ACTA ABP BIOCHIMICA POLONICA

Vol. 65, No 4/2018
521–534
https://doi.org/10.18388/abp.2018_2351

Regular paper

In silico analysis of different signal peptides to discover a panel of appropriate signal peptides for secretory production of Interferon-beta 1b in *Escherichia coli*

Shahrokh Ghovvati[1]✉, Zahra Pezeshkian[2] and Seyed Ziaeddin Mirhoseini[2]

[1]Department of Biotechnology, Faculty of Agriculture, University of Guilan, Rasht, Guilan, Iran; [2]Department of Animal Science, Faculty of Agriculture, University of Guilan, Rasht, Guilan, Iran

Signal peptides (SPs) are one of the most important factors for suitable secretion of the recombinant heterologous proteins in *Escherichia coli* (*E. coli*). The objective of this study was to identify a panel of signal peptides (among the 90 biologically active SPs) required for the secretory production of interferon-beta 1b (IFN-beta 1b) recombinant protein into the periplasmic space of *E. coli* host. In the initial step, after predicting the accurate locations of the cleavage sites of signal peptides and their discrimination scores using SignalP 4.1 server, 31 SPs were eliminated from further analysis because their discrimination scores were less than 0.5 or their cleavage sites were inappropriately located. Therefore, only 59 SPs could be theoretically applied to secrete IFN-beta 1b into the periplasmic space of *E. coli*. The physico-chemical and the solubility properties, which are necessary parameters for selecting appropriate SPs, were predicted using ProtParam and SOLpro servers using the 59 remaining signal peptides. The final subcellular localization of IFN-beta 1b in combination with different SPs was predicted using ProtComB server. Consequently, according to the ranking of 59 confirmed SPs, the obtained results revealed that SPs Flagellar P-ring protein (flgI), Glucan 1,3-beta-glucosidase I/II (EXG1) and outer membrane protein C (OmpC) were theoretically the most potent and desirable SPs for secretion of recombinant IFN-beta 1b into the periplasmic space of *E. coli*. For further studies in the future, the experimental investigations on the obtained results will be considered.

Key words: bioinformatics, *in silico*, signal peptide, Interferon-beta 1b, recombinant protein, *E. coli*

Received: 11 February, 2017; revised: 06 July, 2018; accepted: 06 July, 2018; available on-line: 31 October, 2018

✉e-mail: Ghovvati@guilan.ac.ir

Abbreviations: c-region, cleavable region; D-score, discriminating score; EXG1, glucan 1,3-beta-glucosidase I/II; flgI, flagellar P-ring protein; GRAVY, grand average of hydropathicity; h-region, hydrophobic region; IFN-beta 1b, interferon-beta 1b; MS, multiple sclerosis; n-region, N-terminal region; OmpC, outer membrane protein C; SPDs, sec-dependent proteins; SP, signal peptide; SVM, support vector machine; Tat, twin arginine translocation; Uniprot, universal protein resource

INTRODUCTION

Cytokines are a group of proteins used for communication between cells which helps the protective mechanisms of the immune system to destroy pathogens (Cohen & Parkin, 2001). Among cytokines, interferons (IFNs) are secreted proteins which induce an antiviral state in their target cells. The IFNs are the first line of defense against viral diseases and possibly against viral infectious agents, and they are accepted universally as therapeutic agents (Sen & Lengyel, 1992). There are two different types of recombinant IFN-beta. The first type, IFN-beta 1a, is a glycosylated form of pharmaceutical recombinant protein and is produced in mammalian cells, such as CHO cell line (Sørensen, 2010; Morowvat *et al.,* 2014). The second type, IFN-beta 1b, is not a glycosylated protein and is produced in *Escherichia coli* recombinant cells (Runkel *et al*., 1998; Morowvat *et al.,* 2014). IFN beta-1b is a synthetic analogue (165 amino acids, 18500 Da) of human IFN in which cysteine 17 is substituted with serine. IFN beta-1b was the first disease-modifying drug approved by FDA in 1993 for treatment of (relapsing/remitting) multiple sclerosis (MS) (Marziniak & Meuth, 2014; Gasteiger *et al.,* 2005). This disease is a chronic inflammatory disorder of the central nervous system affecting mainly young adults (Kolb-Maurer *et al.,* 2015).

The vast majority of recombinant proteins have been produced in the Gram-negative bacteria. Furthermore, *E. coli* is a Gram-negative bacterium which has numerous benefits for heterologous recombinant protein production due to its ability of growing rapidly at high density on inexpensive substrates, the profound genetic and physiological characterization, the large number of compatible tools available for biotechnology especially cloning vectors and host strains, and the simple process scale up (Baneyx, 1999; Babaeipour *et al.,* 2013). The inability of proteins to rapidly fold into native structures causes their degradation into insoluble aggregates known as inclusion bodies or inactive proteins (Baneyx & Mujacic, 2004; Choi & Lee, 2004; Ventura & Villaverde, 2006). One approach to solve these problems is transferring the heterologous proteins into the periplasmic space of the bacterial host, using a proper signal peptide at the N-terminal of the protein (Choi & Lee, 2004). The periplasm has many benefits which make it a preferable space for protein storage and folding as well as protein purification and N-terminal processing. Also, it provides more stability for expressed heterologous protein because of proper folding and decreased protein degradation. Moreover, it ensures an oxidizing environment to promote proper folding of the produced proteins *(Mergulha˜o et al.,* 2005; Morowvat *et al.,* 2014). Some reports revealed the toxic effect of human IFN-beta expressed in *E. coli* host (Gross et *al.,* 1985; Morowvat *et al.,* 2014). To solve the problem, periplasmic production of this recombinant protein was considered as the most important strategy for its easy accumulation in the safe space to overcome its lethal effects on the host cells (Morowvat *et al.,* 2014).

The general secretory pathway is one of the mechanisms for protein secretion. It is found in both eukaryotic and prokaryotic cells. The entrance into the general secretory pathway is controlled by the signal peptide. It is an N-terminal peptide which is normally between 15 and 40 amino acids long and detached from the mature section of the protein during translocation across the membrane (Emanuelsson *et al.,* 2007; Zhang *et al.,* 2013). The most characteristic feature of the signal peptides is a segment of hydrophobic amino acids called the h-region containing 7 to 15 residues (Haeuptle *et al.,* 1989). The area between the first methionine and the h-region is called the n-region and it typically contains one to five amino acids and normally carries a positive charge. The c-region is between the h-region and the cleavage site and it consists of three to seven polar, but mostly uncharged, amino acids (Nielsen & Krogh, 1998). Obviously, for selecting an optimal signal sequence, which is compatible with the recombinant secretory protein, it is normally necessary to consider and analysis the feature of signal peptides for the efficient secretory production of the heterologous recombinant proteins (Gupta & Shukla, 2016; Forouharmehr *et al.,* 2018).

In Gram-negative bacteria, the secretion system can transport proteins into periplasmic space by the Sec-dependent pathway or Twin-arginine translocation (Tat) pathway. In Sec-dependent pathway, the secreted preproteins carry a signal peptide that is unfolded in this step and directed through a protein-conducting channel, whereas in the Tat pathway, the folded preproteins are translocated across the inner membrane (Bagos *et al.,* 2010; Yoon *et al.,* 2010).

A number of signal sequences have been used for the efficient secretory production of heterologous recombinant proteins in *E. coli* host. Over the past years, different signal peptides have been used and analyzed for transportation of proteins into the periplasmic or extracellular space. For example, OmpA and PelB (Ramanan *et al.,* 2010), and PelB (Morowvat *et al.,* 2014) were used for IFN-alpha 2b and IFN-beta 1b secretory production in *E. coli*, respectively.

Nowadays, with progress in genetic engineering and computer technology, conducting reliable and fast computational programs for predicting the signal peptides are required for improving the production level and minimizing the production expenses. In this study, some important features of IFN-beta 1b and the other 89 different signal peptides were studied using bioinformatics tools, in order to theoretically identify the efficient signal peptides that can be suitably used for secretion of IFN-beta 1b in *E. coli* host.

MATERIALS AND METHODS

Signals sequence collection and study design. The amino acid sequences of 90 signal peptides that were usually used in secretory proteins production in previous studies were retrieved from database resources of the Universal Protein Resource (UniProt) at www.uniprot.org. Signal sequences are tabulated in Table 1. Afterward, in the next phase, *in silico* methods were utilized to analyze and characterize the collected signal peptide sequences. Eventually, after trimming, predicting the subcellular localization site and excluding inappropriate signal peptides, the elected signal peptides were compared and measured to gain a high level of secretory expression of IFN-beta 1b protein in the *E. coli* host.

Prediction of signal peptides cleavage sites. Among the proposed data mining models and bioinformatics tools used for identifying signal peptides sequences and their precise cleavage sites, SignalP was the most accurate and reliable tool, which provided high-through processing of protein sequences with an accuracy of 87% (Dyrløv Bendtsen *et al.,* 2004). The SignalP 4.1 is available at http://www.cbs.dtu.dk/services/SignalP, and is based on ANN method (Petersen *et al.,* 2011).

Physico-chemical parameters of signal peptides. ProtParam online server at http://web.expasy.org/protparam/ (Gasteiger *et al.,* 2005) was used to evaluate different physico-chemical properties of the chosen signal peptides. The physico-chemical parameters were evaluated, including molecular weight, instability index, amino acid composition, theoretical pI, aliphatic index and grand average of hydropathicity (GRAVY).

***In silico* analysis for protein solubility prediction.** The solubility of recombinant protein that is expressed in *E. coli* most often represents the production yield. SOLpro server at http://scratch.proteomics,ics.uci.edu/ with an accuracy of above 74% was used to predict the tendency of protein solubility in *E. coli*. This bioinformatics tool employed a two-phase support vector machine (SVM) architecture according to multiple representations of the primary sequence (Magnan *et al*., 2009).

Prediction of protein localization site. Predicting the final destination of secreted proteins from the primary sequence is a major component of automated protein annotation and is critical to a wide range of studies. ProtComp B server was used for *in silico* analysis and prediction of the final destination of IFN-beta 1b protein in fusion with different signal peptides (http://www.softberry.com). Softberry reports 86% correct prediction of extracellular proteins as tested with approximately 200 extracellular proteins (Klee & Ellis, 2005).

RESULTS

In silico analysis of signal peptides cleavage sites' prediction

In this study, SignalP 4.1 was used to predict the most suitable amino acid sequences as a signal peptide for connecting to IFN-beta 1b in order to secrete that protein inside the periplasmic space in *E. coli*. The suitable signal peptides were predicted based on their potential D-score (discrimination score). The output of SignalP 4.1 reported five scores. The C and S-score recognized cleavage sites and signal peptide positions respectively. Y-score was a derivative of the C and S-score resulting in the more precise prediction of the cleavage sites than the raw C-score. The average of the S-score was S-mean. D-score was the average of the S-mean and Y-max which indicated the primary distinction between secretory and non-secretory proteins. Sequences with D-score >0.5 have a high probability of being signal peptides. In SignalP 4.1, there is an option for the user to adjust the cut-off values in order to increase the sensitivity of the program. In this research, a default SignalP D-score of 0.5 was used. The *in silico* analysis results of SignalP including three regions of signal peptides (n, h and c), cleavage probabilities and cleavage sites and C, Y and S scores, S means and D-score, are presented in Table 2. Cleavage probability is the maximum cleavage site probability at the beginning of the protein. The *in silico* analysis results of SignalP server indicated that the highest D-score belonged

Table 1. List of signal peptides

No	Full name	Signal peptide	Accession no.	Source	Amino acid sequence
1	interferon beta	IFN-beta	P01574	*Homo sapiens*	**MTNK**CLLQIALLLCFSTTALS
2	human growth hormone	hGH	P01241	*Homo sapiens*	**MATGSRTS**LLLAFGLLCLPWLQEGSA
3	pectate lyase B	Pelb	P0C1C1	*Erwinia Carotovorum*	**MKYLLP**TAAAGLLLLAAQPAMA
4	Alkaline phosphatase	PhoA	P00634	*Escherichia coli* (strain K 12)	**MKQST**IALALLPLLFTPVTKA
5	Outer membrane pore protein E	PhoE	P02932	*Escherichia coli* (strain K 12)	**MKKST**LALVVMGIVASASVQA
6	outer membrane protein A	OmpA	P0A910	*Escherichia coli* (strain K 12)	**MKKT**AIAIAVALAGFATVAQA
7	outer membrane protein F	OmpF	P02931	*Escherichia coli* (strain K 12)	**MMKRNI**LAVIVPALLVAGTANA
8	outer membrane protein C	OmpC	P06996	*Escherichia coli* (strain K 12)	**MKVK**VLSLLVPALLVAGAANA
9	protease VII	OmpT	P09169	*Escherichia coli* (strain K 12)	**MRAK**LLGIVLTTPIAISSFA
10	maltose-binding periplasmic protein	MalE	P0AEX9	*Escherichia coli* (strain K 12)	**MKIKTGAR**ILALSALTTMMFSASALA
11	maltoporin	LamB	P02943	*Escherichia coli* (strain K 12)	**MMITLRKLPL**AVAVAAGVMSAQAMA
12	major outer membrane lipoprotein	LPP	P69776	*Escherichia coli* (strain K 12)	**MKATK**LVLGAVILGSTLLAG
13	heat-stable entrotoxin II	STII	P22542	*Escherichia coli*	**MKKNI**AFLLASMFVFSIATNAYA
14	disulfide interchange protein DsbA	DsbA	P0AEG4	*Escherichia coli* K-12	**MKKIW**LALAGLVLAFSASA
15	disulfide interchange protein DsbC	DsbC	P0AEG6	*Escherichia coli*	**MKKG**FMLFTLLAAFSGFAQA
16	Beta-lactamase	ampC	P00811	*Escherichia coli* (strain K12)	**MFKT**TLCALLITASCSTFA
17	heat-labile entrotoxin subunit B	LTB	P13811	*Escherichia coli*	**MNKVK**CYVLFTALLSSLYAHG
18	L-Asparaginase II	L-Asparaginase II	P00805	*Escherichia coli* K-12	**MEFFKK**TALAALVMGFSGAALA
19	Endo-1,4-beta-xylanase	Endo-1,4-beta-xylanase	Q59256	Bacillus sp. YA-14	**MFKFKKK**FLVGLTAAFMSISMFSATASA
20	fimbrial chaperone SfmC	sfmC	P77249	*Escherichia coli* K-12	**MMTKIK**LLMLIIFYLIISASAHA
21	heat-stable entrotoxin ST-IA/ST-P	ST-IA/ST-P	P01559	*Escherichia coli*	**MKKL**MLAIFISVLSFPSFS
22	D-galactose-binding periplasmic protein	MglB	P0AEE5	*Escherichia coli* K-12	**MNKK**VLTLSAVMASMLFGAAAHA
23	beta-lactamase TEM	bla	P62593	*Escherichia coli*	**MSIQHFR**VALIPFFAAFCLPVFA
24	neutral protease	npr	P06832	*Bacillus amyloliquefaciens*	**MGLGKK**LSVAVAASFMSLTISLPGVQA
25	protein TolB	TolB	P0A855	*Escherichia coli* K-12	**MKQALR**VAFGFLILWASVLHA
26	periplasmice protein TorT	TorT	P38683	*Escherichia coli* K-12	**MRV**LLFLLLSLFMLPAFS
27	Interleukin-2	IL-2	P60568	*Homo sapiens*	**MYRMQ**LLSCIALSLALVTNS
28	Chaperone protein Skp	SKp	P0AEU7	*Escherichia coli* (strain K12)	**MKKW**LLAAGLGLALATSAQA
29	D-ribose-binding periplasmic protein	rbsB	P02925	*Escherichia coli* (strain K12)	**MNMKKL**ATLVSAVALSATVSANAMA
30	Glucan 1,3-beta-glucosidase I/II	EXG1	P23776	*Saccharomyces cerevisiae*	**MLSLKT**LLCTLLTVSSVLA
31	Heat-labile enterotoxin IIB	LT-IIB	P43529	*Escherichia coli*	**MSFKKIIK**AFVIMAALVSVQAHA
32	Periplasmic serine endoprotease DegP	degP	P0C0V0	*Escherichia coli* (strain K12)	**MKKTT**LALSALALSLGLALSPLSATA
33	Spheroplast protein Y	spy	P77754	*Escherichia coli* (strain K12)	**MRKLT**ALFVASTLALGAANLAHA

34	Flagellar P-ring protein	flgI	P0A6S3	*Escherichia coli* (strain K12)	**MIKF**LSALILLLVTTAAQA
35	Secreted 45 kDa protein	usp45	P22865	*Lactococcus lactis*	**MKKKII**SAILMSTVILSAAAPLSGVYA
36	Immunoglobulin G-binding protein A	spa	P38507	*Staphylococcus aureus*	**MKKKNIYSIRK**LGVGIASVTLGTLLISGGVT-PAANA
37	Endoglucanase	egIS	P10475	*Bacillus subtilis*	**MKRSIS**IFITCLLITLLTMGGMIASPASA
38	Outer membrane protein G	ompG	P76045	*Escherichia coli* (strain K12)	**MKKL**LPCTALVMCAGMACAQA
39	Dr hemagglutinin structural subunit	draA	P24093	*Escherichia coli*	**MKKL**AIMAAASMVFAVSSAHA
40	Glutamine-binding periplasmic protein	glnH	P0AEQ3	*Escherichia coli* (strain K12)	**MKSVLK**VSLAALTLAFAVSSHA
41	Ribonuclease I	rna	P21338	*Escherichia coli* (strain K12)	**MKAFWRN**AALLAVSLLPFSSANA
42	L-arabinose-binding periplasmic protein	araF	P02924	*Escherichia coli* (strain K12)	**MHKFTK**ALAAIGLAAVMSQSAMA
43	Putative outer membrane porin protein	nmpC	P21420	*Escherichia coli* (strain K12)	**MKKLT**VAISAVAASVLMAMSAQA
44	Penicillin-insensitive mureinendopeptidase	mepA	P0C0T5	*Escherichia coli* (strain K12)	**MNKT**AIALLALLASSASLA
45	Peptidyl-prolyl cis-trans isomerase A	ppiA	P0AFL3	*Escherichia coli* (strain K12)	**MFKSTL**AAMAAVFALSALSPAAMA
46	Carcinoembryonic antigen-related cell adhesion molecule 5	CEACAM5	P06731	*Homo sapiens* (Human)	**MESPSAPPHRWCIPWQR**LLLTASLLT-FWNPPTTA
47	Beta-defensin 103	DEFB103A	P81534	*Homo sapiens* (Human)	**MRIHY**LLFALLFLFLVPVPGHG
48	Interferon omega-1	IFNW1	P05000	*Homo sapiens* (Human)	**MA**LLFPLLAALVMTSYSPVGS
49	Azurocidin	AZU1	P20160	*Homo sapiens* (Human)	**MTR**LTVLALLAGLLASSRAGSSPLLD
50	Interferon alpha-2	IFNA2	P01563	*Homo sapiens* (Human)	**MAL**TFALLVALLVLSCKSSCSVG
51	C-C motif chemokine 20	CCL20	P78556	*Homo sapiens* (Human)	**MCCTKS**LLLAALMSVLLLHLCGESEA
52	Cystatin-11	CST11	Q9H112	*Homo sapiens* (Human)	**MMAEPWQA**LQLLLAILLTLMALPYQA
53	Interferon alpha-4	IFNA4	P05014	*Homo sapiens* (Human)	**MALS**FSLLMAVLVLSYKSICSLG
54	Heat-stable enterotoxin receptor	GUCY2C	P25092	*Homo sapiens* (Human)	**MKT**LLLDLALWSLLFQPGWLSFS
55	Kallikrein-7	KLK7	P49862	*Homo sapiens* (Human)	**MARS**LLLPLQILLLSLALETAG
56	Periplasmic beta-glucosidase	bglX	P33363	*Escherichia coli* (strain K12)	**MKWL**CSVGIAVSLALQPALA
57	Plasma protease C1 inhibitor	SERPING1	P05155	*Homo sapiens* (Human)	**MASRL**TLLTLLLLLLAGDRASS
58	Interferon alpha-1/13	IFNA1	P01562	*Homo sapiens* (Human)	**MASPF**ALLMVLVVLSCKSSCSLG
59	Protransforming growth factor alpha	TGFA	P01135	*Homo sapiens* (Human)	**MVPSAGQ**LALFALGIVLAACQAL
60	Collagen alpha-1(IV) chain	COL4A1	P02462	*Homo sapiens* (Human)	**MGPRLS**VWLLLLPAALLLHEEHSRAAA
61	Ecotin	eco	P23827	*Escherichia coli* (strain K12)	**MKTI**LPAVLFAAFATTSAWA
62	Serum albumin	ALB	P02768	*Homo sapiens* (Human)	**MKW**VTFISLLFLFSSAYS
63	MHC class I polypeptide-related sequence A	MICA	Q29983	*Homo sapiens* (Human)	**MGLGP**VFLLLAGIFPFAPPGAAA
64	L-asparaginase 2	ansB	P00805	*Escherichia coli* (strain K12)	**MEFFKK**TALAALVMGFSGAALA
65	Uncharacterized fimbrial-like protein YehD	yehD	P33343	*Escherichia coli* (strain K12)	**MKRS**IIAAAVFSSFFMSAGVFA
66	Uncharacterized protein YfjT	yfjT	P52135	*Escherichia coli* (strain K12)	**MKIRSLSR**FVLASTMFASFTASA
67	Outer membrane protease OmpP	ompP	P34210	*Escherichia coli* (strain K12)	**MQTK**LLAIMLAAPVVFSSQEASA
68	Iron uptake system component EfeO	efeO	P0AB24	*Escherichia coli* (strain K12)	**MTINFRRN**ALQLSVAALFSSAFMANA
69	D-alanyl-D-alanine endopeptidase	pbpG	P0AFI5	*Escherichia coli* (strain K12)	**MPKFR**VSLFSLALMLAVPFAPQAVA

70	Probable fimbrial chaperone YadV	yadV	P33128	*Escherichia coli* (strain K12)	**MFFNTKHTT**ALCFVTCMAFSSSSIA
71	Glucose-1-phosphatase	agp	P19926	*Escherichia coli* (strain K12)	**MNKT**LIAAAVAGIVLLASNAQA
72	Endonuclease-1	endA	P25736	*Escherichia coli* (strain K12)	**MYRYL**SIAAVVLSAAFSGPALA
73	Threonine-rich inner membrane protein GfcA	gfcA	P75885	*Escherichia coli* (strain K12)	**MKHK**LSAILMAFMLTTPAAFA
74	Protein GltF	gltF	P28721	*Escherichia coli* (strain K12)	**MFFKKNL**TTAAICAALSVAAFSAMA
75	Endoglucanase	bcsZ	P37651	*Escherichia coli* (strain K12)	**MNVLR**SGIVTMLLLAAFSVQA
76	Thiamine-binding periplasmic protein	thiB	P31550	*Escherichia coli* (strain K12)	**MLKK**CLPLLLLCTAPVFA
77	D-xylose-binding periplasmic protein	xylF	P37387	*Escherichia coli* (strain K12)	**MKIKNI**LLTLCTSLLLTNVAAHA
78	UPF0412 protein YaaI	yaaI	P28696	*Escherichia coli* (strain K12)	**MKSVFT**ISASLAISLMLCCTAQA
79	Uncharacterized protein YaaX	yaaX	P75616	*Escherichia coli* (strain K12)	**MKKMQS**IVLALSLVLVAPMAAQA
80	Uncharacterized fimbrial-like protein YadC	yadC	P31058	*Escherichia coli* (strain K12)	**MKTIFRY**ILFLALYSCCNTVSA
81	Uncharacterized fimbrial-like protein YadN	yadN	P37050	*Escherichia coli* (strain K12)	**MSKK**LGFALSGLMLAMVAGTASA
82	Protein YdgH	ydgH	P76177	*Escherichia coli* (strain K12)	**MKLKN**TLLASALLSAMAFS
83	Uncharacterized fimbrial-like protein YfcQ	yfcQ	P76500	*Escherichia coli* (strain K12)	**MRKT**FLTLLCVSSAIAHA
84	Uncharacterized protein YhcF	yhcF	P45422	*Escherichia coli* (strain K12)	**MNNVK**LLIAGSAFFAMSAQA
85	Uncharacterized protein YiiX	yiiX	P32167	*Escherichia coli* (strain K12)	**MKNR**LLILSLLVSVPAFA
86	Uncharacterized protein YpeC	ypeC	P64542	*Escherichia coli* (strain K12)	**MFRS**LFLAAALMAFTPLAANA
87	Uncharacterized fimbrial-like protein YraH	yraH	P42913	*Escherichia coli* (strain K12)	**MNKVTK**TAIAGLLALFAGNAAA
88	Uncharacterized fimbrial-like protein YraK	yraK	P43319	*Escherichia coli* (strain K12)	**MKRA**PLITGLLLISTSCAYA
89	Sigma-E factor regulatory protein RseB	rseB	P0AFX9	*Escherichia coli* (strain K12)	**MKQLWF**AMSLVTGSLLFSANASA
90	Cyclic di-GMP-binding protein	bcsB	P37652	*Escherichia coli* (strain K12)	**MKRKLF**WICAVAMGMSAFPSFMTQA

to DEFB103A, nmpC, ppiA, mepA, SKp and rna respectively (0.906, 0.886, 0.875, 0.875, 0.865 and 0.852).

Moreover, the SignalP results demonstrated that the D-scores of 23 signal peptides were less than 0.5 and cleavage sites of 8 signal peptides such as DsbA, AZU1 were inappropriately located; hence, they were eliminated from further analysis. Nonetheless, further analysis was conducted on the other 59 signal peptides.

Physico-chemical parameters of signal peptides

The 59 remaining signal peptides were analyzed for their different physico-chemical parameters using ProtParam server (Table 3). The length range of the 59 remaining signal peptides was between 18 and 36 amino acids. The parameters, computed by ProtParam, included the molecular weight (daltons), theoretical pI, aliphatic index, instability index and grand average of hydropathicity (GRAVY). In ProtParam, the molecular weight of a protein is calculated by the addition of average isotopic masses of amino acids in the provided protein and the average isotopic mass of one water molecule.

The *in silico* analysis results indicated that the highest molecular weight belonged to CEACAM5, spa, Endo-1, 4-beta-xylanase, eglS and bcsB (3919.58, 3643.30, 3061.70, 3040.70 and 2853.53, respectively).

The grand average of hydropathy (GRAVY) value for a peptide or protein was calculated by the sum of hydropathy values of all the amino acids, divided by the number of residues in the sequence. The GRAVY index was normally used to compare the general hydropathy of the signal peptide. The *in silico* analysis results showed that the highest GRAVY scores belonged to flgI, EXG1, IFNA1, OmpC and MICA (1.816, 1.726, 1.604, 1.552 and 1.53, respectively).

The aliphatic index is one of the major factors which indicate the hydrophobicity value of peptides and proteins. This index is specified as the relative

Table 2. Analysis of the signal peptides sequences using SignalP 4.1 server

Signal peptides	n-region	h-region	c-region	Cleavage probability (%)	Cleavage site	C-score	Y-score	S-score	S-mean	D-score
IFN-beta	1-4 (4)	5-16 (12)	17-21(5)	71.1	ALS	0.8	0.685	0.877	0.67	0.679
hGH	1-8 (8)	9-18(10)	19-26(8)	80.2	GSA	0.465	0.627	0.96	0.858	0.752
PelB	1-6 (6)	7-16(10)	17-22(6)	99.6	AMA	0.804	0.627	0.82	0.603	0.618
PhoA	1-5 (5)	6-14(9)	15-21(7)	–	–	0.309	0.273	0.523	0.326	0.292
PhoE	1-5 (5)	6-15(10)	16-21(6)	90.0	VQA	0.695	0.55	0.738	0.526	0.541
OmpA	1-4 (4)	5-14(10)	15-21(7)	95.8	AQA	0.717	0.563	0.761	0.551	0.558
OmpF	1-6 (6)	7-16 (10)	17-22 (6)	–	–	0.732	0.475	0.748	0.45	0.466
OmpC	1-4 (4)	5-15 (11)	16-21 (6)	84.9	ANA	0.73	0.591	0.849	0.614	0.599
OmpT	1-4 (4)	5-15 (11)	16-20 (5)	–	–	0.334	0.258	0.454	0.276	0.265
MalE	1-8 (8)	9-18 (10)	19-26 (8)	96.7	ALA	0.632	0.548	0.873	0.636	0.58
LamB	1-8 (8)	9-18 (10)	19-25 (7)	97.6	AMA	0.719	0.566	0.842	0.613	0.583
LPP	1-5 (5)	6-14 (9)	15-20 (6)	–	–	0.133	0.194	0.533	0.378	0.262
STII	1-5 (5)	6-16 (11)	17-23(7)	96.7	AYA	0.731	0.526	0.816	0.547	0.534
DsbA	1-5 (5)	6-14 (9)	15-19 (5)	76.3	AMS	0.565	0.51	0.817	0.566	0.531
DsbC	1-4 (4)	5-13 (9)	14-20 (7)	98.7	AQA	0.694	0.792	0.966	0.9	0.842
DsbD	1-4 (4)	5-13 (9)	14-19 (6)	–	–	0.776	0.464	0.617	0.372	0.43
LTB	1-5 (5)	6-14 (9)	15-21 (7)	98.4	AHG	0.711	0.543	0.724	0.507	0.53
L-Asparaginase II	1-6 (6)	7-15 (9)	16-22 (7)	–	–	0.798	0.504	0.679	0.431	0.477
Endo-1,4-beta--xylanase	1-7 (7)	8-21 (14)	22-28 (7)	79.7	ASA	0.377	0.49	0.948	0.772	0.6
sfmC	1-6 (6)	7-16 (10)	17-23 (7)	–	–	0.759	0.458	0.515	0.323	0.408
ST-IA/ST-P	1-4 (4)	5-13 (9)	14-19 (6)	–	–	0.27	0.309	0.664	0.438	0.356
MglB	1-4 (4)	5-16 (12)	17-23 (7)	98.6	AHA	0.718	0.664	0.885	0.714	0.682
Bla	1-7 (7)	8-16 (9)	17-23 (7)	–	–	0.509	0.29	0.388	0.245	0.274
Npr	1-6 (6)	7-20 (14)	21- 27(7)	76.9	VQA	0.464	0.505	0.791	0.631	0.554
TolB	1-6 (6)	7-15 (9)	16-21 (6)	–	–	0.426	0.269	0.407	0.236	0.256
TorT	1-3 (3)	4-12 (9)	13-18 (6)	–	–	0.319	0.292	0.589	0.329	0.306
IL-2	1-5 (5)	6-17 (12)	18-20 (3)	64.7	TNS	0.449	0.472	0.837	0.587	0.518
SKp	1-4 (4)	5-14 (10)	15-20 (6)	86.8	AQA	0.76	0.834	0.95	0.9	0.865
rbsB	1-6 (6)	7-18 (12)	19-25 (7)	84.4	AMA	0.726	0.784	0.974	0.895	0.836
EXG1	1-6 (6)	7-14 (8)	15-19 (5)	93.1	VLA	0.732	0.644	0.837	0.625	0.637
LT-IIB	1-8 (8)	9-17 (9)	18-23 (6)	99.7	AHA	0.831	0.629	0.768	0.568	0.606
degP	1-5 (5)	6-19 (14)	20-26 (7)	91.1	ATA	0.57	0.534	0.9	0.688	0.591
spy	1-5 (5)	6-17 (12)	18-23 (6)	99.3	AHA	0.453	0.491	0.846	0.659	0.553
flgI	1-4 (4)	5-13 (9)	14-19 (6)	88.8	AQA	0.732	0.558	0.8	0.532	0.548
usp45	1-6 (6)	7-20 (14)	21-27 (7)	79.8	VYA	0.518	0.46	0.907	0.646	0.532
spa	1-11(11)	12-29 (18)	30-36 (7)	78.9	ANA	0.505	0.479	0.751	0.591	0.523
eglS	1-6 (6)	7-22 (16)	23-29 (7)	79.6	ASA	0.54	0.456	0.856	0.586	0.507
yebF	1-4 (4)	5-15 (11)	16-21 (6)	–	–	0.446	0.436	0.811	0.563	0.483
LivK	1-7 (7)	8-17 (10)	18-23 (6)	–	–	0.594	0.468	0.619	0.458	0.464
glnH	1-6(6)	7-15 (9)	16-22 (7)	94.9	SHA	0.654	0.765	0.957	0.905	0.831
rna	1-7 (7)	8-16 (9)	17-23 (7)	84.5	ANA	0.712	0.799	0.963	0.912	0.852
araF	1-6 (6)	7-16(10)	17-23 (7)	99.3	AMA	0.738	0.811	0.941	0.867	0.837
nmpC	1-5 (5)	6-16 (11)	17-23 (7)	89.3	AQA	0.778	0.848	0.979	0.929	0.886
mepA	1-4 (4)	5-13 (9)	14-19 (6)	85.8	SLA	0.696	0.816	0.971	0.941	0.875
ppiA	1-6 (6)	7-18 (12)	19-24 (6)	99.2	AMA	0.71	0.812	0.986	0.946	0.875

CEACAM5	1-17 (17)	18-27 (10)	28-34 (7)	68.6	TTA	0.568	0.518	0.769	0.556	0.533
DEFB103A	1-5 (5)	6-15 (10)	16-22 (6)	71	GHG	0.829	0.878	0.978	0.930	0.906
IFNW1	1-2 (2)	3-13 (11)	14-21 (8)	51.1	VGS	0.283	0.467	0.943	0.761	0.626
AZU1	1-3 (3)	4-14 (11)	15-26 (12)	79.7	SRA	0.628	0.753	0.945	0.899	0.832
IFNA2	1-3 (3)	4-14 (11)	15-23 (9)	27	CKS	0.383	0.606	0.984	0.956	0.795
CCL20	1-6 (6)	7-18 (12)	19-26 (8)	85.1	SEA	0.677	0.766	0.980	0.878	0.826
CST11	1-8 (8)	9-17 (9)	18-26 (9)	60	AMS	0.620	0.546	0.868	0.625	0.578
IFNA4	1-4 (4)	5-14 (10)	15-23 (9)	35.9	ICS	0.283	0.493	0.958	0.859	0.690
GUCY2C	1-3 (3)	4-15 (12)	16-23 (8)	–	–	0.268	0.364	0.811	0.571	0.446
KLK7	1-4 (4)	5-15 (11)	16-22 (7)	45.6	SLA	0.395	0.608	0.973	0.945	0.790
bglX	1-4 (4)	5-13 (9)	14-20 (7)	–	–	0.568	0.453	0.663	0.452	0.452
SERPING1	1-5 (5)	6-15 (10)	16-22 (7)	50.8	DRA	0.576	0.736	0.978	0.942	0.847
IFNA1	1-5 (5)	6-15 (10)	16-23 (8)	32.3	SLG	0.472	0.623	0.964	0.826	0.733
TGFA	1-7 (7)	8-17 (10)	18-23 (6)	–	–	0.383	0.405	0.751	0.569	0.471
COL4A1	1-6 (6)	7-18 (12)	13-21 (9)	44.5	SRA	0.296	0.489	0.957	0.820	0.667
eco	1-4 (4)	5-14 (10)	15-20 (6)	99.8	AWA	0.772	0.530	0.810	0.499	0.519
ALB	1-3 (3)	4-13 (10)	14-18 (5)	72.5	AYS	0.639	0.595	0.823	0.606	0.599
MICA	1-5 (5)	6-15 (10)	16-23 (8)	78.4	AAA	0.724	0.744	0.890	0.768	0.757
ansB	1-6 (6)	7-15 (9)	16-22 (7)	–	–	0.798	0.504	0.679	0.431	0.477
yehD	1-4 (4)	5-15 (11)	16-22 (7)	–	–	0.734	0.480	0.692	0.444	0.467
yfjT	1-8 (8)	9-17 (9)	18-23 (6)	72.7	ASA	0.528	0.536	0.861	0.665	0.584
ompP	1-4 (4)	5-15 (11)	16-23 (8)	72.9	ASA	0.510	0.560	0.853	0.689	0.621
efeO	1-8 (8)	9-20 (12)	21-26 (6)	76.2	ANA	0.479	0.645	0.963	0.879	0.755
pbpG	1-6 (6)	7-18 (12)	19-25 (7)	92.2	AVA	0.468	0.473	0.843	0.636	0.533
yadV	1-9 (9)	10-18 (9)	19-25 (7)	76.7	SIA	0.493	0.649	0.926	0.838	0.738
agp	1-4 (4)	5-16 (12)	17-22 (6)	98.8	AQA	0.804	0.612	0.734	0.540	0.585
endA	1-5 (5)	6-15 (10)	16-22 (7)	99.8	ALA	0.836	0.542	0.736	0.471	0.516
gfcA	1-4 (4)	5-14 (10)	15-21 (7)	96.4	AFA	0.748	0.607	0.841	0.611	0.609
gltF	1-7 (7)	8-18 (11)	19-25 (7)	92.5	AMA	0.683	0.657	0.921	0.764	0.697
bcsZ	1-5 (5)	6-16 (10)	17-22 (6)	–	–	0.732	0.442	0.570	0.362	0.413
thiB	1-4 (4)	5-12 (8)	13-18 (6)	67.7	VFA	0.404	0.452	0.812	0.588	0.502
xylF	1-6 (6)	7-16 (10)	17-23 (7)	97.6	AHA	0.624	0.748	0.966	0.900	0.819
yaaI	1-6 (6)	7-17 (11)	18-23 (6)	77.8	AQA	0.620	0.749	0.953	0.908	0.824
yaaX	1-6 (6)	7-17 (11)	18-23 (6)	91.4	AQA	0.620	0.564	0.846	0.631	0.589
yadC	1-7 (7)	8-16 (9)	17-22 (6)	–	–	0.420	0.412	0.742	0.517	0.451
yadN	1-4 (4)	5-16 (12)	17-23 (7)	90	ASA	0.639	0.518	0.851	0.599	0.548
ydgH	1-5 (5)	6-15 (10)	16-19 (4)	–	–	0.161	0.315	0.879	0.691	0.454
yfcQ	1-4 (4)	5-12 (8)	13-18 (6)	97.4	AHA	0.612	0.734	0.946	0.860	0.793
yhcF	1-5 (5)	6-14 (9)	15-20 (6)	73	AQA	0.644	0.703	0.889	0.769	0.734
yiiX	1-4 (4)	5-12 (8)	13-18 (6)	–	–	0.703	0.482	0.647	0.414	0.457
ypeC	1-4 (4)	5-14 (10)	15-21 (7)	92.4	ANA	0.554	0.712	0.983	0.933	0.816
yraH	1-6 (6)	7-15 (9)	16-22 (7)	96.7	AAA	0.626	0.569	0.810	0.621	0.588
yraK	1-4 (4)	5-13 (9)	14-20 (7)	–	–0	0.612	0.454	0.658	0.429	0.444
rseB	1-6 (6)	7-16 (10)	17-23 (7)	95.4	ASA	0.626	0.554	0.800	0.599	0.571
bcsB	1-6 (6)	7-17 (11)	18-25 (8)	55.6	TQA	0.353	0.535	0.980	0.877	0.696

Table 3. Analysis of physicochemical properties of the signal peptides using ProtParam

No.	Signal peptides	Amino acids number	MW (Da)	PI	Net positive charge	GRAVY	Aliphatic index	Instability (Separately)	Instability (in fusion with IFN-bata 1b)	Solubility
1	IFN-beta	21	2,284.80	7.82	1	1.238	139.52	36.97	52.09	Insoluble
2	hGH	26	2,736.20	5.75	1	0.777	116.54	56.9	54.40	Insoluble
3	pelb	22	2,228.70	8.34	1	1.191	138.18	41.42	52.53	Insoluble
4	PhoE	21	2,104.50	10	2	1.195	130	1.44	48.10	Insoluble
5	OmpA	21	2,046.50	10	2	1.295	121.43	9.52	49.01	Insoluble
6	OmpC	21	2,078.60	10	2	1.552	171.9	14.37	49.56	Insoluble
7	MalE	26	2,698.30	11.17	3	1.012	113.08	2.85	47.08	Insoluble
8	LamB	25	2,545.20	11	2	1.332	125.2	42.97	52.56	Insoluble
9	STII	23	2,552.00	9.7	2	1.026	102.17	32.43	51.38	Insoluble
10	DsbC	20	2,179.60	10	2	0.842	78.5	5.25	48.77	Insoluble
11	LTB	21	2,358.80	9.1	2	0.695	111.43	26.85	50.96	Insoluble
12	Endo-1,4-beta-xylanase	28	3,061.70	10.48	4	0.871	70	11.57	47.88	Insoluble
13	MglB	23	2,362.80	10	2	0.952	102.17	14.15	49.16	Insoluble
14	npr	27	2,677.20	10	2	1.041	119.26	26.75	50.19	Insoluble
15	IL-2	20	2,227.7	7.98	1	1.085	141.5	13.65	49.67	Insoluble
16	SKp	20	2,014.4	10	2	0.860	127.5	15.67	49.89	Insoluble
17	rbsB	25	2,494.00	10	2	0.948	109.6	11.14	48.40	Insoluble
18	EXG1	19	2,006.50	7.98	1	1.726	179.47	32.3	51.78	Insoluble
19	LT-IIB	23	2,504.10	10.3	3	1.23	127.39	16.43	49.43	Insoluble
20	degP	26	2,543.10	10	2	1.077	143.08	26.52	50.28	Insoluble
21	spy	23	2,340.80	11	2	0.991	127.83	-4.79	46.85	Insoluble
22	flgI	19	2,017.50	8.5	1	1.816	180	10.67	49.56	Insoluble
23	usp45	27	2,777.40	10	3	1.174	141.11	50.14	53.47	Insoluble
24	spa	36	3,643.30	10.58	5	0.511	121.94	16.56	47.33	Insoluble
25	eglS	29	3,040.70	9.5	2	1.31	131.38	68.74	56.20	Insoluble
26	glnH	22	2,244.7	10	2	1.209	133.18	10.58	48.93	Insoluble
27	rna	23	2,478.9	11	2	0.757	106.52	40.05	52.31	Insoluble
28	araF	23	2,348.8	10	2	0.878	93.91	96.71	59.20	Insoluble
29	nmpC	23	2,292.8	10	2	1.243	119.13	30.34	51.13	Insoluble
30	mepA	19	1,859.2	8.5	1	1.353	154.74	32.07	51.76	Insoluble
31	ppiA	24	2,371.9	8.5	1	1.438	98.33	39.94	52.53	Insoluble
32	CEACAM5	34	3919.58	8.02	2	-0.121	77.65	72.52	57.15	Insoluble
33	DEFB103A	22	2554.18	8.54	1	1.436	155.00	53.25	53.92	Insoluble
34	IFNW1	21	2181.68	5.27	0	1.495	134.76	49.25	53.47	Insoluble
35	CCL20	26	2750.40	5.38	1	1.296	142.69	50.97	53.60	Insoluble
36	IFNA1	23	2356.94	7.82	1	1.604	131.30	79.13	57.07	Insoluble
37	eco	20	2110.54	8.50	1	1.265	103.00	30.88	51.52	Insoluble
38	ALB	18	2140.57	8.34	1	1.233	108.33	17.57	50.44	Insoluble
39	MICA	23	2227.74	5.28	0	1.530	119.13	44.05	52.80	Insoluble
40.	yfjT	23	2523.01	12.01	3	0.757	80.87	33.35	51.49	Insoluble
41	ompP	23	2406.88	5.75	1	0.904	114.78	44.47	52.85	Insoluble
42	efeO	26	2845.33	12.00	2	0.654	94.23	54.20	54.03	Insoluble
43	pbpG	25	2705.36	11.00	2	1.228	117.20	57.99	54.53	Insoluble
44	yadV	25	2746.22	7.85	1	0.852	54.80	40.88	52.29	Insoluble

45	agp	22	2140.57	8.50	1	1.227	146.82	12.59	49.16	Insoluble
46	endA	22	2271.70	8.34	1	1.286	124.55	17.95	49.79	Insoluble
47	gfcA	21	2293.87	10.00	2	1.019	98.10	40.98	52.54	Insoluble
48	gltF	25	2579.13	9.31	2	1.244	90.40	8.52	48.05	Insoluble
49	xylF	23	2482.08	9.31	2	1.083	161.30	33.61	51.53	Insoluble
50	yaaI	23	2389.93	7.82	1	1.365	114.78	23.74	50.32	Insoluble
51	yaaX	23	2414.06	10.00	2	1.261	144.35	46.33	53.07	Insoluble
52	yadN	23	2255.78	10.00	2	1.078	102.17	21.29	50.03	Insoluble
53	yfcQ	18	1962.40	9.50	2	1.006	119.44	13.91	50.08	Insoluble
54	yhcF	20	2084.48	8.50	1	0.915	98.00	25.79	50.97	Insoluble
55	ypeC	21	2227.71	9.50	1	1.376	107.62	36.32	52.02	Insoluble
56	yraH	22	2146.58	10.00	2	0.927	115.91	8.73	48.71	Insoluble
57	rseB	23	2473.93	8.50	1	0.865	97.83	41.88	52.53	Insoluble
58	bcsB	25	2853.53	10.06	3	0.688	58.80	48.06	53.23	Insoluble

Table 4. Prediction of protein localization

Signal peptide	Sub-cellular location Score				
	Cytoplasmic	Membrane	Secreted	Periplasmic	Final Prediction Site
flgI	0.00	3.34	0.00	6.66	Periplasmic
EXG1	0.32	9.55	0.02	0.11	Outer Membrane
OmpC	0.20	9.64	0.00	0.16	Outer Membrane
mepA	0.00	6.26	0.00	3.74	Outer Membrane
degP	0.07	7.32	0.00	2.61	Outer Membrane
IL-2	0.00	9.84	0.00	0.16	Outer Membrane
usp45	0.18	9.78	0.00	0.04	Outer Membrane
IFN-beta	0.07	9.78	0.00	0.15	Outer Membrane
pelb	0.57	8.27	1.16	0.00	Outer Membrane
glnH	0.26	4.69	0.00	5.05	Periplasmic
eglS	0.15	9.48	0.00	0.36	Outer Membrane
PhoE	0.21	9.75	0.00	0.04	Outer Membrane
spy	0.04	9.58	0.00	0.38	Outer Membrane
SKp	0.00	7.65	0.00	2.35	Outer Membrane
LT-IIB	0.11	9.86	0.00	0.02	Outer Membrane
LamB	0.22	9.74	0.00	0.04	Outer Membrane
spa	0.00	4.05	5.95	0.00	Secreted (Extracellular)
OmpA	0.13	9.67	0.00	0.20	Outer Membrane
npr	0.29	9.62	0.00	0.09	Outer Membrane
nmpC	0.23	9.77	0.00	0.00	Outer Membrane
hGH	0.24	9.67	0.00	0.09	Outer Membrane
MalE	0.00	1.47	0.00	8.53	Periplasmic
LTB	0.14	9.84	0.00	0.02	Outer Membrane
rbsB	0.55	6.56	0.00	2.88	Outer Membrane
rna	0.25	9.55	0.00	0.20	Outer Membrane
STII	0.10	9.89	0.00	0.02	Outer Membrane
MglB	0.00	5.53	0.00	4.47	Outer Membrane
ppiA	0.00	6.07	0.00	3.93	Outer Membrane
araF	0.09	3.61	0.00	6.30	Periplasmic

DsbC	0.00	5.51	0.00	4.49	Outer Membrane
Endo-1,4-beta-xylanase	0.02	9.91	0.00	0.07	Outer Membrane
CEACAM5	0.86	8.69	0.35	0.10	Outer Membrane
DEFB103A	0.10	9.85	0.00	0.05	Outer Membrane
IFNW1	0.00	9.86	0.00	0.14	Outer Membrane
CCL20	0.81	9.14	0.05	0.00	Outer Membrane
IFNA1	0.04	9.96	0.00	0.00	Outer Membrane
eco	0.00	3.47	0.00	6.53	Periplasmic
ALB	0.16	9.84	0.00	0.00	Outer Membrane
MICA	0.18	9.79	0.00	0.03	Outer Membrane
yfjT	0.20	9.80	0.00	0.00	Outer Membrane
ompP	0.27	9.66	0.00	0.08	Outer Membrane
efeO	0.00	0.93	0.00	9.07	Periplasmic
pbpG	0.00	2.74	0.00	7.26	Periplasmic
yadV	0.00	0.41	0.00	9.59	Periplasmic
agp	0.00	7.91	0.00	2.09	Outer Membrane
endA	0.00	6.84	0.00	3.16	Outer Membrane
gfcA	0.16	9.83	0.00	0.01	Outer Membrane
gltF	0.21	9.79	0.00	0.00	Outer Membrane
xylF	0.25	1.90	0.00	7.85	Periplasmic
yaaI	0.18	9.66	0.00	0.16	Outer Membrane
yaaX	0.16	9.64	0.00	0.20	Outer Membrane
yadN	0.33	9.67	0.00	0.00	Outer Membrane
yfcQ	0.00	9.77	0.00	0.23	Outer Membrane
yhcF	0.11	9.38	0.00	0.51	Outer Membrane
ypeC	0.13	9.63	0.00	0.24	Outer Membrane
yraH	0.31	9.56	0.12	0.00	Outer Membrane
rseB	0.00	6.10	0.00	3.90	Outer Membrane
bcsB	0.00	9.86	0.00	0.14	Outer Membrane

Table 5. Sorting the signal peptides according to the aliphatic index

Signal peptides	Aliphatic index	GRAVY	D-score	h-region length	SP Status
flgI	180	1.816	0.548	9	Confirmed
EXG1	179.47	1.726	0.637	8	Confirmed
OmpC	171.9	1.552	0.599	11	Confirmed
xylF	161.3	1.083	0.819	10	Confirmed
DEFB103A	155	1.436	0.906	10	Confirmed
mepA	154.74	1.353	0.875	9	Confirmed
agp	146.82	1.227	0.585	12	Confirmed
yaaX	144.35	1.261	0.589	11	Potential
degP	143.08	1.077	0.591	14	Confirmed
CCL20	142.69	1.296	0.826	12	Confirmed
IL-2	141.5	1.085	0.518	12	Confirmed
usp45	141.11	1.174	0.532	14	Confirmed
IFN-beta	139.52	1.238	0.679	12	By similarity
pelb	138.18	1.191	0.618	10	Confirmed
IFNW1	134.76	1.495	0.626	11	Confirmed
glnH	133.18	1.209	0.831	9	Confirmed

eglS	131.38	1.31	0.507	16	Confirmed
IFNA1	131.3	1.604	0.733	10	Confirmed
PhoE	130	1.195	0.541	10	Confirmed
spy	127.83	0.991	0.553	12	Confirmed
SKp	127.5	0.86	0.865	10	Confirmed
LT-IIB	127.39	1.23	0.606	9	Potential
LamB	125.2	1.332	0.583	10	Confirmed
endA	124.55	1.286	0.516	10	Potential
spa	121.94	0.511	0.523	18	Confirmed
OmpA	121.43	1.295	0.558	10	Confirmed
yfcQ	119.44	1.006	0.793	8	Potential
npr	119.26	1.041	0.554	14	Confirmed
nmpC	119.13	1.243	0.886	11	Confirmed
MlCA	119.13	1.53	0.757	10	Confirmed
pbpG	117.2	1.228	0.533	12	Confirmed
hGH	116.54	0.777	0.752	10	Potential
yraH	115.91	0.927	0.588	9	Potential
ompP	114.78	0.904	0.621	11	Confirmed
yaaI	114.78	1.365	0.824	11	Potential
MalE	113.08	1.012	0.58	10	Confirmed
LTB	111.43	0.695	0.53	9	Confirmed
rbsB	109.6	0.948	0.836	12	Potential
ALB	108.33	1.233	0.599	10	Confirmed
ypeC	107.62	1.376	0.816	10	Potential
rna	106.52	0.757	0.852	9	Confirmed
eco	103	1.265	0.519	10	Confirmed
STII	102.17	1.026	0.534	11	Confirmed
MglB	102.17	0.952	0.682	12	Confirmed
yadN	102.17	1.078	0.548	12	Potential
ppiA	98.33	1.438	0.875	12	Potential
gfcA	98.1	1.019	0.609	10	Potential
yhcF	98	0.915	0.734	9	Potential
rseB	97.83	0.865	0.571	10	Potential
efeO	94.23	0.654	0.755	12	probable
araF	93.91	0.878	0.837	10	Confirmed
gltF	90.4	1.244	0.697	11	Potential
yfjT	80.87	0.757	0.584	9	Potential
DsbC	78.5	0.842	0.842	9	Confirmed
CEACAM5	77.65	-0.121	0.533	10	Confirmed
Endo-1,4-beta-xy-lanase	70	0.871	0.6	14	Confirmed
bcsB	58.8	0.688	0.696	11	Potential
yadV	54.8	0.852	0.738	9	Potential

volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine) in an amino acid sequence. The *in silico* analysis results demonstrated that the highest aliphatic index values belonged to flgI, EXG1, OmpC, xylF and DEFB103A (180, 179.47, 171.9, 161.3 and 155 respectively).

The stability of a recombinant protein in a test tube is estimated by the instability index. A protein with instability index of below 40 is predicted as stable, whereas a value higher than 40 implies that the protein may be unstable. The instability of signal peptides alone and in the connection with IFN-beta 1b was evaluated by instability index. The analysis results demonstrated

that Spy, PhoE, MalE, DsbC and gltF were the most stable signal peptides among the 59 remaining signal peptides (–4.79, 1.44, 2.85, 5.25 and 8.52 respectively). All the signal peptides in connection with IFN-beta 1b were unstable (Table 3).

Protein solubility prediction

The solubility of IFN-beta 1b in connection with the different signal peptides was evaluated. The results showed that IFN-beta 1b was insoluble in connection with all the signal peptides (Table 3).

Prediction of the protein localization

The predicted localization site of the protein with different signal peptides is shown in Table 4. The *in silico* analysis results indicated that the final localization site of most of the signal peptides in fusion with IFN-beta was in the outer membrane space. Furthermore, final subcellular localization of IFN-beta 1b with 9 signal peptides such as flgI, glnH and etc. was predicted to be inside the periplasmic space. However, only the final predicted localization for spa peptide was extracellular.

DISCUSSION

Computational methods are being used in the large variety of biological fields in order to decrease the costs and increase the accuracy of experimental research (Zamani *et al.,* 2015). The aggregation and misfolding can be occurring as the result of a high expression level of heterologous proteins during intracellular expression (Baradaran *et al.,* 2013). Inserting a signal peptide at the N-terminus of the DNA sequence of these proteins caused secretion of the proteins into the *E. coli* periplasmic space and eventually solved the problem (Zamani *et al.,* 2015). For a successful secretion, there are different factors that should be carefully balanced during the secretory pathway. One of the most important factors for production of recombinant heterologous protein in the prokaryotic system is the signal peptide. Therefore, the physicochemical and structural features of a signal peptide are an important property in the functionality of secretion. For this purpose, various computational tools have been applied to predict and characterize the physicochemical properties of the signal peptides. They compute different features like the number of amino acids and the physicochemical properties of a signal peptide such as molecular weight, isoelectric point, GRAVY, aliphatic index and instability index (Baradaran *et al.,* 2013).

In this study, different signal peptides including the natural IFN-beta 1b and 89 signal peptides were evaluated with different bioinformatics tools. The theory behind the suitability of the eukaryotic signal peptides for prokaryotic expression in the current study is that the Sec-dependent proteins (SPDs) are homologous for both mentioned groups, and in SPDs, the homology between the translocating machinery ensures the cross-kingdom signal peptide compatibility. On the other hand, some eukaryotic sequences can be suitable for TAT purely due to a sequence convergence/chance because of the presence of twin arginine that has somehow evolved. Surprisingly, this study results indicated that EXG1 which belongs to *Saccharomyces cerevisiae* ranked at the second place as proper SPs to fuse with IFN-beta 1b. And theoretically, it is one of the most suitable signal peptides for translocating IFN-beta 1b protein into the periplasmic space of *E. coli*. The signal peptide of IFN-beta, Human growth hormone and Interleukin-2 signal peptides were used because of their ability to contribute to the secretion of recombinant protein to periplasmic space in other studies (Dalton & Barton, 2014; Zamani *et al.,* 2015).

The results of SignalP 4.1 analysis showed the discrimination between secreted and non-secreted sequences using signal peptide amino acid sequence in combination with the IFN-beta 1b protein. SignalP server has the capability to identify the main regions of the signal peptides. The C and S-score could recognize cleavage sites and signal peptide positions, respectively. Y-score is derived from the C and S-score resulting in the more precise prediction of the cleavage sites than the raw C-score. The average of the S-score is the S-mean. D-score is the average of the S-mean and Y-max which indicates a primary distinction between secretory and non-secretory proteins. *In silico* evaluation results indicated that the signal peptidase cannot recognize the cleavage sites of 23 signal peptides in connection with IFN-beta 1b in *E. coli* host. Therefore, they were not predicted as an appropriate choice for IFN-beta 1b secretion. On the other hand, the other 67 signal peptides can be used for expression of the protein in *E. coli*. Moreover, the result of SignalP analysis demonstrated that the D-score values of 8 signal peptides such as DsbA and AZU1 were higher than 0.5, so these signal peptides were suitable signal peptides for further analysis. More so, they have multiple cleavage sites for signal peptides enzymes or their cleavage sites were inappropriately located. So, it was concluded that using these signal peptides in fusion with IFN-beta 1b protein can cause problems after expression in *E. coli* cytoplasm and it is strongly possible that the final three-dimensional structure of IFN-beta 1b protein will be changed. And this prominent change might decrease the functionality, potency and efficiency of the IFN-beta 1b protein. Therefore, these 8 signal peptides were eliminated from the further analysis too.

It was suggested that by increasing the hydrophobicity levels and length of the h-region, the rate of the protein secretion could be improved (Chen *et al.,* 1996). The hydrophobicity levels of the signal peptides can be indicated by aliphatic index and GRAVY (Table 3). An efficient signal peptide cleavage in the c-region considerably influences the protein secretion levels. There is also a famous rule in the c-region called (-3, -1) or AXA motif. According to this rule, the amino acids at positions of -1 and -3 relative to the cleavage site must be small and neutral like alanine, glycine, and serine (Choi & Lee, 2004). In contrast, there were large bulky residues at the position of -2 that would not fit into either the -3 or -1 position (Pratap & Dikshit, 1998). Most of the signal peptides in this study have AXA motif in their cleavage sites (Table 2). The most important parameter that should be considered to select the most appropriate sequences is hydrophobicity (aliphatic index, GRAVY and h-region length). Therefore, data were sorted with priority of aliphatic index, GRAVY, h-region length and D-scores respectively. The results were presented in Table 5.

According to this statistical sorting of the main parameters, flgI, EXG1, OmpC, xylF and DEFB103A were considered as the most effective signal peptides, respectively. In contrast, yadV, bcsB, Endo-1, 4-beta-xylanase, CEACAM5 and DsbC showed a weak level of the features needed for the secretion process. However, the status of most of the signal peptides was confirmed.

Some signal peptides were applied by the researchers for the secretory production of recombinant proteins in *E. coli* such as PhoA, OmpA, PelB, etc (Choi & Lee,

2004). Previous studies reported the high yield of IFN-beta 1b secretion in the presence of PelB (Morowvat *et al.,* 2014) and IFN-beta 1b signal peptides (Krishna Rao *et al.,* 2009). Moreover, using PelB signal peptide in combination with IFN-beta 1b facilitated the expression of the protein fully in the periplasmic space (Mobasher *et al.,* 2016).

ProtComp, from Softberry, Inc., is an online bioinformatics tool that could predict protein localization, including extracellular proteins, using a combination of neural networks methods and sequence homology. ProtCompB combines different methods of protein localization prediction. For Gram-negative bacteria proteins, four locations are predicted: Cytoplasmic, Membrane (outer and inner), Periplasmic and Extracellular (secreted).

On the other hand, translocating the protein of interest into the periplasmic space or into the culture medium has several advantages when compared with the lower content of bacterial proteins (Forouharmehr *et al.,* 2018). In an experimental investigation by Steiner *et al.* (2006), they evaluated 10 different SPs substitution effects at N-terminal of the protein of interest (POI) for translocation of polypeptides through the cytoplasmic membrane into the periplasm of Gram-negative bacteria. This study results indicated that the substitution of SPs improved the enrichment of phage display selection from 10 fold to more than 1000 fold per each round of panning (Steiner *et al.,* 2006). Klatt and Konthur (Klatt & Konthur, 2012) accomplished SPs trimming for optimizing secretory expression of recombinant protein in *Leishmania tarentolae*. They applied *in silico* approach and used SignalP server to identify the most potent SPs. To evaluate the signal peptide cleavage site and changes of expression rate, SPs were N-terminally linked to POI. The obtained results demonstrated the importance of SPs optimization for efficient secretory expression of recombinant proteins. These results indicated that minor modifications in SPs structure, based on *in silico* investigations, increased the yield of recombinant protein secretory production (Klatt & Konthur, 2012).

Purification of the expressed recombinant protein which is translocated into the periplasm compartment, enables not only the provision of downstream processing much easier than cytosolic production but also could decrease the processing cost and running time, too. Hence, due to the reducing impurity of different cellular components and the circumvention of proteolytic degradation by intracellular proteases, isolation and purification of the over-expressed products which is translocated into the periplasm compartment can be much simplified. Overall, the obtained results showed that the final localization of IFN-beta 1b protein in combination with most of the signal peptides was in the periplasmic space of *E. coli*. However, only in combination with Spa signal peptide IFN-beta 1b was predicted to be found in the culture media (extracellularly).

CONCLUSIONS

Nowadays, the emergence of *in silico* approaches such as artificial neural network, computational biology, bioinformatics and data analysis using mathematical language in theoretical biology accelerated the process of analyzing SPs for the production of pharmaceutical recombinant proteins. Moreover, it reduced the costs of the expression and purification of recombinant proteins as well the time required for the process. So, predicting the best SPs by *in silico* approach would help biologist and protein engineers to accelerate and facilitate the vital projects. In this research, in order to select an optimal signal peptide, which is compatible with the Interferon-beta 1b recombinant secretory protein, we applied *in silico* approach to build a computational prediction model for virtual screening. In conclusion and with regards to the obtained results of the *in silico* analysis, the flgI, EXG1 and OmpC signal peptides seemed good candidates for the secretion of IFN-beta into the proper location. Therefore, they can be used for further experimental analysis of IFN-beta 1b secretion in the future. Furthermore, the test that was developed in this research could be useful and applied by researchers for evaluating and controlling various SPs in combination with other pharmaceutical recombinant proteins, to verify and introduce the best SPs for successful scale-up of the production of the pharmaceutical recombinant protein of choice.

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgements

The authors would like to especially thank Vice Chancellor for Research and Technology, University of Guilan, for providing necessary facilities and support to optimize and complete this study.

REFERENCES

Babaeipour V, Shojaosadati SA, Maghsoudi N (2013) Maximizing production of human interferon-γ in HCDC of recombinant *E. coli*. *Iran J Pharm Res* **12**: 563–572. PMID: 24250663

Bagos PG, Nikolaou EP, Liakopoulos TD, Tsirigos KD (2010) Combined prediction of Tat and Sec signal peptides with hidden Markov models. *Bioinformatics* **26**: 2811–2817. doi: 10.1093/bioinformatics/btq530

Baneyx F (1999) Recombinant protein expression in *Escherichia coli*. *Curr Opin Biotechnol* **10**: 411–421. PMID: 10508629

Baneyx F, Mujacic M (2004) Recombinant protein folding and misfolding in *Escherichia coli*. *Nat Biotechnol* **22**: 1399–1408. doi: 10.1038/nbt1029

Baradaran A, Sieo CC, Foo HL, Illias RM, Yusoff K, Rahim RA (2013) Cloning and *in silico* characterization of two signal peptides from *Pediococcus pentosaceus* and their function for the secretion of heterologous protein in *Lactococcus lactis*. *Biotechnol Lett* **35**: 233–238. doi: 10.1007/s10529-012-1059-4

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**: 783–795. doi: 10.1016/j.jmb.2004.05.028

Chen H, Kim J, Kendall DA (1996) Competition between functional signal peptides demonstrates variation in affinity for the secretion pathway. *J Bacteriol* **178**: 6658–6664. PMID: 8955279

Choi J, Lee S (2004) Secretory and extracellular production of recombinant proteins using *Escherichia coli*. *Appl Microbiol Biotechnol* **64**: 625–635. PMID: 14966662

Cohen B, Parkin J (2001) An overview of the immune system. *Lancet* **357**: 1777–1789. doi: 10.1016/S0140-6736(00)04904-7

Dalton AC, Barton WA (2014) Over-expression of secreted proteins from mammalian cell lines. *Protein Science: A Publication of the Protein Society* **23**: 517–525. doi: 10.1002/pro.2439

Emanuelsson O, Brunak S, Von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* **2**: 953–971. doi: 10.1038/nprot.2007.131

Forouharmehr A, Nassiri M, Ghovvati S, Javadmanesh, A (2018) Evaluation of different signal peptides for secretory production of recombinant bovine pancreatic ribonuclease A in gram negative bacterial system: an *in silico* study. *Curr Proteomics* **15**: 24–33. doi: 10.2174/1570164614666170725144424

Gasteiger E, Bairoch A, Sanchez JC, William, KL, Wilkins MR, Appel RD, Hochstrasser DF (2005) Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol* **112**: 531–552. doi: 10.1385/1-59259-890-0:571

Gupta SK, Shukla P (2016) Advanced technologies for improved expression of recombinant proteins in bacteria: perspectives and applications. *Crit Rev Biotechnol* **36**: 1089–1098. doi: 10.3109/07388551.2015.1084264

Gross G, Mayr U, Stüber D, Westphal W, Zacher M, Frank R, Blöcker H (1985) Stabilisation of human interferon beta synthesis in *Escherichia coli* by zinc ions. *Biochim Biophys Acta* **825**: 207–213. PMID: 3890952

Haeuptle MT, Flint N, Gough NM, Dobberstein B (1989) A tripartite structure of the signals that determine protein insertion into the endoplasmic reticulum membrane. *J Cell Biol* **108**: 1227. PMID: 2784443

Klatt S, Konthur Z (2012) Secretory signal peptide modification for optimized antibody-fragment expression-secretion in *Leishmania tarentolae*. *Microb Cell Fact* **11**: 97. doi: 10.1186/1475-2859-11-97

Klee EW, Ellis LB (2005) Evaluating eukaryotic secreted protein prediction. BMC *Bioinformatics* **6**: 256. doi: 10.1186/1471-2105-6-256

Kolb-Maurer A, Goebeler M, Maurer M (2015) Cutaneous Adverse Events Associated with Interferon-beta Treatment of Multiple Sclerosis. *Int J Mol Sci* **16**: 14951–14960. doi: 10.3390/ijms160714951

Krishna Rao DV, Ramu CT, Venkateswara Rao J, Narasu ML, Bhujanga Rao AKS (2009) Cloning, High Expression and Purification of Recombinant Human Intereferon-β-1b in *Escherichia coli*. *Appl Biochem Biotechnol* **158**: 140–154. doi: 10.1007/s12010-008-8318-9

Magnan CN, Randall A, Baldi P (2009) SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics* **25**: 2200–2207. doi: 10.1093/bioinformatics/btp386

Marziniak M, Meuth S (2014) Current perspectives on interferon Beta-1b for the treatment of multiple sclerosis. *Adv Ther* **31**: 915–931. doi: 10.1007/s12325-014-0149-1

Mergulha˜o FJM, Summers DK, Monteiro GA (2005) Recombinant protein secretion in *Escherichia coli*. *Biotechnol Adv* **23**: 177–202. doi: 10.1016/j.biotechadv.2004.11.003

Mobasher MA, Ghasemi Y, Montazeri-Najafabady N, Tahzibi A (2016) Expression of recombinant IFN beta 1-b a comparison between soluble and insoluble forms. *Minerva Biotecnologica* **28**: 39–43

Morowvat MH, Babaeipour V, Rajabi-Memari H, Vahidi H, Maghsoudi N (2014) Overexpression of recombinant human beta interferon (rhIFN-beta) in periplasmic space of *Escherichia coli*. *Iran J Pharm Res* **13**: 151–160. PMID: 24711841

Nielsen H, Krogh A (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc Int Conf Intell Syst Mol Biol* **6**: 122–130. PMID: 9783217

Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8**: 785–786. doi: 10.1038/nmeth.1701

Pratap J, Dikshit K (1998) Effect of signal peptide changes on the extracellular processing of streptokinase from *Escherichia coli*: requirement for secondary structure at the cleavage junction. *Mol Gen Genet MGG* **258**: 326–333. PMID: 9648736

Ramanan RN, Tik WB, Memari HR, Azaman SNA, Ling TC, Tey BT, et al. (2010) Effect of promoter strength and signal sequence on the periplasmic expression of human interferon-a2b in *Escherichia coli*. *Afr J Biotechnol* **9**: 285–292

Runkel L, Meier W, Pepinsky RB, Karpusas M, Whitty A, Kimball K, et al. (1998) Structural and functional differences between glycosylated and non-glycosylated forms of human interferon-beta (IFN-beta). *Pharm Res* **15**: 641–649. PMID: 9587963

Sen GC, Lengyel P (1992) The Interferon system. *J Biol Chem* **267**: 5017–5020

Sørensen PS (2010) Interferon-beta-1a: Therapeutic effects, tolerability, current and future status in multiple sclerosis. *Hot Topics Neurol Psychiatr* **3**: 7–16. doi: 10.4147/HTN-100907

Steiner D, Forrer P, Stumpp MT, Pluckthun A (2006) Signal sequences directing cotranslational translocation expand the range of proteins amenable to phage display. *Nat Biotechnol* **24**: 823–831. doi: 10.1038/nbt1218

Talele TT, Khedkar SA, Rigby AC (2010) Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Curr Top Med Chem* **10**: 127–141. PMID: 19929824

Ventura S, Villaverde A (2006) Protein quality in bacterial inclusion bodies. *Trends Biotechnol* **24**: 179–185. doi: 10.1016/j.tibtech.2006.02.007

Yoon S, Kim S, Kim JF (2010) Secretory production of recombinant proteins in *Escherichia coli*. *Recent Pat Biotechnol* **4**: 23–29. PMID: 20201800

Zamani M, Nezafat N, Negahdaripour M, Dabbagh F, Ghasemi Y (2015) *In Silico* evaluation of different signal peptides for the secretory production of human growth hormone in *E. coli*. *Int J Pept Res Ther* **21**: 261–268. https://doi.org/10.1007/s10989-015-9454-z

Zhang C, Marcia M, Langer JD, Peng G, Michel H (2013) Role of the N-terminal signal peptide in the membrane insertion of *Aquifex aeolicus* F1F0ATP synthase c-subunit. *FEBS J* **280**: 3425–3435. doi: 10.1111/febs.12336