International Journal of Peptide Research and Therapeutics (2019) 25:1279–1290
https://doi.org/10.1007/s10989-018-9775-9

In silico Analysis of Different Signal Peptides for the Excretory Production of Recombinant NS3-GP96 Fusion Protein in *Escherichia coli*

Shiva Mohammadi[1] · Zohreh Mostafavi-Pour[2] · Younes Ghasemi[1,4] · Mahdi Barazesh[1] · Soudabeh Kavousi Pour[1] · Amir Atapour[1] · Pooneh Mokarram[2,3] · Mohammad Hossein Morowvat[1,4]

Accepted: 6 October 2018 / Published online: 10 October 2018

Abstract

Escherichia coli is one of the simplest hosts which is widely being used to express heterologous proteins. However, without appropriate signal peptide, this host cannot be applied for secretory proteins. Secretory production of recombinant proteins in *E. coli* has been an issue of interest because of its diverse advantages including cost and time savings, as well as reduction of endotoxin. NS3 from hepatitis C virus (HCV) was chosen as an antigen for vaccine development against HCV virus infections and it connected to gp96 as an adjuvant for stimulating Toll-like receptors (TLRs) to stimulate cytokines secretion by T cells. It was successfully produced in *E. coli* without using signal peptide previously. In this study, in order to increase the expression level of recombinant NS3-gp96 fusion protein (rNS3-gp96) in periplasmic space, we selected a series of signal peptides. Therefore, to foretell the best signal peptides for expression of NS3-gp96 recombinant protein in *E. coli*, 52 signal peptides from gram-negative bacteria were chosen and the most important physicochemical features of them were investigated. Therefore, n, h and c regions and signal peptide probability of them were evaluated by signalP software "version 4.1", and physicochemical features were assessed by ProtParam and PROSO II tools. Eventually, prsK protein, outer membrane pore protein E (*pho*E), and fimbrial adapter papK protein were determined as the best candidates for the secretory production of rNS3-gp96 in *E. coli* in our study (with D score 0.899, 0.806, 0.797, respectively). Although, in the experimental investigation, should be considered other influencing parameters.

Keywords Bioinformatics · *E. coli* · NS3 -gp96 · Signal peptides

Introduction

Hepatitis C virus (HCV) infection leads to acute and chronic liver diseases in humans such as cirrhosis, chronic hepatitis (Atapour et al. 2017). It is one of the major health problems that has been infected about 200 million people all over the world, and the majority of HCV exposed individuals become steadily unhealthy (Alter et al. 1989). HCV is a single-stranded, RNA Virus that has positive polarity and, is encoded a single open reading frame. Upon translation, the polyprotein is processed by both viral and cellular proteases into individual nonstructural and structural proteins. Although; there is no available vaccine against HCV at the moment HCV, as is one of the protein molecules encoding with RNA, can process into at least ten distinct structural proteins, for instance, C, E1 and E2 and nonstructural proteins such as NS2, NS3, NS4A, NS4B, NS5A and NS5B (Simmonds 2013). Each of them is considered as a potential

✉ Pooneh Mokarram
mokaram2@gmail.com

✉ Mohammad Hossein Morowvat
mhmorowvat@sums.ac.ir

[1] Department of Medical Biotechnology, School of Advanced Medical Sciences and Technologies, Shiraz University of Medical Sciences, Shiraz, Iran

[2] Department of Biochemistry, School of Medicine, Shiraz University of Medical Science, Shiraz, Iran

[3] Colorectal Research Center, Shiraz University of Medical Sciences, Shiraz, Iran

[4] Pharmaceutical Sciences Research Center, Shiraz University of Medical Sciences, P.O. Box 71345-1583, Shiraz, Iran

target for screening of antiviral compounds. Efforts done for developing HCV vaccine have been hindered by several factors including the prone to high-error replication of HCV, lack of suitable animal models and the absence of well-established in vitro knowledge of protective immunity (Singh and Raghava 2001). Novel vaccines are based on molecular technology for eliciting a proper immune response against HCV, including both broadly neutralizing antibodies and effective T-cell response (Naika et al. 2015). Proteins such as NS3, because of stimulation of strong immunity and the existence of conserved epitopes, are attractive for vaccine design; several studies have now shown that T-cell immune responses against NS3 associate with resolution of the infection. Despite the advantages and safety of the recombinant protein vaccines, other strategies to improve their immunogenicity are needed (Pouriayevali et al. 2016). Heat shock proteins (HSPs) facilitate cellular immune responses to antigenic peptides or proteins bound to them. In the present study, we used (HSP gp96) as an adjuvant for creating fusion protein as a candidate vaccine for HCV disease so designed NS3-gp96 fusion protein by connecting the N-terminal NS3 to the N-terminal gp96.

The Prokaryotic system, in particular, *Escherichia coli* is being employed for production of recombinant protein, in fact, *E. coli* is one of the best hosts for the expression of recombinant proteins since not only is less expensive but also is very simple to apply (Idicula-Thomas and Balaji 2005; Magnan et al. 2009). Although NS3-gp96 as a recombinant protein can be expressed in *E. coli*, an important issue is to be considered here; High-level production of functional and soluble recombinant proteins is the major purpose of their expression in bacterial host. Recombinant proteins can be expressed in *E. coli* as intracellular inclusion bodies; but secretion into the extracellular compartment is a priority as it simplifies downstream purification processes, protects heterologous proteins from proteolysis by cytoplasmic or periplasmic proteases, decrease endotoxin levels and contamination of the product by others host proteins, also improve biological activity and solubility (Gottesman 1996). In *E. coli*, proteins usually do not secrete into the extracellular compartment except for a few numbers of proteins. Although, small proteins are commonly released into the culture medium depends on the characteristics of signal peptide sequences and proteins (Choi and Lee 2004; Tong et al. 2000). So, we need a tool to direct NS3-gp96 to extracellular compartment of *E. coli*. In gram-negative bacteria, there are three fate for targeting of expressed protein, including secretion into periplasmic compartment, secretion into outer membrane and extracellular release from outer membrane by common secretory pathway (Desvaux et al. 2004).

The best approach for transfer of rNS3-gp96 to extracellular compartment is using a suitable signal peptide. In fact, in bacteria signal peptides can translocate proteins to periplasmic circumstance by different pathways. In general, there are three main pathways in bacteria for translocation of a secretory protein to periplasmic circumstance that have been classified to the universal secretion pathway (Sec-pathway); the signal recognition particle pathway (SRP pathway) and the twin-arginine translocation (TAT-pathway). Furthermore, among this TAT pathway can transfer folded proteins to periplasmic compartment (Kumari and Chaurasia 2015), whereas Sec and SRP pathways transfer unfolded proteins to periplasmic compartment (De Marco 2009; Natale et al. 2008). Therefore, the researchers are widely using these tools to express secretory protein in which the identification of suitable SP for each protein appears very indispensable to express (De Marco 2009; Gardy and Brinkman 2006; Müller and Bernd Klösgen 2005). There have been some differences particularly in the length and composition of SPs, but in general, any SP is a N-terminal peptide with three key regions; N-terminal region (n-region), a hydrophobic region (h-region) and a cleavable site (c-region). The h-region generally has 7–15 residues while n and c regions have 3–5 residues in length. N and h-regions play a critical role in transferring recombinant proteins into periplasmic space (Emanuelsson et al. 2007; Zimmermann et al. 2011), while c-region plays a vital role as a cleavable site which can be distinguished by signal peptidase enzyme. In spite of SPs key role in the secretion of heterologous proteins, there have been no universal principles to detect them (Emanuelsson et al. 2007; Zhang et al. 2013). In recent decades with the increase in biological tools, biologists are mostly applying method such as machine learning to evaluate the data (Ezziane 2006), as in today, bioinformatics tools have attracted unique attention in biology, because they not only decline the high cost of experiments also provide trustworthy results (Zhang et al. 2013). Our aim was to identify a suitable SP for secretory expression of NS3-gp96 protein in *E. coli*, therefore, most important features of 52 numbers of SPs from gram-negative bacteria were evaluated and compared using in silico methods and the best of which are introduced for experimental applications.

Materials and Methods

Signal Sequence Collection and Study Design

In this study, amino acid sequences of 52 numbers of SPs were taken from national center of biotechnology information (NCBI) as shown in Table 1. In silico methods such as machine learning techniques were employed to evaluate and characterize the collected signal sequences. Eventually, after trimming and prediction of sub-cellular localization site and also after excluding inappropriate signal peptides, the selected signal peptides were then evaluated to observe

Table 1 Amino acid sequences of bacterial signal peptides used in this study

	Full name	Signal peptide	Accession number	Source	n-region h-region n-region
1[a]	L-asparaginase 2	*AnsB*	P00805	*E. coli*	MEFFKKTALAALVMGFSGAALA
2	Beta-lactamase TEM	*Bla*	P62593	*E. coli*	MSIQHFRVALIPFFAAFCLPVFA
3[a]	Thiol:disulfide interchange protein	*DsbA*	P0AEG4	*E. coli*	MKKIWLALAGLVLAFSASA
4	Heat-labile enterotoxin B chain	*EltB*	P0CK94	*E. coli*	MNKVKFYVLFTALLSPLCAHG
5	FKBP-type peptidyl-prolyl *cis–trans* isomerase	*FkpA*	P45523	*E. coli*	MKSLFKVTLLATTMAVALHAPITFA
6[a]	Maltoporin	*LamB*	P02943	*E. coli*	MMITLRKLPLAVAVAAGVMSAQAMA
7[a]	Major outer membrane lipoprotein	*Lpp*	P69776	*E. coli*	MKATKLVLGAVILGSTLLAG
8[a]	Maltose-binding periplasmic protein	*MalE*	P0AEX9	*E. coli*	MKIKTGARILALSALTTMMFSASALA
9[a]	D-galactose-binding periplasmic protein	*MglB*	P0AEE5	*E. coli*	MNKKVLTLSAVMASMLFGAAAHA
10[a]	Outer membrane protein A	*OmpA*	P0A910	*E. coli*	MKKTAIAIAVALAGFATVAQA
11	Periplasmic *appA* protein	*appA*	EHN88412	*E. coli*	MKAILIPFLSLLIPLTPQSAFA
12	Cytochrome c-type biogenesis protein	*ccmH*	AEJ57359	*E. coli*	MRFLLGVLMLMISGSALA
13	Protein *cexE*	*cexE*	WP_001687026	*E. coli*	MKKYILGVILAMGSLSAIA
14	Thiosulfate-binding protein	*cysP*	WP_033801079	*E. coli*	MAVNLLKKNSLALVASLLLAGHVQA
15	Dr hemagglutinin structural subunit	*draA*	P24093	*E. coli*	MKKLAIMAAASMVFAVSSAHA
16	Thiol:disulfide interchange protein dsbD	*dsbD*	WP_058033897	*E. coli*	MAQRIFTLILLLCSTSVFA
17	Thiol:disulfide interchange protein *dsbG*	*dsbG*	ETJ26382	*E. coli*	MLKKILLLALLPAIAFA
18	K88 fimbrial protein AD	*faeG*	WP_001380745	*E. coli*	MKKTLIALAIAASAASGMAHA
19	Iron(III) dicitrate-binding periplasmic protein	*fecB*	KDW96130	*E. coli*	MLAFIRFLFAGLLLVISHAFA
20	F107 fimbrial protein	*fedA*	ACY05963	*E. coli*	MKRLVFISFVALSMTAGSAMA
21	F41 fimbrial protein	*FimF41a*	AAA23421	*E. coli*	MKKTLIALAVAASAAVSGSVMA
22	Flagellar P-ring protein	*flgI*	EFJ97486	*E. coli*	MVIKFLSALILLLVTTAAQA
23	Protein transport protein hofQ	*hofQ*	EDV85112	*E. coli*	MKQWIAALLLMLIPGVQA
24	Outer-membrane lipoprotein carrier protein	*lolA*	WP_016247003	*E. coli*	MKKIAITCALLSSLVASSVWA
25	Lipopolysaccharide export system protein *lptA*	*lptA*	EHV68281	*E. coli*	MKFKTNKLSLNLVLASSLLAASIPAFA
26	Penicillin-insensitive murein endopeptidase	*mepA*	WP_001043836	*E. coli*	MNKTAIALLALLASSVSLA
27	Nickel-binding periplasmic protein	*appA*	WP_021568845	*E. coli*	MLSTLRRTLFALLACASFIVHA
28	Cytochrome c-552	*nrfA*	CTU12334	*E. coli*	MTRIKINARRIFSLLIPFFFFTSVHA
29	Outer membrane protease *ompP*	*ompP*	WP_041124237	*E. coli*	MQTKLLAIMLAAPVVFSSQEASA
30	Outer membrane protein W	*ompW*	EKW81199	*E. coli*	MKKLTVAALAVTTLLSGSAFA
31	Fimbrial adapter *papK*	*papK*	WP_020239066	*E. coli*	MIKSTGALLLFAALSAGQAIA
32	D-alanyl-D-alanine endopeptidase	*pbpG*	WP_032295491	*E. coli*	MPKFRVSLFSLALMLAVPFAPQAVA
33	Alkaline phosphatase	*phoA*	AAA24362	*E. coli*	MKQSTIALALLPLLFTPVTKA
34	Outer membrane pore protein E	*phoE*	EIO69468	*E. coli*	MKKSTLALVVMGIVASASVQA
35	Protein *prs*K	*prsK*	EQN57820	*E. coli*	MIKSTGALLLFAALSAGQAMA
36	Phage shock protein E	*pspE*	KDU08780	*E. coli*	MFKKGLLALALVFSLPVFA
37	Protease 3	*ptrA*	EIL66839	*E. coli*	M*PRS*TWFKALLLLVALWAPLSQA
38	S-fimbrial adhesin protein sfaS	*sfaS*	WP_021524832	*E. coli*	MKLKAIILATGLINCIAFSAQA
39	Taurine-binding periplasmic protein	*tauA*	WP_032218149	*E. coli*	MAISSRNTLLAALAFIAFQAQA
40	Thiamine-binding periplasmic protein	*thiB*	WP_032307836	*E. coli*	MLKKCLPLLLLCTAPVFA
41	Periplasmic protein torT	*torT*	WP_029487908	*E. coli*	MRVLLFLLLSLFMLPAFS

Table 1 (continued)

	Full name	Signal peptide	Accession number	Source	n-region h-region n-region
42	sn-glycerol-3-phosphate-binding periplasmic protein *ugpB*	*ugpB*	ELJ77555	*E. coli*	MKPLHYTASALALGLALMGNAQA
43	D-xylose-binding periplasmic protein	*xylF*	EOV74805	*E. coli*	MKIKNILLTLCTSLLLTNVAAHA
44	Uncharacterized protein *yfeK*	*yfeK*	WP_053887217	*E. coli*	MKKIICLVITLLMTLPVYA
45	UPF0379 protein *yhcN*	*yhcN*	WP_058905387	*E. coli*	MKIKTTVAALSVLSVLSFGAFA
46	Uncharacterized protein yncJ	*yncJ*	EYB53638	*E. coli*	MFTKALSVVLLTCALFSGQLMA
47	UPF0482 protein *ynfB*	*ynfB*	WP_000705210	*E. coli*	MKITLSKRIGLLAILLPCALALSTTVHA
48	Zinc resistance-associated protein	*zraP*	WP_042082503	*E. coli*	MKRNTKIALVMMALSAMAMGST-SAFA
49	–	*ASPG_ERWCH*	P06608	*Erwinia chrysanthemi*	MERWFKSLFVLVLFFVFTASA
50	–	*AGAR_ALTAT*	P13734	*Alteromonas atlantica*	MLKVIPWLLVTSSLVAIPTYIHA
51	Chaperone protein *Caf1M*	*Caf1M*	P26926	*Yersinia pestis*	MILNRLSTLGIITFGMLSFAPGPPPGP-PRVS
52	Pectate lyase 2	*Pel2*	Q6CZT3	*Erwinia carotovora*	MKYL*LPTA*AAGLLLLAAQPAMA

In amino acid sequence of SPs, contrary to h region, n and c regions have been shown in red color

[a]*E. coli* (strain K12)

whether they have gained high level of secretory expression of rNS3-gp96 protein in *E. coli*.

In Silico Prediction of n, h and c Regions and Signal Peptide Probability

In order to predict n, h and c regions and signal peptide probability SignalP server version 4.1 (http://www.cbs.dtu.dk/services/SignalP/) was used. These are based on a combination of several artificial neural networks and hidden Markov models (Bendtsen et al. 2004; Petersen et al. 2011). In order to use the server, each SP was connected to N-terminal of NS3-gp96 amino acid sequence and methionine residues were inserted between each SP and NS3-gp96 amino acid sequence.

Physico-Chemical properties and Sub-Cellular Localization of Signal Peptides

In silico study of physicochemical features of signal peptides such as amino acid composition, molecular weight, theoretical PI, Aliphatic Index, solubility index, grand average of hydropathicity (GRAVY) and positively and negatively charged residues were all evaluated by ProtParam server (Walker 2005) (http://web.expasy.org/cgi-bin/protparam/protparam). Prediction of protein solubility upon expression in *E. coli* was done by the PROSO II software at http://mips.helmholtzmuenchen.de/prosoII. This server uses minute differences between soluble proteins from TargetDB and PDB and undisputedly insoluble proteins from TargetDB, and also literature mining for performing the predictions. In addition, a solubility score between 0 and 1 with a default threshold of 0.6 is given (Smialowski et al. 2012). PROSO II has the maximum prediction accuracy percentage (64.35) compared to some other similar servers, such as CCSOL (54.20), SOLpro (59.95), PROSO (57.85), and recombinant protein solubility (51.4). More importantly, it can be used for heterologous proteins in *E. coli* (Chang et al. 2013). The solubility tests were performed for SPs linked to rNS3-gp96. In order to sort SPs based on the secretion properties, PRED-TAT server (Bagos et al. 2010) was used (http://www.compgen.org/tools/PRED-TAT/submit). PRED-TAT operates based on hidden Markov models (Bagos et al. 2010). For study of signal peptides sub-cellular location, ProtComp server was used. It merges several methods of protein localization prediction, neural networks-based prediction; direct comparison with updated base of homologous proteins of known localization; and also, comparisons of pentamer distributions calculated for query and DB sequences (http://www.softberry.com). Average accuracy of ProtCompB is 86–100% which depends on space of sub-cellular location, for example, this accuracy in membrane is 100% but in extracellular is 86%. In order to apply PROSO II, PRED-TAT and ProtCompB, each SP was linked to N-terminal of rNS3-gp96 amino acid sequence

so that methionine residues were put in between SPs and rNS3-gp96 amino acid sequence (Magnan et al. 2009; Mousavi et al. 2017; Zamani et al. 2015).

Results

In Silico Prediction of n, h and c-Regions and Signal Peptide Probability

The results showed that SPs' D-scores were between 0.540 (*ASPG_ERWCH*) and 0.929 (*lptA*) (Table 2). The most significant parameter for the diagnosis of a SP is the discriminating score (D-score) which is usually described with a cut-off value of 0.5. Actuality only when the SP has a D-score more than 0.50, it is considered. The in silico analysis results of SignalP server has also shown that the highest D-score belonged to *lptA, pel2, flgI* and *ptrA*, respectively. Having D-scores < 0.5, Signal peptides *AGAR_ALTAT, Lpp* and *Caf1M* were not suitable candidates for the excretion of rNS3-gp96 protein. These signal peptides were deleted among other signal peptides. Then, next analyses were performed on the 49 remaining signal peptides.

As it was mentioned before that n and h regions are important in cleaving SPs from protein, therefore a reliable SP sequence should have obvious n, h and c regions. All of the collected signal peptides have the n-region, h region and c region length between 4 and 11, 8 and 14, and 3 and 13 amino acids respectively. All SP sequences in our study (except three of them) not only had D-score more than 0.50 but also contained obvious n, h and c regions.

Physico-Chemical Properties of Signal Peptides

The in silico results exhibited that the studied SPs length variation was between 17 (*dsb*G) and 28 (*ynf*B) amino acid, the lowest and the highest Mw belonged to *dsb*G (3167.8) and *ynf*B (2948.7), respectively (Table 3). The results also demonstrated that the range of Net positive charge was between 0 and 4, whereas the range of PI was between 5.75 (*omp*P) and 12.3 (*nrf*A). The grand average of hydropathy score (GRAVY) is used to compare SPs overall hydropathy, in fact, this parameter is defined as the sum of hydropathy of amino acids (Zamani et al. 2015). As it is observed the lowest GRAVY belonged to *ugp*B (0.622) and the highest GRAVY belonged to *fec*B (2.076). Another factor used to show hydrophobicity is aliphatic index, this factor is defined as the relative volume occupied by aliphatic side chain in an amino acid sequence. According to in silico outcome, the variation in range of aliphatic index was between 79.23 (*zra*P) and 207.06 (*dsb*G). Instability index evaluated as another factor too, in general when instability is more than 40, possible proteins is considered unstable, whereas when instability is < 40, it shows the stability of the protein (Zamani et al. 2015). The instability of signal peptides alone and also in connection with rNS3-gp96 was evaluated by instability index. The in silico analysis results showed that the variation in range of instability index was between − 2.6 (*pap*K) and 65.64 (*thi*B). Instability index of 11 signal peptides including, *bla, lam*B, *app*A, *app*A, *omp*P, *pbpG, pho*A, *ptr*A, *thi*B, *yfeK* and *Pel*2, was more than 40, so they were predicted as unstable. in fact, the analysis results demonstrated that *pap*K (− 2.6) and *yhc*N (− 2.03) were the most stable signal peptides among the 49 studied signal peptides, respectively (The most unstable signal peptides in connection with rNS3-gp96 were *thi*B (65.6), *app*A (60.45) and *pbp*G (57.99), respectively). The PROSO II server was applied for characterization of rNS3-gp96 solubility in connection with the 49 studied signal peptides. It has been said, solubility of passenger proteins seems essential for secretion, considering that the insoluble proteins tend to aggregate in the inclusion bodies (Baneyx 1999). Considering the solubility of all the tested sequences, this criterion does not look a limiting factor in our analysis, so was not selected as a main decisive factor (Baneyx 1999; Chang et al. 2013). Overexpression of rNS3-gp96 such as other recombinant proteins in *E. coli* host leads to formation of inclusion body. The inclusion body is a bulk containing the insoluble, nonfunctional and misfolded form of heterologous proteins. To solve this problem, several strategies have been developed. The first is extracellular production of recombinant rNS3-gp96 in *E. coli* accomplished via attaching signal peptides to N-terminal or C-terminal of gene of interest. The secretory production efficacy of recombinant proteins is different. Therefore, it is essential to assess and evaluate novel signal peptides for optimum selection of proper secretion pathway that is the most effective for the production, processing and secretion of the interested protein (Singh and Panda 2005). The availability of many biological data and advances in computational techniques enable biologist users to study biological systems at different fields from design vaccine to protein engineering, which not only has confidently reduced the time and costs consuming experimental process but has also improved the accuracy of practical studies (Gholami et al. 2015; Zamani et al. 2015). Consequently, the results have indicated that all SPs connected to rNS3-gp96 protein could make a soluble protein, theoretically.

Secretion Sorting and Sub-Cellular Localization

In this study, Sec, SRP and TAT pathways were evaluated by PRED-TAT software and the results revealed that all 49

Table 2 In silico analysis of the signal peptide sequences by SignalP version 4.1

No.	Signal peptides	n-Region	h-Region	c-Region	Cleavage site	C-Score	Y-Score	S-Score	S-Mean	D-Score
1[a]	*AnsB*	1–7(7)	8–17(9)	18–22(6)	ALA	0.870	0.879	0.953	0.872	0.876
2	*Bla*	1–7(7)	8–19(12)	20–23(4)	VFA	0.684	0.591	0.601	0.499	0.557
3[a]	*DsbA*	1–3(3)	4–15(12)	16–19(4)	ASA	0.744	0.846	0.971	0.951	0.895
4	E*ltB*	1–5(5)	6–14(9)	15–21(6)	AHG	0.653	0.752	0.945	0.869	0.807
5	*FkpA*	1–6(6)	7–16(10)	17–25(9)	TFA	0.706	0.773	0.981	0.896	0.831
6[a]	*LamB*	1–7(7)	8–19(12)	20–25(6)	AMA	0.824	0.864	0.982	0.928	0.894
7[a]	*Mal*E	1–8(8)	9–20(12)	21–26(6)	ALA	0.783	0.861	0.988	0.948	0.902
8[a]	*MglB*	1–4(4)	5–17(13)	18–23(6)	AHA	0.816	0.877	0.980	0.946	0.909
9[a]	*OmpA*	1–4(4)	5–16(12)	17–21(5)	AQA	0.840	0.878	0.960	0.918	0.897
10[a]	*appA*	1–4(4)	5–16(12)	17–22(6)	AFA	0.836	0.829	0.945	0.850	0.839
11	*ccmH*	1–3(3)	4–12(9)	13–18(6)	ALA	0.803	0.726	0.780	0.654	0.699
12	*cex*E	1–4(4)	5–13(9)	14–19(6)	AIA	0.742	0.704	0.800	0.663	0.689
13	*cysP*	1–10(10)	11–19(9)	20–25(6)	VQA	0.812	0.841	0.929	0.880	0.859
14	*draA*	1–4(4)	5–15(11)	16–21(6)	AHA	0.778	0.853	0.970	0.940	0.894
15	*dsbD*	1–4(4)	5–13(9)	14–19(6)	VFA	0.843	0.756	0.738	0.666	0.722
16	*dsbG*	1–4(4)	5–14(10)	15–17(3)	AFA	0.494	0.594	0.789	0.718	0.640
17	*faeG*	1–4(4)	5–15(11)	16–21(6)	AMA	0.814	0.861	0.969	0.919	0.888
18	*fecB*	1–6(6)	7–15(9)	16–21(6)	AFA	0.667	0.592	0.692	0.525	0.567
19	*fedA*	1–4(4)	5–14(10)	15–21(7)	AMA	0.796	0.861	0.978	0.934	0.895
20	*FimF41a*	1–4(4)	5–16(12)	17–22(6)	VMA	0.885	0.900	0.978	0.928	0.913
21	*flgI*	1–4(4)	5–14(10)	15–20(6)	AQA	0.852	0.901	0.974	0.947	0.923
22	*hofQ*	1–4(4)	5–13(9)	14–18(5)	VQA	0.699	0.638	0.762	0.528	0.597
23	*lolA*	1–4(4)	5–15(11)	16–21(6)	VWA	0.815	0.876	0.973	0.938	0.905
24	*lptA*	1–11(11)	12–21(10)	22–27(6)	AFA	0.876	0.908	0.987	0.952	0.929
25	*mepA*	1–4(4)	5–13(9)	14–19(6)	SLA	0.849	0.898	0.974	0.941	0.918
26	*appA*	1–7(7)	8–16(9)	17–22(6)	VHA	0.858	0.900	0.960	0.936	0.917
27	*nrfA*	1–10(10)	11–19(9)	20–26(7)	VHA	0.614	0.577	0.649	0.523	0.557
28	*ompP*	1–4(4)	5–15(11)	16–23(8)	ASA	0.699	0.714	0.889	0.784	0.747
29	*ompW*	1–5(5)	6–15(10)	16–21(6)	AFA	0.849	0.896	0.962	0.941	0.917
30	*papK*	1–5(5)	6–14(9)	15–21(7)	AIA	0.797	0.841	0.940	0.893	0.866
31	*pbpG*	1–6(6)	7–18(12)	19–25(7)	AVA	0.750	0.812	0.985	0.920	0.863
32	*phoA*	1–5(5)	6–14(9)	15–21(7)	TKA	0.584	0.694	0.891	0.822	0.754
33	*pho*E	1–5(5)	6–15(10)	16–21(6)	VQA	0.806	0.851	0.947	0.885	0.867
34	*prs*K	1–5(5)	6–14(9)	15–21(7)	AMA	0.863	0.887	0.956	0.912	0.899
35	*psp*E	1–4(4)	5–13(9)	14–19(6)	VFA	0.833	0.761	0.793	0.693	0.736
36	*ptrA*	1–8(8)	9–17(9)	18–23(6)	SQA	0.836	0.897	0.975	0.951	0.922
37	*sfaS*	1–4(4)	5–16(12)	17–22(6)	AQA	0.754	0.816	0.958	0.879	0.845
38	*tauA*	1–7(7)	8–16(9)	17–22(6)	AQA	0.858	0.861	0.943	0.876	0.868
39	*thiB*	1–4(4)	5–12(8)	13–18(6)	VFA	0.701	0.814	0.962	0.933	0.870
40	*torT*	1–4(4)	5–13(9)	14–18(5)	AFS	0.506	0.564	0.723	0.627	0.587
41	*ugpB*	1–7(7)	8–17(10)	18–23(6)	AQA	0.854	0.861	0.930	0.871	0.866
42	*xylF*	1–6(6)	7–16(10)	17–23(7)	AHA	0.789	0.855	0.974	0.930	0.890
43	*yfeK*	1–4(4)	5–13(9)	14–19(6)	VYA	0.755	0.658	0.747	0.566	0.624
44	*yhcN*	1–6(6)	7–16(10)	17–22(6)	AFA	0.758	0.756	0.847	0.773	0.762
45	*yncJ*	1–4(4)	5–15(11)	16–22(7)	LMA	0.840	0.887	0.952	0.928	0.906
46	*ynfB*	1–10(10)	11–22(12)	23–28(6)	VHA	0.881	0.895	0.989	0.937	0.915
47	*zraP*	1–7(7)	8–18(11)	19–26(8)	AFA	0.827	0.878	0.994	0.951	0.912
48	*ASPG_ERWCH*	1–6	7–17(11)	18–21(4)	ASA	0.680	0.579	0.663	0.473	0.540
49	*Pel2*	1–3	4–17(14)	18–22(5)	AMA	0.886	0.913	0.972	0.937	0.924

In Signalp4.1 output, the C-score and S-score determine the cleavage sites and location respectively. Y-score distinct the geometric average between the C-score and a smoothed derivative of the S-score. S-mean is arithmetic average of the S-score from position 1 to location where the Y-score is the highest. D-score is the mean of the S-mean and Y-max which discriminates secretory and non-secretory proteins with cut-off value of 0.5. Signal peptides with D-score < 0.5 are determined as signal peptide

[a]*E. coli* (strain k12)

Table 3 Physico-chemical properties of the signal peptides determined by ProtParam and PROSO II

No.	Signal peptides	Amino acid length	MW (Da)	PI	Net positive charge	Charge GRAVY	Aliphatic Index	Instability	Solubility
1*	*AnsB*	22	2274.76	8.35	1	1.136	93.64	− 1.15	Soluble
2	*Bla*	23	2626.22	8.02	1	1.539	110.43	56.40	Soluble
3*	*DsbA*	19	1990.48	10.00	2	1.416	144.21	11.5	Soluble
4	E*ltB*	21	2352.88	9.19	2	0.89	111.43	31.1	Soluble
5	*FkpA*	25	2676.31	10.00	2	1.212	121.20	14.37	Soluble
6*	*LamB*	25	2545.22	11.00	2	1.332	125.2	42.97	Soluble
7*	*MglB*	23	2362.89	10.00	2	0.952	102.17	14.15	Soluble
8*	*OmpA*	21	2046.50	10.00	2	1.295	121.43	9.52	Soluble
9*	*appA*	22	2384.9	8.5	1	1.405	155.45	53.16	Soluble
10*	*ccmH*	18	1923.4	9.5	1	1.828	157.22	5.26	Soluble
11	*cex*E	19	1979.5	9.7	2	1.411	154.21	29.75	Soluble
12	*cysP*	25	2575.1	10	2	1.064	164	11.14	Soluble
13	*draA*	21	2135.6	10	2	1.162	98.1	16.49	Soluble
14	*dsbD*	19	2127.6	8	1	1.632	148.95	26.11	Soluble
15	*dsbG*	17	1839.4	10	2	2.018	207.06	33.41	Soluble
16	*faeG*	21	2027.4	10	2	1.005	112.38	11.36	Soluble
17	*fecB*	21	2350.9	9.52	1	2.076	162.86	9.52	Soluble
18	*fedA*	21	2231.7	11	2	1.29	102.38	29.55	Soluble
19	*FimF41a*	22	2090.5	10	2	1.355	124.55	15.15	Soluble
20	*flgI*	20	2116.6	8.5	1	1.935	185.5	10.64	Soluble
21	*hofQ*	18	1996.5	8.5	1	1.322	162.78	21	Soluble
22	*lolA*	21	2192.7	9.31	2	1.324	139.52	16.67	Soluble
23	*lptA*	27	2849.4	10.3	3	0.881	130.37	17.32	Soluble
24	*mal*E	26	2698.3	11.17	3	1.012	113.08	2.85	Soluble
25	*mepA*	19	1887.3	8.5	1	1.479	164.74	32.07	Soluble
26	*appA*	22	2434.9	10.35	2	1.35	137.37	60.45	Soluble
27	*nrfA*	26	3126.8	12.3	4	0.792	108.85	30.31	Soluble
28	*ompP*	23	2406.8	5.75	0	0.904	114.78	44.47	Soluble
29	*ompW*	21	2093.5	10	2	1.21	125.71	1.44	Soluble
30	*papK*	21	2047.4	8.5	1	1.39	140	− 2.6	Soluble
31	*pbpG*	25	2705.3	11	2	1.228	117.2	57.99	Soluble
32	*phoA*	21	2256.8	10	2	0.971	139.52	56.02	Soluble
33	*pho*E	21	2104.5	10	2	1.195	130	1.44	Soluble
34	*prs*K	21	2065.5	8.5	1	1.267	121.43	3.27	Soluble
35	*psp*E	19	2065.6	10	2	1.711	148.95	17.37	Soluble
36	*ptrA*	23	2613.2	11	2	0.857	131.74	51.93	Soluble
37	*sfaS*	22	2290.8	9.31	2	1.314	146.82	5.41	Soluble
38	*tauA*	22	2308.7	9.5	1	1.055	120.45	34.41	Soluble
39	*thiB*	18	1974.6	8.89	2	1.589	157.22	65.64	Soluble
40	*torT*	18	2111.7	9.5	1	2.061	173.33	26.66	Soluble
41	*ugpB*	23	2342.8	8.37	1	0.622	110.87	18.01	Soluble
42	*xylF*	23	2482	9.31	2	1.083	161.3	33.61	Soluble
43	*yfeK*	19	2163.8	9.19	2	1.742	179.47	42.39	Soluble
44	*yhcN*	22	2254.7	10	2	1.418	128.64	− 2.03	Soluble
45	*yncJ*	22	2344.9	7.98	1	1.541	128.64	15.15	Soluble
46	*ynfB*	28	2948.7	10.06	3	1.239	163.93	29.32	Soluble
47	*zraP*	26	2733.3	11.17	3	0.746	79.23	28.75	Soluble
48	*ASPG_ERWCH*	21	2539.08	8.50	1	1.352	106.67	29.64	Soluble
49	*Pel2*	22	2228.78	8.34	1	1.191	138.18	41.42	Soluble

The instability index provides an estimate of the stability of evaluated protein, Proteins with instability index < 40 is predicted as stable and above that as unstable; MW molecular weight, average isotopic masses of amino acids in the provided protein and the average isotopic mass of one water molecule. Aliphatic index: the relative volume occupied by the amino acids such as alanine, valine, isoleucine and leucine, which have an aliphatic side chain in their structure. pI isoelectric point: pKa values of amino acids. The pKa value of amino acids depends on its side chain.

Table 3 (continued)

It has an important role in defining the pH dependent characteristics of a protein. GRAVY grand average of hydropathicity: the sum of hydropathy of amino acids, increasing positive score indicates a greater hydrophobicity

**E. coli* (strain k12)

studied SPs belonged to Sec-pathway. This, in turn, could transfer the expressed rNS3-gp96 recombinant protein to different compartments. Sub-cellular localization analysis showed (by ProtCompB server) that among 49 SPs, 42 SPs can localize rNS3-gp96 in cytoplasm, four SPs can transfer this heterologous protein into extracellular space, and three SPs can localize this heterologous protein into plasma membrane (Table 4).

Discussion

NS3-gp96, as a monomeric protein, lacking disulfide bonds seems a good candidate for secretory production in *E. coli*. Considering the decisive role of SPs in directing the protein through the membrane, the selection of an appropriate SP is critical. A total number of 52 SPs were selected from several organisms, and their sequences were retrieved from the UniProt server. All 52 numbers of SPs are prokaryotic. Since the native SPs of each host may be more suitable for protein production in that microorganism, 48 SPs were selected from *E. coli* proteins. Four other SPs from other gram-negative bacteria were also chosen. TAT, Sec and SRP are the main pathways in prokaryote cells directing nascent protein to periplasmic compartment. Furthermore, these pathways operate based on signal peptide recognition, hence it is easily inferred that signal peptides play an important role in folding secretory protein in prokaryote cells (Baneyx and Mujacic 2004; Keller et al. 2012). As mentioned earlier, *E. coli* is the cheapest and simplest host to express recombinant proteins but the success in using it entirely depends on employing the suitable SPs (Rosano and Ceccarelli 2014). Consequently, the identification of suitable SPs is one of the most vital steps to produce secretory proteins as a recombinant protein in *E. coli*. Today bioinformatics tools are widely being used in different parts of biological studies largely because they reduce the cost of experiments and they also provide more exact results (Ghasemi et al. 2012; Zamani et al. 2015). As it is observed in this study, it was attempted to employ the most accurate and recent version of bioinformatics tools to predict the variety of SP features. Among various features of SP, net positive charge, aliphatic index, GRAVY, D-score, h-region length, cleavable site and sub-cellular location are more important (Table 5). Accordingly, these features were expected to make the final decision of selecting the best possible SPs. D score is the first parameter in diagnosing an SP, therefore, SPs have all been sorted on the basis of D-score. When D score is more than 0.50, a signal sequence can be considered SP (Zamani et al. 2015). Since all SPs' D-score in this study is more than 0.50, (except three of them) thereby all of them could be SP but for optimum screening, other features of selection should be considered. N-region is a crucial area in an SP which interferes translocation of a secretory protein, in fact, for maintaining its function, n-region needs a positive charge and this charge is directly linked to the existence of one or more basic residues such as lysine at the beginning of an SP (Zamani et al. 2015). It is believed that switching the basic residues with neutral or acidic residues have an impact on translocation of nascent protein because of the significant role of this positive charge in interacting between SP of nascent protein and membrane phospholipid of RER (Low et al. 2013). As the results show, the variety of net positive charge is considered between 0 and 4, thereby it seems in this stage we do not have enough justification to decide whether to select any SP since all the selected ones have appropriate net positive charge. Another important region which plays a vital role in translocation is h-region, in fact, the most important factor enabling h- region, is hydrophobicity. It has been reported this factor extremely relies on the length of h-region. In fact, the increase in the length of h-region would improve the level of hydrophobicity. Accordingly, there has not been a significant diversity in the length of SPs h-region (9–12) thereby other important factors were used such as aliphatic index and GRAVY in recognition of hydrophobicity. Aliphatic index and GRAVY are the two parameters with direct association with hydrophobicity, in fact, the boost in these parameters, lead to the increase of hydrophobicity (Low et al. 2013; Zamani et al. 2015). As it has been reported in Table 5, among 49 SPs only *zraP* has low aliphatic index (79.23) and GRAVY (0.746) while in the case of other SPs, no significant difference was observed; therefore, it seems *zraP* is not a suitable SP to express NS3-gp96 protein. C-region, particularly the three terminal residues that are also named −3, −2, −1 box, are extremely significant in detaching SPs and the secretory proteins after translocation, in fact, −3, −2, −1 boxes are recognized and cleaved by the signal peptidase. Previous studies have indicated that there are typically small or neutral residues such as alanine in −1 and −3 positions, whereas there are often big residues in −2 position which is different with the residues in −1 and −3 positions, this residue is illustrated with X (Choi and Lee 2004; Payne et al. 2012; Zamani et al. 2015). As shown in Table 3 all SPs are following this rule and are almost

Table 4 Secretion sorting and sub-cellular location of SPs

No.	Signal peptides	Type of SP	Reliability Score (%)	Cytoplasmic	Membrane	Sub-Cellular Location Score		
						Secreted (extracellular)	Periplasmic	Final prediction site
1	*AnsB*	Sec	99.9	7.5	2	0	0.5	Cytoplasmic
2	*Bla*	Sec	99.9	5.3	3	1	0.7	Cytoplasmic
3	*DsbA*	Sec	99.9	4.6	3.6	1	0.9	Cytoplasmic
4	E*ltB*	Sec	99.5	9.1	0.8	0.00	0.06	Cytoplasmic
5	*FkpA*	Sec	99.9	8.7	1	0.00	0.3	Cytoplasmic
6	*LamB*	Sec	99.9	6.4	0.4	2.4	0.7	Cytoplasmic
7	*MglB*	Sec	100	8.1	1.2	0.06	0.6	Cytoplasmic
8	*OmpA*	Sec	100	7	0.4	2.3	0.4	Cytoplasmic
9	*appA*	Sec	100	8.6	1	0.00	0.3	Cytoplasmic
10	*ccmH*	Sec	99.9	6.2	2.5	0.6	0.7	Cytoplasmic
11	*cex*E	Sec	99.6	8.1	1.7	0.00	0.3	Cytoplasmic
12	*cysP*	Sec	100	7.8	0.4	1.2	0.6	Cytoplasmic
13	*draA*	Sec	100	6.5	0.1	2.9	0.4	Cytoplasmic
14	*dsbD*	Sec	99.9	7.1	2	0.3	0.7	Cytoplasmic
15	*dsbG*	Sec	98.9	2.3	7.2	0.4	0.03	Outer Membrane
16	*faeG*	Sec	100	7.3	2.1	0.1	0.5	Cytoplasmic
17	*fecB*	Sec	100	8.6	1.3	0.00	0.1	Cytoplasmic
18	*fedA*	Sec	100	8.7	1	0.00	0.3	Cytoplasmic
19	*FimF41a*	Sec	100	8.7	1	0.00	0.3	Cytoplasmic
20	*flgI*	Sec	100	8.3	1.3	0.00	0.4	Cytoplasmic
21	*hofQ*	Sec	100	6.2	0.2	3.2	0.4	Cytoplasmic
22	*lolA*	Sec	100	6.8	0.6	1.6	1.00	Cytoplasmic
23	*lptA*	Sec	100	6.2	0.8	1.8	1.2	Cytoplasmic
24	*mal*E	Sec	99.9	7.1	1.1	0.6	1.8	Cytoplasmic
25	*mepA*	Sec	99.7	8.6	0.9	0.00	0.5	Cytoplasmic
26	*appA*	Sec	99.9	8.4	1	0.03	0.5	Cytoplasmic
27	*nrfA*	Sec	100	7.7	1.2	0.3	1.04	Cytoplasmic
28	*ompP*	Sec	99.9	7.4	0.1	2.1	0.3	Cytoplasmic
29	*ompW*	Sec	100	7.8	1	0.4	0.7	Cytoplasmic
30	*papK*	Sec	99.9	2.6	0.00	7.1	0.3	Secreted (Extracellular)
31	*pbpG*	Sec	99.9	9	0.8	0.00	0.2	Cytoplasmic
32	*phoA*	Sec	99.9	7.9	0.6	0.7	0.8	Cytoplasmic
33	*pho*E	Sec	99.9	4	0	5.5	0.4	Secreted (Extracellular)
34	*prs*K	Sec	100	2.3	0	7.5	0.1	Secreted (Extracellular)
35	*psp*E	Sec	100	6	2.5	0.7	0.8	Cytoplasmic
36	*ptrA*	Sec	100	3.4	0.00	5.6	0.6	Secreted (Extracellular)
37	*sfaS*	Sec	99.9	8.9	1.1	0.00	0.0	Cytoplasmic
38	*tauA*	Sec	100	8.9	1.01	0.0	0.0	Cytoplasmic
39	*thiB*	Sec	99.9	7.2	0.3	1.8	0.6	Cytoplasmic
40	*torT*	Sec	99.4	3.7	4.6	1.1	0.5	Inner Membrane
41	*ugpB*	Sec	99.4	8.5	0.8	0.0	0.7	Cytoplasmic
42	*xylF*	Sec	100	7	0.5	1.5	1.0	Cytoplasmic
43	*yfeK*	Sec	100	7.4	2.3	0.0	0.3	Cytoplasmic
44	*yhcN*	Sec	100	6.8	2.5	0.2	0.5	Cytoplasmic
45	*yncJ*	Sec	100	5.9	0.0	3.9	0.2	Cytoplasmic
46	*ynfB*	Sec	100	8.4	1.5	0.0	0.04	Cytoplasmic
47	*zraP*	Sec	100	6.9	1.1	0.5	1.5	Cytoplasmic
48	*ASPG_ERWCH*	Sec	99.3	0.9	9	0.0	0.0	Outer Membrane
49	*Pel2*	Sec	100	8.1	1.0	0.1	0.8	Cytoplasmic

Table 5 Sorting the signal peptides according to aliphatic index, GRAVY, h-region length and D-score respectively

No.	Signal peptides	Net positive charge	Aliphatic Index	D-score	Gravy	h–Region length	Final prediction site
1	*lptA*	3	130.37	0.929	0.881	12–21(10)	Cytoplasmic
2	*Pel2*	1	138.18	0.924	1.191	4–17(14)	Cytoplasmic
3	*flgI*	1	185.5	0.923	1.935	5–14(10)	Cytoplasmic
4	*ptrA*	2	131.74	0.922	0.857	9–17(9)	Secreted (extracellular)
5	*mepA*	1	164.74	0.918	1.479	5–13(9)	Cytoplasmic
6	*appA*	2	137.37	0.917	1.35	8–16(9)	Cytoplasmic
7	*ompW*	2	125.71	0.917	1.21	6–15(10)	Cytoplasmic
8	*ynfB*	3	163.93	0.915	1.239	11–22(12)	Cytoplasmic
9	*FimF41a*	2	124.55	0.913	1.355	5–16(12)	Cytoplasmic
10	*zraP*	3	79.23	0.912	0.746	8–18(11)	Cytoplasmic
11	*MglB*	2	102.17	0.909	0.952	5–17(13)	Cytoplasmic
12	*yncJ*	1	128.64	0.906	1.541	5–15(11)	Cytoplasmic
13	*lolA*	2	139.52	0.905	1.324	5–15(11)	Cytoplasmic
14	*mal*E	3	113.08	0.902	1.012	9–18(10)	Cytoplasmic
15	*prs*K	1	121.43	0.899	1.267	6–14(9)	Secreted (extracellular)
16	*OmpA*	2	121.43	0.897	1.295	5–14(10)	Cytoplasmic
17	*DsbA*	2	144.21	0.895	1.416	4–15(12)	Cytoplasmic
18	*fedA*	2	102.38	0.895	1.29	5–14(10)	Cytoplasmic
19	*LamB*	2	125.2	0.894	1.332	8–19(12)	Cytoplasmic
20	*draA*	2	98.1	0.894	1.162	5–15(11)	Cytoplasmic
21	*xylF*	2	161.3	0.890	1.083	7–16(10)	Cytoplasmic
22	*faeG*	2	112.38	0.888	1.005	5–15(11)	Cytoplasmic
23	*AnsB*	1	93.64	0.876	1.136	8–17(9)	Cytoplasmic
24	*thiB*	2	157.22	0.870	1.589	5–12(8)	Cytoplasmic
25	*tauA*	1	120.45	0.868	1.055	8–16(9)	Cytoplasmic
26	*pho*E	2	130	0.867	1.195	6–15(10)	Secreted (extracellular)
27	*papK*	1	140	0.866	1.39	6–14(9)	Secreted (extracellular)
28	*ugpB*	1	110.87	0.866	0.622	8–17(10)	Cytoplasmic
29	*pbpG*	2	117.2	0.863	1.228	7–18(12)	Cytoplasmic
30	*cysP*	2	164	0.859	1.064	11–19(9)	Cytoplasmic
31	*sfaS*	2	146.82	0.845	1.314	5–16(12)	Cytoplasmic
32	*appA*	1	155.45	0.839	1.405	5–16(12)	Cytoplasmic
33	*FkpA*	2	121.20	0.831	1.212	7–16(10)	Cytoplasmic
34	E*ltB*	2	111.43	0.807	0.89	6–14(9)	Cytoplasmic
35	*yhcN*	2	128.64	0.762	1.418	7–16(10)	Cytoplasmic
36	*phoA*	2	139.52	0.754	0.971	6–14(9)	Cytoplasmic
37	*ompP*	0	114.78	0.747	0.904	5–15(11)	Cytoplasmic
38	*psp*E	2	148.95	0.736	1.711	5–13(9)	Cytoplasmic
39	*dsbD*	1	148.95	0.722	1.632	5–13(9)	Cytoplasmic
40	*ccmH*	1	157.22	0.699	1.828	4–12(9)	Cytoplasmic
41	*cex*E	2	154.21	0.689	1.411	5–13(9)	Cytoplasmic
42	*dsbG*	2	207.06	0.640	2.018	5–14(10)	Outer Membrane
43	*yfeK*	2	179.47	0.624	1.742	5–13(9)	Cytoplasmic
44	*hofQ*	1	162.78	0.597	1.322	5–13(9)	Cytoplasmic
45	*torT*	1	173.33	0.587	2.061	5–13(9)	Inner Membrane
46	*fecB*	1	162.86	0.567	2.076	7–15(9)	Cytoplasmic
47	*Bla*	1	110.43	0.557	1.539	8–19(12)	Cytoplasmic
48	*nrfA*	4	108.85	0.557	0.792	11–19(9)	Cytoplasmic
49	*ASPG_ERWCH*	1	106.67	0.540	1.352	7–17(11)	Outer Membrane

similar to AXA box, therefore we have avoided mentioning this parameter in Table 5. In general the bacteria which uses Sec and SRP pathways translocate unfolded proteins to periplasmic compartment where folding and accumulation are both occurring, on the contrary by the use of TAT pathway they tend to fold secretory proteins in cytoplasm compartment and then translocate the folded proteins to periplasmic compartment for accumulation (De Marco 2009), it seems Sec and SRP pathways are more essential than TAT pathway because folding and purification of secretory proteins in periplasmic or extracellular are easier than in cytoplasm. Since degradation of secretory proteins is less than cytoplasm, it can be concluded that the SPs using these pathways can be more appropriate than SPs which use TAT pathways (Pugsley and Schwartz 1985; Talmadge and Gilbert 1982). As it is shown in Table 4, all SPs in this study belonged to Sec pathway and none could be deleted using this analysis, subsequently other analysis was performed here (it has been reported in previous sections). Finally, it was clarified that among 48 SPs (without *zraP*), 41 of them can translocate rNS3-gp96 protein to cytoplasmic compartment which could confirm the previous analysis (sec pathway), four SPs could translocate NS3-gp96 to extracellular compartments while three of them translocate rNS3-gp96 protein to membrane compartments. Therefore, it seems only these four signal sequences can be introduced as reliable SP. Therefore, according to D-score (the most important feature), Protein prsK protein, Outer membrane pore protein E (phoE), and Fimbrial adapter papK, were introduced (respectively) as the best signal peptides to express rNS3-gp96 protein into extracellular *E. coli. pap*K which is the most famous signal peptide in this analysis.

Conclusion

Due to existing bioinformatics methods for rapid prediction of functional excretory signal peptides, it is essential to use this approach for effective extracellular production of recombinant proteins in heterologous host. In fact, by selecting an appropriate signal peptide for target protein can be reduce the costs and time of the expression and purification of recombinant proteins. This study evaluated 52 different signal peptides and then selected optimum for secretory production of the recombinant NS3-gp96 protein in *E. coli* host. This is the first report in theoretical sequence-based analysis of several signal peptides connected with NS3-gp96 and their efficiency in protein secretion to extracellular medium. So, predicting the best SPs by in silico approach would assist biologist and protein engineers to hasten and facilitate the vital projects. Eventually, prsK protein, outer membrane pore protein E (phoE), and fimbrial adapter papK were introduced (respectively) as the best signal peptides to express rNS3-gp96 protein in to extracellular *E. coli*. Nevertheless, the confirmation of these results needs experimental evaluation.

Acknowledgements This study was financially supported by the office of vice-chancellor for research of Shiraz University of Medical Sciences with the Grant No. 95-01-74-11344. The results described in this research were part of PhD student thesis of Shiva Mohammadi.

Compliance with Ethical Standards

Conflict of interest The authors declare no conflict of interest.

References

Alter HJ, Purcell RH, Shih JW, Melpolder JC, Houghton M, Choo Q-L, Kuo G (1989) Detection of antibody to hepatitis C virus in prospectively followed transfusion recipients with acute and chronic non-A, non-B hepatitis. New Eng J Med 321:1494–1500

Atapour A, Mokarram P, Mostafavi-Pour Z, Ramezani PA (2017) Molecular cloning, expression, and purification of a recombinant fusion protein (rNT-gp96-NT300). Biopharm Int 30:38–44

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

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

Baneyx F, Mujacic M (2004) Recombinant protein folding and misfolding in *Escherichia coli*. Nat Biotechnol 22:1399

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

Chang CCH, Song J, Tey BT, Ramanan RN (2013) Bioinformatics approaches for improved recombinant protein production in *Escherichia coli*: protein solubility prediction. Brief Bioinform 15:953–962

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

De Marco A (2009) Strategies for successful recombinant expression of disulfide bond-dependent proteins in *Escherichia coli*. Microb Cell Fact 8:26

Desvaux M, Parham NJ, Scott-Tucker A, Henderson IR (2004) The general secretory pathway: a general misnomer? Trend Microbiol 12:306–309

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

Ezziane Z (2006) Applications of artificial intelligence in bioinformatics: a review. Expert Syst Appl 30:2–10

Gardy JL, Brinkman FS (2006) Methods for predicting bacterial protein subcellular localization. Nat Rev Microbiol 4:741

Ghasemi Y, Dabbagh F, Rasoul-Amini S, Haghighi AB, Morowvat MH (2012) The possible role of HSPs on Behçet's disease: a bioinformatic approach. Comput Biol Med 42:1079–1085

Gholami A, Shahin S, Mohkam M, Nezafat N, Ghasemi Y (2015) Cloning, characterization and bioinformatics analysis of novel cytosine deaminase from *Escherichia coli* AGH09. Int J Pept Res Ther 21:365–374

Gottesman S (1996) Proteases and their targets in *Escherichia coli*. Annu Rev Genet 30:465–506

Idicula-Thomas S, Balaji PV (2005) Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in *Escherichia coli*. Protein Sci 14:582–592

Keller R, de Keyzer J, Driessen AJ, Palmer T (2012) Co-operation between different targeting pathways during integration of a membrane protein. J Cell Biol 199:303–315

Kumari S, Chaurasia AK (2015) In silico analysis and experimental validation of lipoprotein and novel Tat signal peptides processing in *Anabaena* sp. PCC7120. J Microbiol 53:837–846

Low KO, Mahadi NM, Illias RM (2013) Optimisation of signal peptide for recombinant protein secretion in bacterial hosts. Appl Microbiol Biotechnol 97:3811–3826

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

Mousavi P, Mostafavi-Pour Z, Morowvat MH, Nezafat N, Zamani M, Berenjian A, Ghasemi Y (2017) In silico analysis of several signal peptides for the excretory production of reteplase in *Escherichia coli*. Curr Proteom 14:326–335

Müller M, Bernd Klösgen R (2005) The Tat pathway in bacteria and chloroplasts. Mol Memb Biol 22:113–121

Naika HR, Lingaraju K, Chandramohan V, Krishna V (2015) Evaluation of phytoconstituents and molecular docking against NS3 protease of hepatitis C virus. J Pharm Sci Pharmacol 2:96–103

Natale P, Brüser T, Driessen AJ (2008) Sec-and Tat-mediated protein secretion across the bacterial cytoplasmic membrane—distinct translocases and mechanisms. Biochim Biophys Acta Biomemb 1778:1735–1756

Payne SH et al (2012) Unexpected diversity of signal peptides in prokaryotes. MBio 3:e00339–e00312

Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Method 8:785

Pouriayevali M-H, Bamdad T, Aghasadeghi M-R, Sadat SM, Sabahi F (2016) Construction and immunogenicity analysis of hepatitis C virus (HCV) truncated non-structural protein 3 (NS3) plasmid vaccine. Jundishapur J Microbiol 9:e33909

Pugsley AP, Schwartz M (1985) Export and secretion of proteins by bacteria. FEMS Microbiol Lett 32:3–38

Rosano GL, Ceccarelli EA (2014) Recombinant protein expression in *Escherichia coli*: advances and challenges. Front Microbiol 5:172

Simmonds P (2013) The origin of hepatitis C virus. In: Bartenschlager R (ed) Hepatitis C Virus: from molecular virology to antiviral therapy. Springer, Berlin, pp 1–15

Singh SM, Panda AK (2005) Solubilization and refolding of bacterial inclusion body proteins. J Biosci Bioeng 99:303–310

Singh H, Raghava G (2001) ProPred: prediction of HLA-DR binding sites. Bioinformatics 17:1236–1237

Smialowski P, Doose G, Torkler P, Kaufmann S, Frishman D (2012) PROSO II–a new method for protein solubility prediction. FEBS J 279:2192–2200

Talmadge K, Gilbert W (1982) Cellular location affects protein stability in *Escherichia coli*. Proc Nat Acad Sci USA 79:1830–1833

Tong L et al (2000) Extracellular expression, purification, and characterization of a winter flounder antifreeze polypeptide from *Escherichia coli*. Protein Expr Purif 18:175–181

Walker JM (2005) The proteomics protocols handbook. Springer, New York

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

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* F1F0 ATP synthase c-subunit. FEBS J 280:3425–3435

Zimmermann R, Eyrisch S, Ahmad M, Helms V (2011) Protein translocation across the ER membrane. Biochim Biophys Acta Biomembr 1808:912–924