African Journal of Biotechnology Vol. 9(11), pp. 1561-1567, 15 March, 2010
Available online at http://www.academicjournals.org/AJB
ISSN 1684–5315 © 2010 Academic Journals

Full Length Research Paper

Prediction and analysis of the secreteomic in *Corynebacterium glutamicum* ATCC 13032

Ning Hao, Yan Li, Ming Yan* and Ping Kai Ouyang

State Key Laboratory of Materials-Oriented Chemical Engineering, College of Life Science and Pharmaceutical Engineering, Nanjing University of Technology, Nanjing 210009, Jiangsu, P.R. China.

Accepted 1 February, 2010

***Corynebacterium glutamicum* is an outstanding organism used for amino acid production. Its ability to secrete L-glutamate has been known for almost fifty years now. The complete nucleotide sequence of *C. glutamicum* ATCC 13032 genome was previously determined and allowed the reliable prediction of 3056 protein-coding genes within this genome using computational methods. The 3056 open reading frames (ORFs) of *C. glutamicum* ATCC 13032 were used for the prediction of secreted proteins by bioinformatics approaches, such as SignalP 3.0 and Proteome Analyst. 167 proteins were predicted to be secreted and contain signal peptides, whose amino residues were relatively conserved. Among them, 10 have RR-motif signal peptide and 46 have SignalPaseII signal peptide. Total of 167 secreted proteins have functional descriptions, many of which were enzymes that are involved in metabolism. This prediction method has given good insights into the whole secreted proteome of *C. glutamicum* and provided basis to further studies of its secretomic features at a genome level.**

Key words: *Corynebacterium glutamicum*, secreted proteins, signal peptides.

INTRODUCTION

Corynebacterium glutamicum is a Gram-positive, non-sporulating bacterium that was isolated by Kinoshita and co-workers in a screen for bacteria that secrete L-glutamate. It is used for the industrial production of amino acids such as glutamate and lysine that have been used in human food, animal feed and pharmaceutical products for several decades. In addition, recent studies have indicated the potential of *C. glutamicum* for production of other commercially relevant compounds, such as succinate or ethanol (Bott, 2007; Leuchtenberger, 1996).

Because of the importance of *C. glutamicum* in industrial biotechnology, its 3.3-Mb genome sequence has been determined several times independently. The establishment of a completely annotated *C. glutamicum* genome sequence is a big leap forward to the understanding of the biology of this organism (Kalinowski et al., 2003). The complete genome sequence is the basis for extensive expression analyses by proteome and transcriptome technologies, which will lead to a comprehensive systemic understanding of gene expression and regulatory networks (Becker et al., 2007).

It is well established that *C. glutamicum* can secrete certain proteins to high concentrations in the medium. However, until recently it was difficult to estimate the number of exported proteins belonging to the secretome of *C. glutamicum*. The completion of the *C. glutamicum* genome sequencing project and the availability of programs for the identification of signal peptides and trans-membrane segments in large collections of protein sequences through worldwide web servers have now made it possible to predict the most likely location of all 3,056 annotated proteins (that is, the proteome) of this organism. Computer-assisted studies have indicated that approximately 15% of the proteome of a given organism, such as *C. glutamicum*, contains membrane sorting signals in the form of hydrophobic stretches of amino acids that can integrate in and span the membrane (Saleh et al., 2001). Some of these putative membrane proteins contain amino-terminal signal peptides (SPs) and may in

*Corresponding author. E-mail: yanming@njut.edu.cn. Tel: 86-25- 8358-7355.

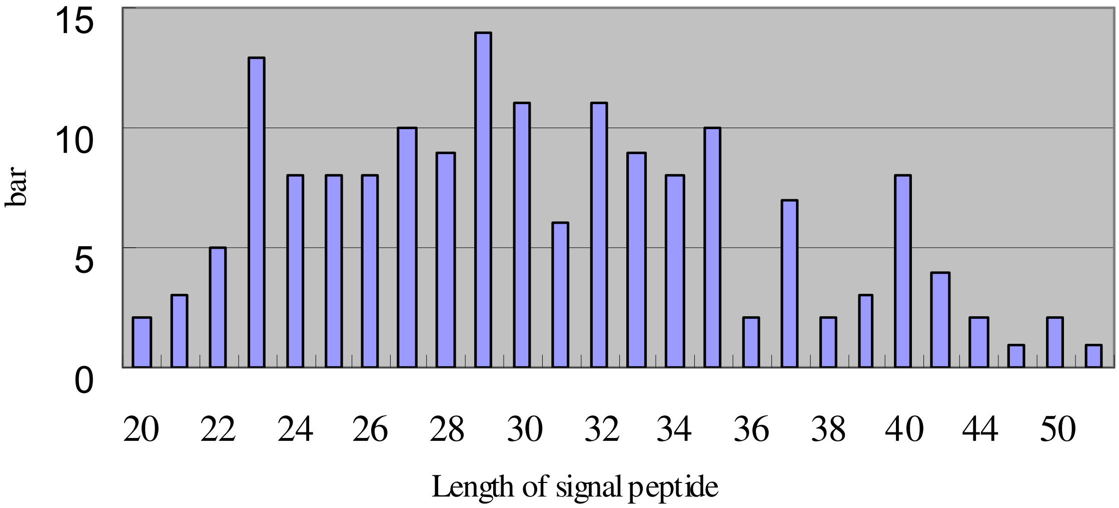

Figure 1. Distribution of signal peptide with different length.

fact be exported proteins.

MATERIALS AND METHODS

Materials

The genome and proteome sequence of *C. glutamicum* ATCC 13032 (accession no. NC_006958) was retrieved from National Center for Biotechnology Information website (http://www.ncbi.nlm.nih.gov). Signal peptides and secreted proteins searches were carried out using SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP) and Proteome Analyst (http: //www.cs.ualberta.ca/~bioinfo/PA), respectively. Lipoprotein SPs and twin-arginine translocation SPs were made with the LipoP 1.0 (http: //www.cbs.dtu.dk/services/LipoP) and TatP 1.0 (http: //www.cbs.dtu.dk/services/TatP).

Prediction of SPs using SignalP 3.0 (Bendtsen et al., 2004; Emanuelsson et al., 2000)

SignalP 3.0 server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks and hidden Markov models. Because of the length restrictions, we divided the sequence into 6 submissions to classify proteins as asignal peptides/non-signal peptides.

Prediction of secreted proteins using proteome analyst 2.5

Proteome analyst is a publicly-available, high-throughput, Web-based system for predicting various properties of each protein in an entire proteome (Szafron et al., 2004). We uploaded a FASTA format file containing the 454 SPs sequences to be classified by SignalP 3.0. With the removal of the transmembrane proteins, a set of 167 secreted proteins were obtained and functionally classified. Motif search of secreted proteins uncharacted was carried out using MYHITS (http: //myhits.isb-sib.ch/ cgi-bin/motif_scan).

Prediction of lipoprotein signal peptides using LipoP 1.0

The hidden Markov model (HMM) was able to distinguish between lipoproteins (SPaseII-cleaved proteins), SPaseI-cleaved proteins, cytoplasmic proteins and transmembrane proteins. The HMM was able to identify 92.9% of the lipoproteins included in a Gram-positive test set. The results obtained were significantly better than those of previously developed methods (Juncker et al., 2003).

Prediction of twin-arginine signal peptides using Tap 1.0

The method is able to discriminate Tat signal peptides from cytoplasmic proteins carrying a similar motif, as well as from Sec signal peptides, with high accuracy (Nielsen et al., 1999).

RESULTS

Length distribution of predicted signal peptides

The 167 predicted signal peptides (Figure 1) had a length varying from 20 to 55 residues, with an average of 31 residues.

The frequency of 20 amino acids residues of signal peptides

The frequency of 20 amino acids residues of predicted signal peptides is shown in Figure 2. It has to be noted that nonpolar amino acid residues can be found more abundant, where alanine (Ala) residue is most abundant (18%). The frequency of positively charged residues arginine (Arg), lysine (Lys) and histidine (His) is respectively 5.0, 3.7 and 0.8%, while that of negatively charged residues asparagine (Asp) and glutamine (Glu) is 1.2 and 1.6% respectively. For uncharged residues, the frequency of serine (Ser) is the most (9.7%) while tyrosine (Tyr) is

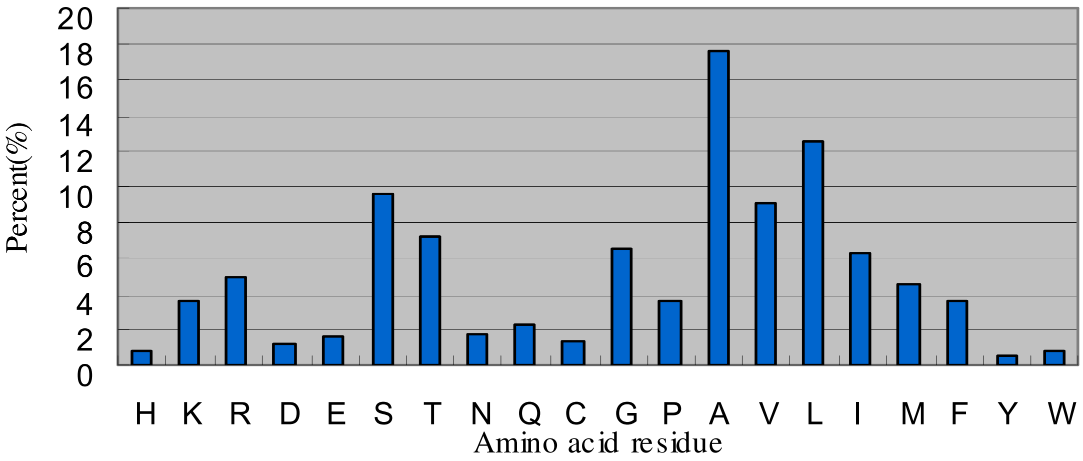

Figure 2. The frequency of 20 amino acids residues of signal peptides in secreted proteins.

Table 1. The frequency of 20 amino acid residues around the signal peptidase cleavage sites (%).

Amino acid	First letter abbreviation	-3	-2	-1	1	2
Histidine	H	0	1.796407	0	1.796407	0
Lysine	K	0	2.39521	0	1.197605	0
Arginine	R	0	1.197605	0.598802	1.197605	1.197605
Aspartic acid	D	1.197605	1.197605	2.994012	8.383234	8.982036
Glutamic acid	E	2.39521	4.790419	0.598802	9.580838	13.17365
Serine	S	7.185629	17.36527	4.191617	10.77844	16.16766
Threonine	T	3.592814	3.592814	0	5.389222	8.982036
Asparagine	N	0.598802	5.389222	1.796407	1.197605	2.994012
Glutamine	Q	1.796407	10.17964	1.796407	16.16766	1.796407
Cysteine	C	0.598802	5.988024	1.197605	2.39521	4.790419
Glycine	G	5.988024	5.389222	5.389222	4.191617	4.191617
Proline	P	1.197605	2.39521	2.39521	2.39521	12.5 7485
Alanine	A	51.49701	7.185629	74.8503	20.95808	4.790419
Valine	V	14.97006	8.982036	0.598802	3.592814	8.383234
Leucine	L	5.389222	7.185629	0.598802	4.790419	2.39521
Isoleucine	I	2.39521	6.586826	1.197605	1.796407	2.994012
Methionine	M	0.598802	1.197605	0	0.598802	1.197605
Phenylalanine	F	0	2.39521	0.598802	2.994012	1.796407
Tyrosine	Y	0	1.197605	0.598802	0.598802	1.796407
Tryptophan	W	0.598802	3.592814	0.598802	0	1.796407

the least (0.5%). The fact that the residues whose frequency is more than 5% are mostly aliphatic amino acids suggests that such residues are involved in the targeting of secreted protein to specific membrane locations in *C. glutamicum*.

Change of amino-acid residues in C-domain

Three distinct regions comprise the N-terminal signal sequence: the charged N-terminus (N-domain), the hydrophobic core (H-domain), and the C-terminal cleavage domain (C-domain) (M Akita, 1990; Paetzel et al., 1998). The C-domain of the predicted signal peptides carries a type I SPase cleavage site, with the consensus sequence A-X-A at position -3 to -1 relative to the SPase I cleavage site (Table 1). The frequency of Ala at -3, -1, +1 position is respectively, 51.5, 74.9 and 21.0%. It is important to note that the C-domain must have an extended (β-sheeted) structure for effective interaction with the active site of type I SPases. Based on the crystal structure of the type I SPase of *Escherichia coli*, the side chains of

Table 2. RR-motif signal peptide.

ID of protein	Signal peptide
gi_62390270	MAQISRRHFLAAATVAGAGATLAA CA
gi_62390439	MLPIWMGLPFKKAGALSRRKAVFSALGAAALIGAALPTIPTAQA QT
gi_62390553	MVSRRGFLGGAGLIAGASALA GC
gi_62391106	MPQLSRRQFLQTTAVTAGLATFAGTPARA EE
gi_62391279	MFKKHRHGLGSPETKPRSITRRFFTAAAATLAGLAVLSGCTAQPSQA ED
gi_62391337	MLNIARNRNMKRRLAIAAFVATATATATMAPASA QT
gi_62391490	MSTTITRRNFLRATGILGVAAGIGATLAACA PD
gi_62391492	MRRKLTTTLENKPGARLGGFRALAPTSKIALVFLLLIFLLAIFAPLIAKY DP
gi_62391811	MTSSFSRRQFLLGGLVLAGTGAVA AC
gi_62391935	MVRSTGSMAIATLLSRITGFLRTVMIGAALSPAIASA FN

residues at the -1 and -3 positions are thought to be bound in two shallow hydrophobic substrate-binding pockets (S1 and S3) of the active site, whereas the side chain of the residue at position -2 is pointing outwards from the enzyme. It is presumably for this reason that residues tolerated at positions -3 and -1 of the signal peptide are generally small and uncharged, while almost all residues seem to be allowed at position -2. Nevertheless, a preference for Ser (17%) at position -2 of the signal peptide seems to exist in *C. glutamicum*.

According to the predictions, an Ala residue is most abundant (21%) at position +1 of the mature protein, but all other residues, with the exception of tryptophan (Trp), seem to be allowed at this position.

Lipoprotein signal peptides

Putative lipoprotein signal peptides were identified through similarity searches in the LipoP 1.0 database. Putative lipoprotein sorting signals identified by the method were combined with those identified by SignalP, resulting in a total number of 46. Signal peptides from lipoproteins differ in several respects from those of secretory signals.

Twin-arginine signal peptides

Proteins containing a signal peptide with the RR-motif (R-R-X-#-#, where # is a hydrophobic residue) may be transported via the Tat pathway. Through TatP 1.0 database search for the presence of this motif in amino-terminal protein sequences, a total number of 10 putative RR-signal peptides were identified (Table 2). Notably, the RR-motif was also found in the signal peptides of two putative lipoproteins, suggesting that these proteins might also be substrates for the Tat pathway.

Functional classification of predicted secreted proteins

A total of 167 secreted proteins have functional descriptions, of which 51 are secreted proteins, while 35 are transport system proteins, 43 are enzymes, 18 are binding proteins and precursors (Figure 3).

In order to identify potential functions of 51 secreted proteins uncharacted, motif analysis was carried out using MYHITS. As a result, five sequences exist as related activity site (Table 3).

Secretion by *C. glutamicum* of heterologous proteins

C. glutamicum has been used for the industrial production of amino acids for several decades. However, there had been only few reports concerning heterologous protein secretion in *C. glutamicum*. Recently, in many studies it has being demonstrated that *Streptomyces mobaraensis* transglutaminase (Date et al., 2004; Kikuchi et al., 2003) another enzyme used in the food industry, and human epidermal growth factor can be efficiently secreted in active form by *C. glutamicum*; so this strain is a potential host for industrial-scale protein production. In addition, the Tat pathway in *C. glutamicum* has been demonstrated to specifically mediate the secretion of *Arthrobacter globiformis* isomaltodextranase and green fluorescent protein (GFP) carrying an *E. coli* TorA signal peptide. Furthermore, the Tat-pathway-dependent secretion of GFP has been shown to be far superior in *C. glutamicum* compared with two other Gram-positive bacteria, *Bacillus subtilis* and *Staphylococcus carnosus* (Meissner et al., 2007). More recently, there was report of secretion of *Streptococcus bovis* α-amylase using cspB promoter and signal sequence by *C. glutamicum* for the efficient utilization of raw starch, identifying *C. glutamicum* as a very useful host for the expression of heterologous proteins (Kikuchi et al., 2006; Tateno and Akihiko, 2007; Kikuchi et al., 2007).

In the present prediction, about 21% of secreted proteins were involved in transport system, while 10 putative RR-signal peptides were identified. Thus, the observations that *C. glutamicum* can efficiently release different heterologous proteins into the culture medium subsequently to their Tat-dependent translocation across

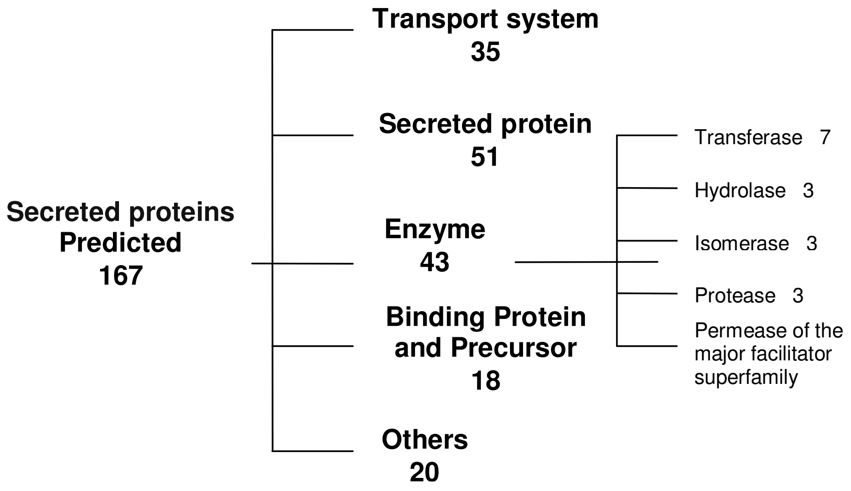

Figure 3. Classification of predicted secreted proteins.

Table 3. Sequences with key motifs in secreted proteins with unknown function.

ID of protein	Description of the motif	Possible function
gi_62389680	Serine proteases, trypsin family, histidine active site.	Serine protease
gi_62390357	Hemopexin domain signature.	Hemopexin
gi_62391502	TonB-dependent receptor proteins signature 1.	TonB-dependent receptor
gi_62391027	Transcription factor TFIIB repeat signature.	a target of gene-specific transcriptional activators
gi_62391174	Aldehyde dehydrogenases glutamic acid active site.	Aldehyde dehydrogenase

the plasma membrane is somewhat unexpected, particularly in view of the fact that the proteins translocated via the Tat pathway usually arrive at the transside of the cytoplasmic membrane in a fully folded state.

Identified enzymes in the predicted secreted proteins

Extensive works have been done to investigate other functions of *C. glutamicum* besides amino acids production binding with the enzyme protein secreted, such as aromatic degradation. In their work, many secreted proteins were identified (Table 4), while some of which were also identified by our prediction.

DISCUSSION

In the present work, efforts were made to identify all genes in the *C. glutamicum* genome database whose deduced proteins would likely be soluble secreted proteins (the secretome). While certain *C. glutamicum* secretory proteins have been studied in detail, such as six mycolyltransferase genes and their gene products, more data on the entire secretome is needed. One approach to rapidly predict the functions of an entire proteome is to utilize genomic database information and prediction algorithms. The use of computer-based prediction algorithms is a powerful, systematic, and rapid tool to obtain preliminary functional information on gene products of an entire genome. Information can then be analyzed in global fashion to organize functional groupings of predicted proteins, or individually, in order to identify genes of particular interest for future experimental study. The *C. glutamicum* genome database was queried in an effort to identify all genes whose deduced proteins would likely be secreted proteins in order to: (a) obtain a global perspective on secreted proteins in *C. glutamicum*; and (b) identify previously uncharacterized genes for further experimental study. A series of prediction algorithms available was therefore used on internet-based servers to analyze the *C. glutamicum* genome database.

In this study, identification was carried out on genes whose proteins have signal peptides and are known to be secreted extracellularly, including: *cop1*, *cmt1*, *cmt2*, *cmt3*, *cmt4*, and *cmt5*. Interestingly, 42 of secreted proteins predicted are of unknown function. In order to gain additional insight into the functional properties of these potential *C. glutamicum* secretory proteins in our dataset, we referred to the extensive motif search and got some related activity site, such as serine protease, Aldehyde dehydrogenase.

Important limitations of this approach are that it relies on prediction algorithms with a defined error rate which

Table 4. Identified enzymes in the predicted secreted proteins.

ID of protein	Enzymes	EC Number	Reference
gi_62389241	Trehalose corynomycolyl transferase	2.3.1.122	(Brand et al., 2003)
gi_62389812	corynomycolyl transferase	2.3.1.122	(Brand et al., 2003)
gi_62389917	corynomycolyl transferase	2.3.1.122	(Brand et al., 2003)
gi_62391021	corynomycolyl transferase	2.3.1.122	(Brand et al., 2003)
gi_62391714	Trehalose corynomycolyl transferase	2.3.1.122	(Brand et al., 2003)
gi_62391716	Trehalose corynomycolyl transferase	2.3.1.122	(Brand et al., 2003)
gi_62391453	putative secreted protein, hypothetical endoglucanase		(Brand et al., 2003)
gi_62390420	secreted cell wall-associated hydrolase		(Brand et al., 2003)
gi_62389427	secreted protein		(Brand et al., 2003)
gi_62390276	putative secreted hydrolase		(Brand et al., 2003)
gi_62389528	ABC-type amino acid transport system, secreted component		(Schluesener et al., 2007)
gi_62390348	NADH Dehydrogenase	1.6.99.3	(Schluesener et al., 2007)
gi_62390845	putative secreted or membrane protein		(Schluesener et al., 2007)
gi_62389441	Probable short-chain dehydrogenase, secreted		(Huang et al., 2008)

could potentially be greater in specific organisms. Furthermore, these prediction algorithms are useful for rapid preliminary analyses of large amounts of genomic data, but it must be emphasized that these are only predictions, which require experimental validation. The present approach was to be inclusive rather than exclusive; so overall, these results probably represent an overestimation of the actual *C. glutamicum* secretome, especially since many open reading frames (ORFs) in the genome database have not been confirmed experimentally and some ORFs may not be expressed.

In conclusion, for further experimental study, we would like to examine novel secreted proteins and identify their function using proteomics-based approaches to analyze *C. glutamicum* secreted proteins.

ACKNOWLEDGEMENT

This work was supported by grants from the National Basic Research Program of China (2009CB724700)

REFERENCES

Becker J, Klopprogge C, Herold A, Zelder O, Bolten CJ, Wittmann C (2007). Metabolic flux engineering of L-lysine production in *Corynebacterium glutamicum*-over expression and modification of G6P dehydrogenase. J. Biotechnol. 132: 99-109.

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

Bott M (2007). Offering surprises: TCA cycle regulation in *Corynebacterium glutamicum*. Trends Microbiol. 15: 417-425.

Brand S, Niehaus K, Puhler A, Kalinowski J (2003). Identification and functional analysis of six mycolyltransferase genes of *Corynebacterium glutamicum* ATCC 13032: the genes cop1, cmt1, and cmt2 can replace each other in the synthesis of trehalose dicorynomycolate, a component of the mycolic acid layer of the cell envelope. Arch. Microbiol. 180: 33-44.

Date M, Umezawa Y, Matsui H, Kikuchi Y (2004). High level expression of Streptomyces mobaraensis transglutaminase in *Corynebacterium glutamicum* using a chimeric pro-region from Streptomyces cinnamoneus transglutaminase. J. Biotechnol. 110: 219-226.

Emanuelsson O, Nielsen H, Brunak S, Von Heijne G (2000). Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. J. Mol. Biol. 300: 1005-1016.

Huang Y, Zhao KX, Shen XH, Jiang CY, Liu SJ (2008). Genetic and biochemical characterization of a 4-hydroxybenzoate hydroxylase from *Corynebacterium glutamicum*. Appl. Microbiol. Biotechnol. 78: 75-83.

Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci. 12: 1652-1662.

Kalinowski J, Bathe B, Bartels D, Bischoff N, Bott M, Burkovski A, Dusch N, Eggeling L, Eikmanns BJ, Gaigalat L, Goesmann A, Hartmann M, Huthmacher K, Kramer R, Linke B, McHardy AC, Meyer F, Mockel B, Pfefferle W, Puhler A, Rey DA, Ruckert C, Rupp O, Sahm H, Wendisch VF, Wiegrabe I, Tauch A (2003). The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. J. Biotechnol. 104: 5-25.

Kikuchi Y, Date M, Yokoyama K, Umezawa Y, Matsui H (2003). Secretion of active-form Streptoverticillium mobaraense transglutaminase by *Corynebacterium glutamicum*: processing of the pro-transglutaminase by a cosecreted subtilisin-Like protease from Streptomyces albogriseolus. Appl. Environ. Microbiol. 69: 358-366.

Kikuchi Y, Date M, Itaya H, Matsui K, Wu LF (2006). Functional analysis of the twin-arginine translocation pathway in *Corynebacterium glutamicum* ATCC 13869. Appl. Environ. Microbiol. 72: 7183-7192.

Kikuchi YCI, Hiroshi I, Masayo D, Kazuhiko M, Long-Fei W (2007). Production of Chryseobacterium proteolyticum protein-glutaminase using the twin-arginine translocation pathway in *Corynebacterium glutamicum*. Appl. Microbiol. Biotechnol. 78: 67-74.

Leuchtenberger W (1996). Amino acids-technical production and use. Biotechnology, 6: 465-502.

M Akita SS, Matsuyama S, Mizushima S (1990). SecA interacts with secretory proteins by recognizing the positive charge at the amino terminus of the signal peptide in *Escherichia coli*. J. Biol. Chem. 265: 8164-8169.

Meissner D, Vollstedt A, van Dijl JM, Freudl R (2007). Comparative analysis of twin-arginine (Tat)-dependent protein secretion of a heterologous model protein (GFP) in three different Gram-positive bacteria. Appl. Microbiol. Biotechnol. 76: 633-642.

Nielsen H, Brunak S, Von Heijne G (1999). Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein Eng. 12: 3-9.

Paetzel M, Dalbey RE, Strynadka NC (1998). Crystal structure of a bacterial signal peptidase in complex with a beta-lactam inhibitor.

Nature, 396: 186-190.

Saleh MT, Fillon M, Brennan PJ, Belisle JT (2001). Identification of putative exported/secreted proteins in prokaryotic proteomes. Gene, 269: 195-204.

Schluesener D, Rogner M, Poetsch A (2007). Evaluation of two proteomics technologies used to screen the membrane proteomes of wild-type *Corynebacterium glutamicum* and an L-lysine-producing strain. Analyt. Bioanalyt. Chem. 389: 1055-1064.

Szafron D, Lu P, Greiner R, Wishart DS, Poulin B, Eisner R, Lu Z, Anvik J, Macdonell C, Fyshe A, Meeuwis D (2004). Proteome Analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations. Nucleic Acids Res. 32: W365-371.

Tateno THF, Akihiko K (2007). Direct production of L-lysine from raw corn starch by *Corynebacterium glutamicum* secreting *Streptococcus bovis* α-amylase using cspB promoter and signal sequence. Appl. Microbiol. Biotechnol. 77: 533-541.