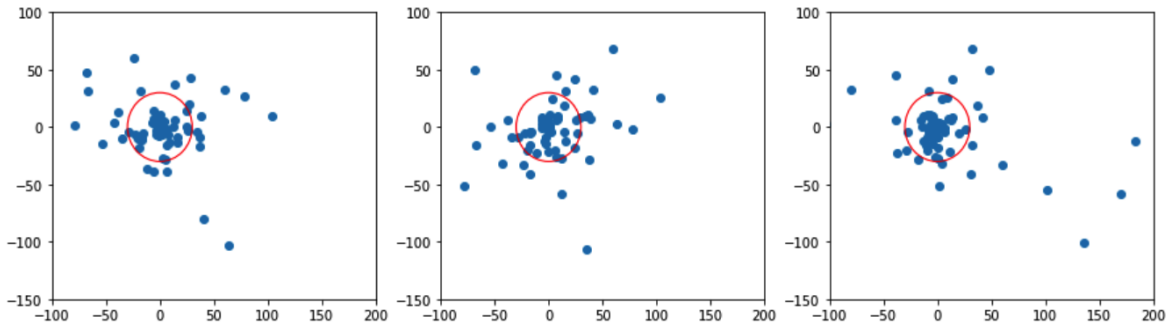


Figure 11: 2D cross-section of t-SNE model applied to 148 sequence set. Seen in red is the neighbourhood with distance of 30.

Applying the model to this dataset of SP yielded similar results to that of of our dataset. 20 SPs outside the neighbourhood range were gathered as maximally diverse SP, which is also similar to that of the 17 SPs gathered. We can thus use these SPs as well as part of the library by cross-referencing with our current 17 SP sequence set, searching for diversity and similarities. This can further optimise and fine-tune the interactions of the SP with the Sec pathway. With this model, we can append it with our own dataset, and extract SPs by potentially increasing the neighbourhood distance beyond 30. This allows us to further maximise diversity, which can then be tested with the sequence alignment as done in Figures 8 and 9.

Improvements and areas for future research

Given the required computational power, analysis can be done to verify the validity of these SPs by checking their secretory activity. This gives us insight on whether they are applicable to be used effectively with the MoClo Golden Gate system. Using an

experimentally validated secreted protein, such as GFP, and appending it to our 17 SP sequence set as well as the 20 sequence set validated from literature - we can test the destination of our SPs. This would be done using three processing tools - TargetP 2.0, BUSAC and MULocDeep. TargetP 2.0, a subcellular prediction homologue to the previously used procssing tool SignalP 5.0, is used to predict the likelihood rate and thus presence that an SP exists at the N-terminus of the protein. This allows us to confidently validate the SPs that have been appended to GFP. To minimise bias, we incorporate BUSAC and MuLocDeep for subcellular and suborganelle prediction of the SPs. BUSAC and MuLocDeep both use deep-learning and neural network algorithms for processing, indicating their robustness. Positive and consistent localisations across these models, given their reliability, can thus give further insight and warrant experimental validation of these SPs. Repeating this process with different screted proteins can then help establish the robustness of our results.

Testing multiple datasets with a different bio-informatic package, such as the *SecretSanta*, could also prove useful. *SecretSanta* allows us to screen *Bacillus* secretomes under stricter conditions, such as their localisation results (Gogleva et al. 2018). We compile datasets of relevant secretomes found in literature, such as the 220 SP set from *Bacillus licheniformis* (Freudl 2018). We can remove proteins that are targeted to the plastid or the mitochondria. Screening the multiple secretomes available from our compilation, we narrow down and improve the reliability of our datasets. We can then apply our t-SNE model to each secretome dataset available and compare the clusters in between secretomes accordingly toolbox. Compiling these different secretomes together and then performing a t-SNE analysis could also be beneficial to see the difference in sequence diversity of SP in their own secretome to the entire genus, further providing information on the accessibility of the SPs. This can provide great insight on the best host to use with the MoClo Cloning system.

Furthermore, understanding the significance of the AA sequences of the post-cleavage and their effect on sequence diversity would help in optimising interactions between SPs and the Sec pathway. Conducting three separate t-SNE analysis' at just the post-cleavage site, at 5+ the post-cleavage site, and at 10+ post-cleavage site can help in seeing this effect. Given a notable effect on sequence diversity, the most optimal post-cleavage site inclusion can be utilised and be included as part of the library to see how it affects protein secretion.

A genome-scale network reconstruction, though extremely computationally expensive and time-intensive, is a cutting-edge approach and would provide complete control of analysis for recombinant protein secretion. Using metabolic knowledge bases, we can computationally construct the biochemical pathways needed for a particular host, such as *Bacillus subtilis* (Fang et al. 2020). This allows us to predict SP interactions with the Sec pathway, and moreso predict the consequences of using diverse SPs (Fang et al. 2020). Given the reaction of the protein in our BEV, we can construct a metabolic network. This allows us to construct a solution space that can be used to optimise this reaction. Thus, using this approach we use the maximally diverse SPs in our library and optimise each SP's solution space, allowing for better control of protein secretion.

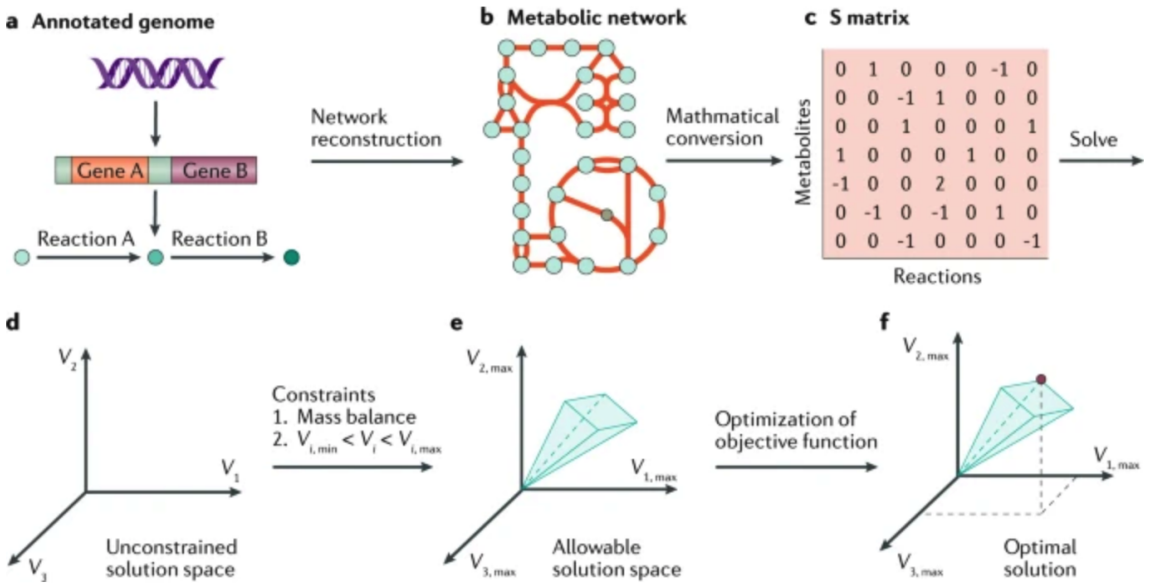


Figure 12: Computational GEM construction and optimization of the solution space using a stoichiometric S matrix.

CONCLUSION

In this project, we aimed to create and analyse a SP library of maximally diverse SPs for the purpose of optimising non-protein specific secretion under the MoClo Golden Gate model, allowing us to achieve combinatorial control. By using the novel machine learning t-SNE technique, *Bacillus* genus, we obtained 17 SPs that are maximally diverse and generally consistent with SP structure. By using likelihood and neighbourhood parameters, as well as sequence alignments to confirm their diversity, we further cross-tested with a validated 148 SP set. We conclude that these 17 SPs are a first step into allowing for the optimisation of recombinant protein secretion for *Bacillus* hosts. Future research is recommended into experimentally validating this library, as well as conducting more bio-informatic analysis. Methods outlined for this purpose are localisation processing, individual secretome analysis, effect of post-cleavage inclusion, and a construction of a genome-scale network for full optimisation of SPs and their reactions.

References

AlphaFold (2021), 'Cyclomaltodextrin glucanotransferase alphafold prediction'.
URL: *https://alphafold.ebi.ac.uk/entry/P05618*

Anné, J., Economou, A. & Bernaerts, K. (2016), 'Protein secretion in gram-positive bacteria: From multiple pathways to biotechnology', *Current Topics in Microbiology and Immunology* p. 267–308.

Burdette, L. A., Leach, S. A., Wong, H. T. & Tullman-Ercek, D. (2018), 'Developing gram-negative bacteria for the secretion of heterologous proteins - microbial cell factories'.
URL: *https://doi.org/10.1186/s12934-018-1041-5*

Cantoia, A., Aguilar Lucero, D., Ceccarelli, E. A. & Rosano, G. L. (2021), 'From the notebook to recombinant protein production in escherichia coli: Design of expression vectors and gene cloning', *Recombinant Protein Expression: Prokaryotic Hosts and Cell-Free Systems* p. 19–35.

Duffaud, G. D., Lehnhardt, S. K., March, P. E. & Inouye, M. (1985), 'Chapter 2 structure and function of the signal peptide', *Current Topics in Membranes and Transport* p. 65–104.

Fang, X., Lloyd, C. J. & Palsson, B. O. (2020), 'Reconstructing organisms in silico: Genome-scale models and their emerging applications', *Nature Reviews Microbiology* **18**(12), 731–743.

Freudl, R. (2013), 'Leaving home ain't easy: Protein export systems in gram-positive bacteria'.
URL: *https://pubmed.ncbi.nlm.nih.gov/23541477/*

Freudl, R. (2018), 'Signal peptides for recombinant protein secretion in bacterial expression systems', *Microbial Cell Factories* **17**(1).

Geddes, B. A., Mendoza-Suárez, M. A. & Poole, P. S. (2019), 'A bacterial expression vector archive (beva) for flexible modular assembly of golden gate-compatible vectors', *Frontiers in Microbiology* **9**.

Gogleva, A., Drost, H.-G. & Schornack, S. (2018), 'Secretsanta: Flexible pipelines for functional secretome prediction', *Bioinformatics* **34**(13), 2295–2296.

Green, E. R. & Mecsas, J. (2016), 'Bacterial secretion systems: An overview', *Microbiology Spectrum* **4**(1).

Heijne, G. & Gavel, Y. (1988), 'Topogenic signals in integral membrane proteins', *European Journal of Biochemistry* **174**(4), 671–678.

Hemmerich, J., Rohe, P., Kleine, B., Jurischka, S., Wiechert, W., Freudl, R. & Oldiges, M. (2016), 'Use of a sec signal peptide library from bacillus subtilis for the optimization of cutinase secretion in corynebacterium glutamicum', *Microbial Cell Factories* **15**(1).

Janda, C. Y., Li, J., Oubridge, C., Hernández, H., Robinson, C. V. & Nagai, K. (2010), 'Recognition of a signal peptide by the signal recognition particle', *Nature* **465**(7297), 507–510.

Moore, S. J., Lai, H.-E., Kelwick, R. J., Chee, S. M., Bell, D. J., Polizzi, K. M. & Freemont, P. S. (2016), 'Ecoflex: A multifunctional moclo kit for e.coli synthetic biology', *ACS Synthetic Biology* **5**(10), 1059–1069.

Nilsson, I., Lara, P., Hessa, T., Johnson, A. E., von Heijne, G. & Karamyshev, A. L. (2015), 'The code for directing proteins for translocation across er membrane: Srp cotranslationally recognizes specific features of a signal sequence', *Journal of Molecular Biology* **427**(6), 1191–1201.

Old, R. W. & Primrose, S. B. (2001), *Principles of gene manipulation: An introduction to genetic engineering*, Blackwell Science.

Prabudiansyah, I. & Driessen, A. J. (2016), 'The canonical and accessory sec system of gram-positive bacteria', *Current Topics in Microbiology and Immunology* p. 45–67.

Uniprot (2022), 'Uniprot'.
URL: *https://www.uniprot.org/*

Weber, E., Engler, C., Gruetzner, R., Werner, S. & Marillonnet, S. (2011), 'A modular cloning system for standardized assembly of multigene constructs', *PLoS ONE* **6**(2).

Zimmermann, R. (2009), *Protein transport into the endoplasmic reticulum*, Landes Bioscience.