American Journal of Biochemistry and Biotechnology

Research Notes

An Artificial Design Technique to Optimize Signal Peptide

[1]Gao Cui-Fang, [1]Wang Sen, [2]Tian Feng-Wei, [1]Zhu Ping and [2]Chen Wei

[1]School of Science, Jiangnan University, Wuxi 214122, China
[2]School of Food Science and Technology, Jiangnan University, Wuxi 214122 China

Article history
Received: 08-06-2017
Revised: 16-06-2017
Accepted: 19-07-2017

Corresponding Author:
Gao Cui-Fang
School of Science, Jiangnan
University, Wuxi 214122,
China
Email: cuifang_gao@163.com

Abstract: To determine optimal artificial signal peptide candidates for the possibility of creating high levels of secretion of heterologous proteins, substitution and redesign of amino acid sequences in the H-domain of the signal peptide was theoretically attempted. The method was based on comprehensive score matrix and Markov transfer matrix, which can make the artificial sequences maintain the structural characteristics and original polarity of signal peptides. For the artificial sequence, the feature vector of Structural Fusion Degree (SFD) is first extracted to quantitatively describe the compatibility of artificial cleaved region, then by comparing with highly secreted natural samples; tendencies of specific substitutions in the amino acid sequence can be identified at certain locations. These substitutions may represent the key amino acids that influence the secretion and expression levels of heterologous proteins.

Keywords: Markov Transition Matrix, Signal Peptide, Feature Vector, Artificial Sequence

Introduction

Proteins that can be exported to other cellular sites from their site of synthesis by traversing the cytoplasmic membrane are generally referred to as secreted proteins. Successful secretion of proteins depends on the presence of a signal peptide, which is generally located at the N-terminus of the amino acid chain and is composed of 15 to 60 amino acids (Nielsen *et al*., 2011). Under the direction of a signal peptide, the synthesized protein is transported through the protein channel and secreted to a targeted destination, following which the signal peptide is cleaved by specific signal peptidases to form the mature protein. It remains a great challenge to industrially synthesize different kinds of poorly secreted natural proteins in organisms.

Should an identifiable artificial signal peptide be designed using bioengineering technology, thus making proteins more highly able to be directly secreted into the culture medium, it will require approaches that exceed the properties of the natural protein resource. A bioengineering approach will only substitute or artificially design the signal peptide in specific host bacteria to guide the heterologous protein as one that is secreted. More importantly, it can maintain an unaltered mature protein sequence and will not have any effect on the biological functions of the synthesized protein. Therefore studies aiming at artificial signal peptide will contribute important technological advances in the industrial production of important natural proteins (Cai *et al*., 2016; Pournejati *et al*., 2014; Romána *et al*., 2014). One important factor that should be taken into account is that the main chain of the protein must retain that found for natural protein after the original signal peptide of the heterologous protein is replace by an artificially synthesized signal peptide. Thus, there must be a high degree of similarity between the artificial sequence and the original sequence, but the secretion and expression levels might be considerably different. Thus, it presents a great challenge to analyze highly similar sequences, including the degree of compatability between the artificial signal peptide and the main chain and some important amino acids that will significantly affect the secretion of the protein and the design of appropriate artificial signal peptides.

According to previous work done in Bacillus subtilis as the host bacteria, some heterologous samples with the artificial signal peptide have successfully achieved high levels of secretion and expression. However, others are poorly secreted. For example, the sequence of Bacillus licheniformis α-amylase (AMY_BACLI) consists of 512 amino acid residues (29 residues are present within the signal peptide), when its original signal peptide is replace by signal peptide SacB (SACB_BACSU), alpha-amylase is non- or poorly secreted. By contrast, when

Science Publications

replace by the signal peptide AprE (SUBT_BACSU), the protein achieves a higher level of secretion (Sloma *et al*., 1988). It should be pointed out that the natural protein SacB and AprE both show high levels of secretion in Bacillus subtilis. Clearly, the heterologous protein in the host bacterial strain Bacillus subtilis can achieve high levels of secretion and expression. Thus, the possible reason might be as a consequence of the mature protein of Bacillus licheniformis α-amylase exhibiting no compatibility with the artificial signal peptide SacB. Such results inform us that the optimal design for the non/poorly secreted signal peptide should take into account the property of compatibility of the cleaved region.

Previous studies have shown that sometimes just a few key amino acids in the signal peptide affect the level of secretion of heterologous proteins, which are significantly different if replacing 2 ~ 3 or even one amino acid residue in the signal peptide sequence (Nijland *et al*., 2007). Thus, it is highly and theoretically possible, to increase the secretion levels of heterologous proteins if the signal peptide sequence is adjusted or somewhat redesigned.

With the rapid development of computational technology, many intelligent algorithms have been developed and applied to the prediction of the signal peptide (Zhang and Wood, 2003; Gao *et al*., 2013; Zheng *et al*., 2012; Tsirigos *et al*., 2015; Zhang *et al*., 2014), such as the Neural Network (NN) (Nielsen *et al*., 2011), the Hidden Markov (HMM) method and the signal-BNF method (Zheng *et al*., 2012) etc. These methods mainly focus on natural protein sequences and there is currently no artificial sample that has undergone replacement or design of a new signal peptide. One research (Gao *et al*., 2010) proposed a Structural Fusion Degree (SFD) feature extraction method and established a mathematical model that took into consideration the signal peptide that was fused into the targeted region of the heterologous protein. The feature vector extracted from the mathematical model could be used to distinguish and characterize the ability of the artificially synthesized proteins to be secreted.

In this research, aiming at designing signal peptides in Gram-positive bacteria, we have developed an optimized design strategy and technique for creating artificial signal peptides based on the characteristics of the Structural Fusion Degree (SFD). By studying the substitution principle and the metastatic pattern of amino acids, we actively redesigned and optimized signal peptide sequences that were otherwise unable to be secreted or were inefficiently secreted. We studied and identified the amino acid assignment trends present on different positions of the signal peptide, with the aim of finding the optimal signal peptide candidate, which could be applied to achieve high levels of secretion and expression of the targeted heterologous protein.

Materials and Methods

In the case of not knowing the key amino acid positions, it is unfeasible to attempt all possible replacement options, even with the use of available computer tools. From a theoretical viewpoint of the biological functions of signal peptide and the characteristics of each amino acid, we will design and analyze the artificial sequence from the following steps: (i) Construct a reasonable comprehensive substitution matrix of the amino acid, (ii) Build a general Markov transition frequency matrix, (iii) Design the artificial sequence according to the above defined matrices, (iv) Extract SFD features of the artificial sequence in an attempt to quantitatively describe the compatibility information and (v) Compare similarity with samples exhibited high levels of secretion in an attempt to determine the sequence of the candidate exhibiting high levels of secretion.

In this paper, we have attempted to adjust/replace partial amino acids of the signal peptide SacB in the permissibility range and connected the artificial signal sequence to the main chain of Bacillus licheniformis α-amylase. By a series of intelligent analyses, we wished to find the amino acid assignment trends of different positions in the signal peptide. It is worthwhile realizing that the optimized design technique will be the same for other signal peptides, depending of course on the different targeted protein.

Construct Comprehensive Score Matrix

The rule of amino acid substitutions in the evolutionary process remains unclear and as a consequence, the determination method of the key amino acid cannot easily be given. However, the signal peptide, as a special segment of protein sequences, possesses a key biological function, which is to guide the target protein and assist its transportation through the protein channel. Accordingly, only if the artificial sequence persists the same characteristic structure and polarities as the natural signal peptide will it be possible to possess its biological function.

BLOSUM 62 matrix (refer to Appendix 1) and hydrophobic matrix (refer to Appendix 2) are frequently-used score matrices in the sequence alignment of protein. The BLOSUM 62 matrix is a statistical pattern based on a likelihood method by estimating the occurrence of each possible pair wise substitution from blocks database. Those pair wise with high score are so called ‘conservative substitution’ in the evolution and such substitution has higher probability to maintain the protein function than ‘random substitution’. Hydrophobic matrix presents the similarity between amino acids from another viewpoint, in which the substitution with high score will cause a small change in hydrophobicity. H-domain is the functional region of signal peptide which primarily consist of hydrophobic amino acids, therefore substitution based on this matrix advantageously persists the characteristic structure of a signal peptide.

So we constructed a comprehensive score matrix based on the amino acid Blosum 62 substitution matrix and Hydrophobic matrix. Firstly, the matrix with different measurement must be standardized to conform to the unified norm. Standardized methods of the Blosum 62 matrix and the hydrophobic matrix are designed according to the following Equation 1:

$$y_{kh} = \frac{x_{kh} - \frac{\sum_{k=1}^{N} x_{kh}}{N}}{\max_k(x_{kh}) - \min_k(x_{kh})}. \tag{1}$$

In which x_{kh} (k = 1,...,20; h = 1,...,20) is the original data and y_{kh} (k = 1,...,20; h = 1,..., 20) is the subsequent standardized data. Then the substitution score can be calculated based on the standardized matrix.

We define the expression of score function as Equation 2, which can indicate the proportion in the different matrix for each substitution amino acid. The hypotheses of the method is that the 'conservative substitution' and "persists hydrophobicity structure" are of equal importance, then we set $w_1 = 0.5$ and $w_2 = 0.5$ in our research. In fact, for different species of protein sequences, w_1 and w_2 may be set at different weight values. For example, signal peptides from Gram-bacteria are not so much various as that from Gram+ bacteria, in other words, they are more conservative, in this case the Blosum 62 matrix can be a little more important. Then in Gram- bacteria, the specific gravity of Blosum 62 matrix can be 60% ($w_1 = 0.6$) and the gravity of hydrophobic matrix can be 40% ($w_2 = 0.4$):

$$f_{ij} = w_1 a_{ij} + w_2 b_{ij} \ . \tag{2}$$

where, a_{ij}, b_{ij} respectively represent the elements in the i-th row and the j-th column of the standardized Blosum 62 matrix and hydrophobic matrix, accordingly, f_{ij} represent the elements in the comprehensive score matrix. We obtain a substitution score matrix as in Table 1.

Construct the General Markova Matrix

Abundant natural signal peptides from one species as a colony generally contain a disciplined pattern of amino acids, including discrepancies, transfer and assignment order and so forth. Under the direction of such patterns, we can adequately utilize prior knowledge to design reasonable artificial signal peptides. Markov chain is a widely applied mathematic model that reveals the collection of state distributions on a peptide sequence. Typically, the signal peptides are described using limited symbols to denote 20 kinds of natural amino acids. Should these residues on the chain be regarded as state parameters, it follows that the sequences of the amino acids will express a series of transition states. In this way, a finite stationary Markov model can be constructed based on symbol distribution to reflect the intrinsic relationship and further to detect the comprehensive information of signal peptide sequence.

Let Θ be a set of complete amino acid symbols in alphabetical order: Θ = {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}, which can be used as the state set. Given a signal peptide sequence containing n amino acid residues: $Q = \{R_1R_2R_3R_4R_5R_6R_7...R_n\}$, where R_i (i = 1, 2, 3…n) denotes one of the residue in set Θ. In order to quantitatively describe the transition behavior state on Q, we defined a 20×20 Markov matrix whose rows and columns were denoted by amino acids to represent the frequency of occurrence of each dipeptide. Assume that $M(i, j) = \{(R_i, R_j), z\}$, That is to say in the frequency matrix M, the element value in the i-th row (denoted by amino acid R_i), j-th column (denoted by amino acid R_j) is numerical z. Where in R_i is the previous residue of a dipeptide and R_j is the latter, z is the transition frequency from R_i to R_j through the full sequence. Thus, we find that the pair-wise residues (R_i,R_j) in the matrix M correspond to their respective denotations and give the assignment $M(i, j) = z$. Thus, the Markov matrix that reflects the composition of the dipeptide and the series of state relations in sequence Q can be obtained.

The general Markov transfer frequency matrix can be constructed if the metastatic behaviors of a large number of signal peptide sequences have had statistical measurements made. Since *Bacillus subtilis* as the host bacteria belongs to the Gram-positive class of bacteria, we thus chose a 140 signal peptide dataset of the secreted protein sequence of Gram positive bacteria in the benchmark dataset (http://www.cbs.dtu.dk/ftp/signalp). The following Table 2 shows the calculated general Markov frequency matrix. So the general characteristic (similarity) among all sequences in the set can be reflected by the statistical value in the matrix. For example, the value 0 appears in the row "K" column "C", which suggests the inexistence that Cysteine followed Lysine in the set. Then according the similarity in Table 2, the reasonable artificial sequences should exclude such occurrence of '…KD…'

Artificial Sequence Designation

Most signal peptides consist of three functional domains (Fan *et al*., 2013): A positively charged N-terminal (N-domain) which is called the alkaline amino terminal; a hydrophobic segment (H-domain) which mainly contains neutral amino acids, can form a section of α-helical structures and is generally viewed as the major functional domain; and a long negatively charged C-terminal (C-domain) which is comprised mainly of small molecule amino acids, is the "cutting area" of the signal peptide, often referred to as the processing zone. Here the major functional H-domain is selected and will be redesigned.

Table 1. Comprehensive score matrix for amino acid substitutions

	R	K	D	E	S	N	Q	G	T	H	A	C	M	P	V	L	I	Y	F	W
R	0.63	0.45	0.15	0.27	0.06	0.12	0.18	-0.05	0.01	0.07	0.01	-0.17	-0.09	-0.15	-0.22	-0.15	-0.22	-0.20	-0.32	-0.37
K	0.43	0.62	0.20	0.32	0.11	0.11	0.17	-0.07	-0.00	-0.00	-0.00	-0.18	-0.10	-0.10	-0.17	-0.17	-0.23	-0.22	-0.33	-0.38
D	0.17	0.22	0.62	0.42	0.16	0.15	0.10	0.05	-0.00	-0.00	-0.05	-0.10	-0.16	-0.06	-0.16	-0.27	-0.22	-0.22	-0.27	-0.38
E	0.25	0.31	0.42	0.58	0.14	0.08	0.19	-0.03	-0.03	0.03	-0.03	-0.19	-0.14	-0.08	-0.14	-0.25	-0.25	-0.19	-0.31	-0.36
S	-0.17	-0.09	-0.01	-0.01	0.52	0.31	0.24	0.24	0.23	0.08	0.23	0.00	0.00	-0.08	-0.15	-0.15	-0.15	-0.24	-0.24	-0.48
N	-0.09	-0.09	-0.04	-0.09	0.30	0.55	0.25	0.25	0.16	0.21	0.06	-0.07	-0.02	-0.02	-0.15	-0.15	-0.15	-0.19	-0.24	-0.45
Q	-0.03	-0.03	-0.10	0.03	0.24	0.24	0.55	0.11	0.09	0.15	0.09	-0.12	0.07	0.01	-0.14	-0.14	-0.20	-0.16	-0.28	-0.39
G	-0.27	-0.27	-0.12	-0.17	0.33	0.33	0.23	0.63	0.13	0.13	0.23	-0.02	-0.02	0.03	-0.02	-0.17	-0.17	-0.22	-0.22	-0.27
T	-0.30	-0.30	-0.30	-0.30	0.25	0.17	0.10	0.03	0.63	0.13	0.28	0.10	0.10	0.00	0.07	0.00	0.00	-0.17	-0.17	-0.37
H	-0.24	-0.29	-0.29	-0.24	0.11	0.20	0.16	0.07	0.17	0.62	0.17	0.02	0.07	0.07	-0.08	-0.08	-0.08	0.05	-0.09	-0.33
A	-0.29	-0.29	-0.37	-0.29	0.25	0.03	0.11	0.18	0.28	0.13	0.56	0.18	0.11	0.11	0.08	0.01	0.01	-0.17	-0.17	-0.44
C	-0.37	-0.37	-0.28	-0.32	0.04	-0.03	-0.03	-0.03	0.13	0.05	0.16	0.59	0.21	0.05	0.13	0.13	0.13	0.00	0.00	-0.25
M	-0.36	-0.36	-0.41	-0.35	-0.00	-0.06	0.06	-0.13	0.07	0.01	0.07	0.14	0.52	0.08	0.27	0.26	0.20	-0.00	0.06	-0.07
P	-0.35	-0.30	-0.23	-0.23	-0.02	0.01	0.06	0.01	0.06	0.08	0.13	0.04	0.15	0.56	0.15	0.04	0.04	0.04	-0.08	-0.15
V	-0.46	-0.39	-0.39	-0.31	-0.10	-0.17	-0.10	-0.10	0.11	-0.10	0.11	0.11	0.33	0.11	0.54	0.33	0.47	0.11	0.04	-0.17
L	-0.36	-0.36	-0.48	-0.42	-0.07	-0.13	-0.07	-0.20	0.06	-0.06	0.06	0.13	0.32	0.01	0.33	0.52	0.39	0.13	0.20	-0.00
I	-0.41	-0.41	-0.41	-0.41	-0.13	-0.13	-0.13	-0.19	0.07	-0.06	0.07	0.14	0.26	0.01	0.46	0.40	0.52	0.14	0.20	-0.06
Y	-0.36	-0.36	-0.34	-0.29	-0.11	-0.11	-0.06	-0.16	-0.04	0.16	-0.04	0.02	0.07	0.03	0.13	0.13	0.13	0.59	0.39	0.22
F	-0.39	-0.39	-0.34	-0.34	-0.07	-0.12	-0.12	-0.12	-0.01	0.04	-0.01	0.05	0.15	-0.05	0.10	0.20	0.20	0.41	0.56	0.25
W	-0.30	-0.30	-0.29	-0.25	-0.10	-0.14	-0.07	-0.02	-0.02	-0.02	-0.05	0.03	0.11	0.01	0.05	0.13	0.10	0.26	0.28	0.66

Table 2. Markov transfer frequency matrix of Gram-positive bacteria

	R	K	D	E	S	N	Q	G	T	H	A	C	M	P	V	L	I	Y	F	W
R	36	28	3	2	12	4	1	8	13	4	24	2	3	5	11	17	9	4	13	3
K	31	65	0	5	12	17	12	11	18	6	25	0	9	3	22	24	23	3	12	4
D	3	6	1	2	1	2	0	2	3	0	5	0	0	0	4	3	2	1	2	0
E	2	6	0	0	3	3	2	1	4	0	13	0	1	3	6	1	3	1	2	0
S	17	12	2	2	25	11	6	16	28	2	51	4	9	13	36	72	34	2	13	1
N	8	15	4	2	8	6	3	2	15	1	15	1	3	7	8	8	6	3	4	1
Q	3	5	4	2	7	5	8	1	6	3	26	0	2	5	2	0	4	2	1	0
G	7	5	0	2	19	6	4	16	24	1	62	1	8	8	27	51	14	1	9	1
T	8	9	5	5	25	5	7	16	23	0	90	1	4	11	28	46	14	4	13	2
H	2	2	1	1	4	1	2	0	5	0	12	0	1	2	1	3	0	1	2	1
A	15	19	6	7	58	13	13	58	48	10	111	4	8	25	59	90	23	3	31	4
C	0	2	0	0	4	0	0	3	0	0	2	2	2	2	5	9	5	0	3	0
M	20	55	1	7	22	17	6	5	12	4	22	1	3	7	11	27	11	3	11	0
P	7	3	1	3	13	6	2	6	12	1	32	0	7	7	16	21	6	4	7	0
V	10	15	4	4	37	6	7	36	18	2	52	5	15	15	35	57	26	7	20	1
L	14	16	1	3	62	8	4	53	50	2	101	9	16	28	59	98	34	5	37	3
I	13	18	0	2	26	5	4	21	22	1	32	6	7	7	21	30	21	1	19	0
Y	1	4	2	2	9	4	0	4	0	1	8	1	1	1	1	4	0	0	2	0
F	8	18	1	2	16	2	6	11	15	3	30	2	5	5	18	37	19	1	10	3
W	0	1	1	0	1	0	0	1	1	0	5	0	1	0	5	5	2	0	1	0

We first ascertained the distribution range of three domains of natural signal peptide SacB using signal P 3.0-HMM (http://www.cbs.dtu.dk/services/SignalP/). According to the online analysis of the results, the H-domain is located in the position 11-22 and then these 12 amino acid residues will be redesigned artificially. We selected the threshold $f \geq 0.28$ in comprehensive score matrix and $f \geq 12$ in Markov transition matrix. The feasible substituted amino acids in each position were filtered and shown in Table 3.

According to Table 3, there are several cadidate amino acids in positions 12, 13, 19, 20, 22. Thus 432 (3×4×3×4×3) artificial signal peptide sequences can be obtained according to these substitutions. Suppose the amino acid V replaced by V in positions 12, but original amino acids changed in other positions, then a new different sequence can be obtained. Only when all of the amino acids in 5 positions replaced by themselves, the original sequence can be obtained. That means there is just one orginal sequence in the 432 artificial sequences. In this way, we have significantly reduced the number of candidate signal peptide sequences. Thus the intelligent analysis and identification of key amino acids will be possible by numerical experiments done with the aid of computer programs.

Extract Numerical Features

From a mathematical viewpoint, the interaction between the artificial signal peptide and neighboring residues in the cleaved region were analyzed.

Table 3. The amino acid can be replaced in the H-domain of signal peptide SacB

Position	11	12	13	14	15	16	17	18	19	20	21	22
Original amino acids	T	V	L	T	F	T	T	A	L	L	A	G
Replaced amino acids	T	VLI	MV L I	T	F	T	T	A	VLI	MV L I	A	SNG

By considering a mathematical approach (Gao *et al*., 2010), suppose the sequence length of the artificial signal peptide is l and suppose we extend the sequence of the signal peptide by adding 15 additional amino acids from its nearest downstream neighbors (as is shown in Fig. 1. Accordingly, the length of the extended signal peptide is l+15, which means it contains 15 adjacent amino acids in the chain.

Then the information set of the extended signal peptide was constructed, which contained all the sub-sequences of the signal peptide fragment. When the extending length is 15, the sub-sequence distribution set is: $\Omega = (U^1\ U^2\ \ldots\ U^{15}\ U^{16})$.

Where:

$$U^1 = \{R_1R_2 \ldots R_l\};$$
$$U^2 = \{R_1R_2 \ldots R_lR_{l+1}\};$$
$$\ldots$$
$$U^k = \{R_1R_2 \ldots R_lR_{l+1} \ldots R_{l+k-1}\};$$
$$\ldots$$
$$U^{15} = \{R_1R_2 \ldots R_lR_{l+1} \ldots R_{l+14}\};$$
$$U^{16} = \{R_1R_2 \ldots R_lR_{l+1} \ldots R_{l+14}R_{l+15}\}$$

where, $R_i(i = 1,2,3\ldots l + 15)$ represents one of the 20 natural amino acids, obviously, U^1 simply represent the signal peptide and U^{16} is the extended signal peptide. Each sequence will contain one additional residue than the former and such elongation might contain discrepancies and interactions among this subsequence. For each sub-sequence in set Ω, a 20-dimensional amino acid component feature vector can be extracted and a total of 16 feature vectors can be obtained. All of these vectors together form a matrix of extended signal peptide, which is denoted as $A = [V_1\ V_2\ \ldots\ V_{16}]$.

Where $V_1 = [v_{1,1}\ v_{1,2}\ \ldots\ v_{1,20}]^T$ is the feature vector of subsequence U^1, $V_2 = [v_{2,1}\ v_{2,2}\ \ldots\ v_{2,20}]^T$ is the feature vector of subsequence U^2 and so on.

There is some overlap between the subsequence in set Ω, so that the related analysis of matrix A is described using different variable covariance. Assume that C is the covariance matrix:

$$C = \begin{bmatrix} c_{1,1} & c_{2,1} & \cdots & c_{20,1} \\ c_{1,2} & c_{2,2} & \cdots & c_{20,2} \\ \cdots & & & \\ c_{1,20} & c_{2,20} & \cdots & c_{20,20} \end{bmatrix}$$

Matrix C is symmetrical, where the element in the position of the subscript (i, j) is the covariance between the row vectors of the ith component and jth component rows in matrix A. For the convenience of computing and the need to not to lose any of the information contained in the covariance matrix, a substitution matrix D consisting of the eigenvectors of matrix C is used to formulate the relationship $DX = B$. Where B is the feature vectors of the entire protein chain. The unknown vector $X = [x_1\ x_2\ \ldots\ x_{20}]'$ represents the requisite features of SFD.

If D^{-1} exists, then the solution vector is $X = D^{-1}B$, otherwise least squares method can be used to obtain the solutions. Therefore, a one-to-one corresponding feature vector between an extended signal and a protein chain can be obtained. These extracted numerical features contain local features and integrated information of the cleaved region, on which the subsequent intelligent analysis of the artificial sequences can be performed.

Numerical Experiments and Results Analysis

We respectively connected signal peptide sequence with the main chain of Bacillus licheniformis α-amylase to derive artificial samples. Next, we extracted the numerical SFD features by the method introduced in materials and methods above and finally we used these numerical vectors to analyze and find the amino acid assignment trend in different positions.

Similarity Analysis of Artificial Sequences

The method needs a reference criterion to evaluate the possible level of secretion of artificial sequences, which is the mean center of all high secreted proteins in the literature (Gao *et al*., 2010). We calculated the similarity distance between the artificial sample and the high secretory protein using kernel-induced metric as shown in Equation 3 (Zhang and Chen, 2003):

$$d(\mathrm{x},\mathrm{y}) = \|\varphi(\mathrm{x}) - \varphi(\mathrm{y})\| = \sqrt{2(1 - K(\mathrm{x},\mathrm{y}))} \qquad (3)$$

In Equation 3, suppose x indicates the numerical SFD feature of artificial sequence, y indicates the mean center of high secretory proteins. The smaller the distance $d(x,y)$ is, the more similar between x and y and the higher possibility of the artificial sequence with high level of secretion.

The function ϕ: $p \in OS \to \phi(p) \in HS$ is a continuous smooth nonlinear mapping function, by which the difference among samples can be extended in the mapped space. Where p denotes an element in the input data space OS and $\phi(p)$ is the corresponding element in the high-dimensional mapped space HS. Here, the most commonly used Gaussian kernel function was adopted:

$$K(\mathrm{x},\mathrm{y}) = \exp\left(\frac{-\|\mathrm{x}-\mathrm{y}\|^2}{\sigma^2}\right) \qquad (4)$$

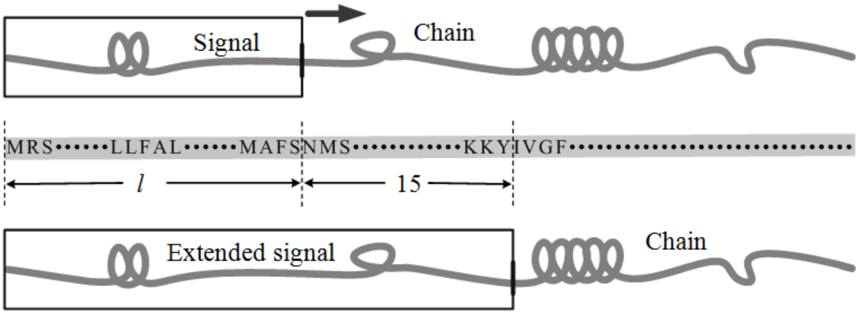

Signal
Chain
MRS••••••LLFAL••••••MAFSNMS•••••••••••KKYIVGF••••••••••••••••••••••••••••
l
15
Extended signal
Chain

Fig. 1. Extended signal peptide which contains some near downstream neighbors

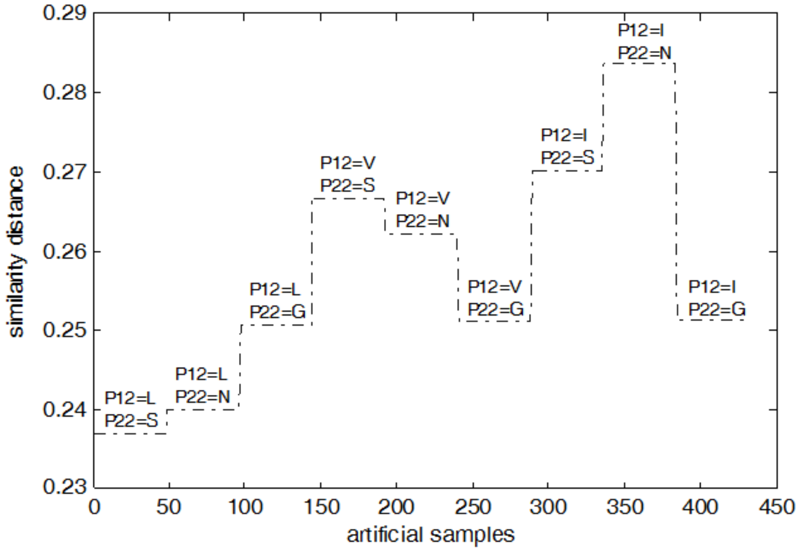


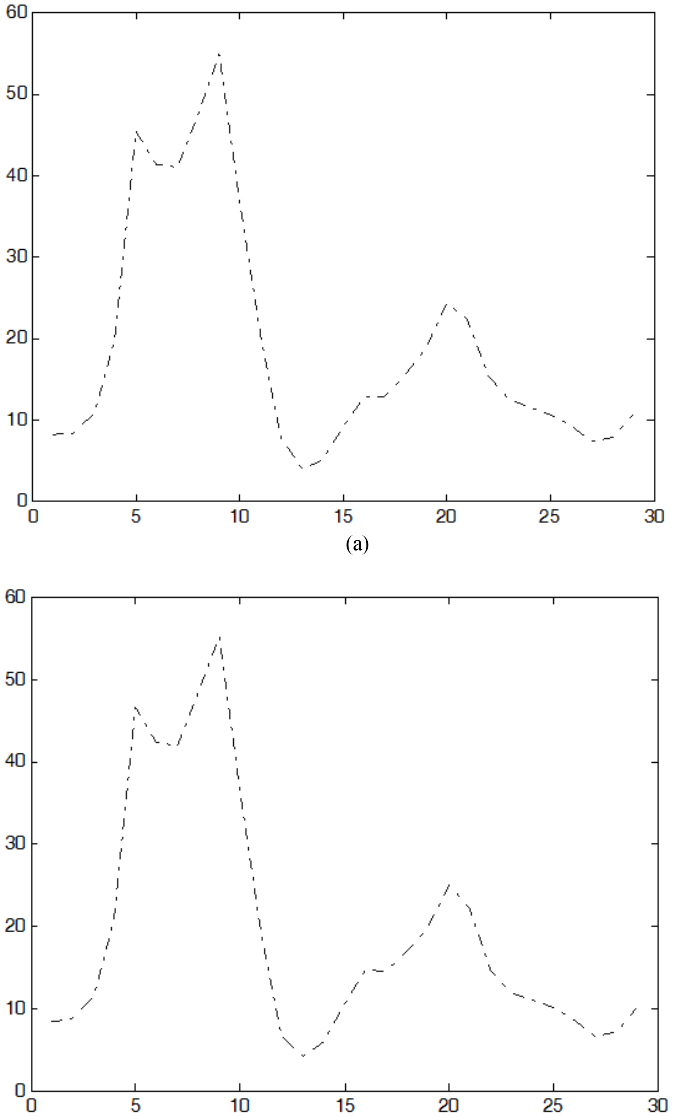
Fig. 2. The similarity of the artificial samples with secreted protein Center

Those unknown samples with smaller distance values will have a high possibility of achieving a high level of secretion. According to the values of distance, we found some sequences with small distances and analyzed their sequence structure. Finally we found that some amino acids have obvious biased assignment trends in different positions. For example, the biased assignment in position 12 is L (leucine) and the biased assignment in position 22 is S (serine) and N (asparagine amide), especially in position 12. The unknown samples and the secreted center, have high similarity with the substitution amino acid L. Such results suggested that the above two positions might represent the key amino acid location. Thus we subsequently substituted the original amino acid with the biased amino acids in these two positions and obtained the artificial sequence SacB-2, then further analyzed the structural characteristic of SacB-2.

Structure Analysis of Artificial Sequences

Wavelet transform is a type of time-frequency analysis method for signals that have been viewed as a "Mathematical microscope", which can provide information of the protein structure which itself is obtained from the wavelet coefficients that can be used to analyze and estimate the H-domain of signal peptides (Li *et al*., 2008). We performed one-dimensional continuous wavelet decomposition for the signal peptide sequences using db2 filter in scale (1:30) and obtained the structural information as shown in Fig. 3.

As the initial segment of a protein sequence, the signal peptide has a certain structure. Therefore, the artificial sequence after redesigning should also maintain the peculiar structure as a signal peptide. As can be seen from the results Fig. 3, the structure of the artificial sequence SacB-2 and the natural high secretion signal peptide SacB are almost consistent. This means that there will be a high probability for SacB-2 to be compatible with the transfer channel of *Bacillus subtilis.* Simultaneously, according to the results of similarity analysis based on the Structural Fusion Degree (SFD), SacB-2 is also compatible with the main chain of *Bacillus licheniformis* α-amylase so that it is likely to achieve both high secretion and expression of the targeted or chosen heterologous proteins.



(a)

(b)

Fig. 3. (a) Natural signal peptide SacB (b) Artificial signal peptide SacB-2 Signal peptide structure obtained by wavelet transform

Table 4. Prediction about the artificial sequences by Signal P 3.0

	P_{12} = L	P_{12} = L	P_{12} = L	P_{12} = V	P_{12} = V	P_{12} = V	P_{12} = I	P_{12} = I	P_{12} = I
Substitutions	P_{22} = S	P_{22} = N	P_{22} = G	P_{22} = S	P_{22} = N	P_{22} = G	P_{22} = S	P_{22} = N	P_{22} = G
Signal peptide probability	0.995	0.990	0.983	0.993	0.985	0.975	0.991	0.980	0.968

Analysis by Successful Software Signal P 3.0

As the currently most popular prediction method for secreted proteins, Signal P 3.0 (Bendtsen *et al*., 2004) has been benchmarked against other available methods and performs significantly better than most prediction schemes. Therefor we use Signal P 3.0 to justify our artificial sequences with substitutions in Fig. 2, which are the biased assignment in position 12 and position 22. The software analyzes the input data (such as artificial sequence: Mnikkfakqatlltfttallasgatqafa) based on hidden Markov models from Gram-positive prokaryotes and then output the signal peptide probability about the input sequence. All the artificial sequences in Fig. 2 were input and the prediction results as Table 4.

The analyses from Signal P 3.0 suggest that these artificial sequences have very high possibility to be signal peptide. Especially, when the substitution in position 12 is L (leucine) and in position 22 is S (serine), the Signal peptide probability is up to 0.995. In short, it seems that some biased assignment exist in the two positions, which is also in line with the results of Fig. 2.

Conclusion

In this research, the H-domain of signal peptide sequences have theoretically redesigned and some key amino acids are determined, located in different positions that have displayed biased assignments. Signal peptide candidates have also been identified that have shown a high degree of possibility to exhibit high levels of secretion and expression of heterologous proteins. This provides a conceptual and theoretical framework that can guide subsequent trials for more efficient biological secretion and expression studies. Without evaluating the key amino acid positions, it is unfeasible to attempt biological experiment, because that all the possible replacement options are enormous. For example, when redesigning the sequence 'TVLTFTTALLAG', there are 20 replacement options for each position and there will be 2012 candidate sequences! It is impossible for biological experiment, therefore most of the sequences should be excluded by the evaluation method in advance.

In addition, it deserved to be mentioned that the comprehensive score matrix and the general Markov transition matrix allow for the artificial sequence to possess the same characteristic structure and polarities as the natural signal peptides and the extracted SFD feature vector can distinguish and characterize the compatibility and similarity of artificial cleaved region. The method based on the 140 signal peptides dataset can get a statistical measurement, at the same time it used the mean center of high secreted proteins as the criterion of evaluation. All of this prior knowledge enables the method to design reasonable artificial signal of SacB, even more the method is suitable for the design of other signal peptides.

Obviously, there are many methods for the optimization of the design of signal peptides, in addition to substituting amino acids in the fixed position, we can also insert or delete several amino acid residues in the signal peptide sequence. Moreover, the amino acid substitution might not be limited to the H-domain, the key amino acid affecting heterologous protein secretion might also be present in other regions. In the future we aim to further broaden the dynamic design range of optimized signal peptides by combining with relevant biological knowledge of the targeted protein.

Acknowledgement

The authors are grateful to National Natural Science Foundation of China (Grant No.: 61402202, 11271163, 31371721), China Postdoctoral Science Foundation (Grant No.: 2015M581724) and Postdoctoral Science Foundation of Jiangsu Province (Grant No.: 1401099C) for financial support.

Author's Contributions

Gao Cui-Fang: Analyzed the data and wrote the paper.

Wang Sen: Performed the numerical experiments.

Tian Feng-Wei: Designed and developed the method.

Zhu Ping: Revised the manuscript.

Chen Wei: Conceived the study.

Ethics

The authors declare their responsibility for any ethical issues that may arise after the publication of this manuscript.

References

Bendtsen, J.D., H. Nielsen, G.V. Heijne and S. Brunak, 2004. Improved prediction of signal peptides: Signal P 3.0. J. Mol. Biol., 340: 783-795.

Cai, D., X. Wei, Y. Qiu, Y. Chen and J. Chen *et al.*, 2016. High-level expression of nattokinase in Bacillus licheniformis by manipulating signal peptide and signal peptidase. J. Applied Microbiol., 121: 704-712.

Fan, Y.X., J.N. Song, C. Xu and H.B. Shen, 2013. Predicting protein N-terminal signal peptides using position-specific amino acid propensities and conditional random fields. Curr. Bioinform., 8: 183-192.

Gao, C., Q. Guan, H. Zhang, W. Chen and F. Tian, 2013. A novel feature extraction method by compressive sensing for signal peptide. J. Chem. Pharm. Res., 5: 212-218.

Gao, C.F., X.J. Wu, F.W. Tian, Y. Xia and W. Chen, 2010. Characterization of protein secretion based on structural fusion degree. Chin. J. Biotech., 26: 687-695.

Li, Y., Z. Wen, C. Zhou, F. Tan and M. Li, 2008. Effects of neighboring sequence environment in predicting cleavage sites of signal peptides. Peptides, 29: 1498-1504.

Nielsen, H., J. Engelbrecht, S. Brunak and G.V. Heijne, 2011. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int. J. Neural. Syst., 8: 581-599.

Nijland, R., R. Heerlien, L.W. Hamoen and O.P. Kuipers, 2007. Changing a single amino acid in Clostridium perfringens β-toxin affects the efficiency of heterologous secretion by Bacillus subtilis. Applied Environ. Microb., 73: 1586-1593.

Pournejati, R., H.R. Karbalaeiheidari and N. Budisa, 2014. Secretion of recombinant archeal lipase mediated by SVP2 signal peptide in Escherichia coli and its optimization by response surface methodology. Protein Expres. Purif., 101: 84-90.

Romána, R., J. Mireta, F. Scaliab, A. Casablancasa and M. Lecinaa *et al.*, 2016. Enhancing heterologous protein expression and secretion in HEK293 cells by means of combination of CMV promoter and IFNα2 signal peptide. J. Biotechnol., 239: 57-60.

Sloma, A., D. Pawlyk and J. Pero, 1988. Development of an Expression and Secretion System in Bacillus Subtilis Utilizing SACQ: Genetics and Biotechnolog of Bacilli. 1st Edn., Academic Press, San Diego.

Tsirigos, K.D., C. Peters, N. Shu, L. Käll and A. Elofsson, 2015. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res., 43: 401-407.

Zhang, D.Q. and S.C. Chen, 2003. Clustering incomplete data using kernel based fuzzy c-means algorithm. Neural Process. Lett., 18: 155-162.

Zhang, S.W., T.H. Zhang, J.N. Zhang and Y. Huang, 2014. Prediction of signal peptide cleavage sites with subsite-coupled and template matching fusion algorithm. Mol. Inform., 33: 230-239.

Zhang, Z. and W.I. Wood, 2003. A profile hidden Markov model for signal peptides generated by HMMER. Bioinformatics, 19: 307-308.

Zheng, Z., Y. Chen, L. Chen, G. Guo and Y. Fan *et al*., 2012. Signal-BNF: A bayesian network fusing approach to predict signal peptides. Biomed Res. Int., 16: 492174.

Appendix 1

Appendix table 1 Blosum 62 amino acid substitution matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	5	-1	0	-2	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	0	-2	-2	-2
P	-3	-1	-1	7	-1	-2	-2	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	0	-1	4	0	-2	-2	-1	-1	-2	-1	-1	-1	-1	-1	0	-2	-2	-3
G	-3	0	-2	-2	0	6	0	-1	-2	-2	-2	-2	-2	-3	-4	-4	-3	-3	-3	-2
N	-3	1	0	-2	-2	0	6	1	0	0	1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	-1	-1	-2	-1	1	6	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	-1	-1	-1	-2	0	2	5	2	0	0	1	-2	-3	-3	-2	-3	-2	-3
Q	-3	0	-1	-1	-1	-2	0	0	2	5	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8	0	-1	-2	-3	-3	-3	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5	-1	-3	-2	-2	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5	1	2	1	0	-1	-1
I	-1	-3	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4	2	3	0	-1	-3
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4	1	0	-1	-2
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	2
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

Appendix table 2 amino acid hydrophobicity substitution matrix

	R	K	D	E	S	N	Q	G	T	H	A	C	M	P	V	L	I	Y	F	W
R	10	10	9	9	6	6	6	5	5	5	5	4	3	3	3	3	3	2	1	0
K	10	10	9	9	6	6	6	5	5	5	5	4	3	3	3	3	3	2	1	0
D	9	9	10	10	7	6	6	6	5	5	5	5	4	4	4	3	3	3	2	1
E	9	9	10	10	7	6	6	6	5	5	5	5	4	4	4	3	3	3	2	1
S	6	6	7	7	10	10	10	10	9	9	9	8	8	7	7	7	7	6	6	4
N	6	6	6	6	10	10	10	10	9	9	9	8	8	8	7	7	7	6	6	4
Q	6	6	6	6	10	10	10	10	9	9	9	8	8	8	7	7	7	6	6	4
G	5	5	6	6	10	10	10	10	9	9	9	8	8	8	8	7	7	6	6	5
T	5	5	5	5	9	9	9	9	10	10	10	9	9	8	8	8	8	7	7	5
H	5	5	5	5	9	9	9	9	10	10	10	9	9	9	8	8	8	7	7	5
A	5	5	5	5	9	9	9	9	10	10	10	9	9	9	8	8	8	7	7	5
C	4	4	5	5	8	8	8	8	9	9	9	10	10	9	9	9	9	8	8	5
M	3	3	4	4	8	8	8	8	9	9	9	10	10	10	10	9	9	8	8	7
P	3	3	4	4	7	8	8	8	8	9	9	9	10	10	10	9	9	9	8	7
V	3	3	4	4	7	7	7	8	8	8	8	9	10	10	10	10	10	9	8	7
L	3	3	3	3	7	7	7	7	8	8	8	9	9	9	10	10	10	9	9	8
I	3	3	3	3	7	7	7	7	8	8	8	9	9	9	10	10	10	9	9	8
Y	2	2	3	3	6	6	6	6	7	7	7	8	8	9	9	9	9	10	10	8
F	1	1	2	2	6	6	6	6	7	7	7	8	8	8	8	9	9	10	10	9
W	0	0	1	1	4	4	4	5	5	5	5	6	7	7	7	8	8	8	9	10