Vol.2012-BIO-28 No.16
2012/3/29

Principal component analysis for bacterial proteomic analysis II

Y-h Taguchi[†1] and Akira Okamoto[†2]

Proteomic analysis is very useful procedure to understand the bacterial behavior changes with reaction to the external environment. This is because most of genomic information of bacteria is devoted to code enzyme to control metabolic networks inside the individual cell. In this paper, we have performed proteomic analysis of *Streptococcus pyogenes*, which is known to be a flesh-eating bacteria and can cause several human life-threatening diseases. Its proteome during growth phase is measured for four time points under two different incubation conditions; with and without shaking. The purpose of it is to understand adaptivity to oxidative stress. Principal component analysis is applied and turns out to be useful to depict biologically important proteins for both supernatant and cell components.

1. Introduction

Streptococcus pyogenes is an important pathogen. There are more than 700 million infections estimated each year and over 650,000 cases of severe, invasive infections that have a mortality rate of 25 %. Although *S. pyogenes* is a normal bacteria flora, *S. pyogenes* can also occasionally cause life-threatening diseases. This means, it will be important to know what the trigger is for it to cause such diseases. There are huge number of researches[1)] to investigate transcrptome responses to external environment, but there are very few researches on how its proteome changes dependent upon external stimulates.

In this paper, we have systematically compared proteome of *S. pyogenes* during growing phases under two distinct incubation conditions; with and without shaking. The later condition was designed to be more oxidative stress condition. The purpose of this research is to know the proteomic response to these two different growth conditions. Using the principal component analysis (PCA)[2)], we have selected representative proteins. Many of the representative proteins play biological roles during the incubation.

†1 Department of Physics, Chuo University
†2 Graduate School of Medicine,Nagoya University

2. Material and Methods

2.1 Proteome Analysis

In this study, *Streptococcus pyogenes* (serotype M1) SF370 of a clinical isolate is investigated. The sample is incubated at 37 °C for 4, 6, 14 and 20 hours ($OD_{660} = 0.40, 0.83, 0.92$, and 0.90, respectively).

Incubated bacterial cells are separated into the supernatant fraction and the bacterial fraction by centrifugation. The reason why the cell fraction is not divided into soluble/insoluble fraction in contrast to the previous researches[3)–4)] is because these two do not differ from each other so much in the preliminary investigations (not shown here). Proteins contained in each fraction is partially purified by ethanol-chloroform purification. After reduced alkylation, they are fragmented by Lysyl Endopeptidase and Trypsin and are provided as sample for mass spectrometry Detection of fragmented proteins are performed by LTQ-Orbitrap XL (Thermo Fisher Sceintific Inc.) attached with Paradigm MS4 LC system (Michrom BioResources Inc.). Obtained spectrum by LTQ is identified by MASCOT program based upon in-house amino acid database which consists with coding-sequence predicted by genomic analysis[5)] and re-evaluation of genome[6)]. To be identified, at least two unique amino acid sequences for each protein is required. False discovery rate is estimated by decoy databases constructed by randomized amino acid sequences. Each of two fractions is measured three times for each of four time points under two distinct incubation conditions separately. Analyzed quantity by PCA is %emPAI[7)–8)], which expresses amount of proteins and %emPAI is its normalized value. %emPAI is normalized to have zero mean and unit variance before any analyses.

Hereafter, each sample is denoted by the tag ID in the form of XXXYY_Z, where XXX is either "sha" (the incubation under the shaking condition) or "sta" (the incubation under the static condition", YY denotes the duration time of the incubation (05, 07, 14, and 20 hours for the shaking incubation condition, and 04, 06, 14 and 20 hours for the static incubation condition), and Z is "wc" (the whole cell fraction) or "snt" (the supernatant fraction), respectively.

2.2 Transcriptome data

Transcriptome data set[9] with the accession number GSE5179 is downloaded from Gene Expression Omnibus (GEO). Raw data files GSM1167X.csv (X ranges from 67 to 79) are loaded into analysis program and column data named as F532.Median is used for further analyses. Each sample is normalized so as to have zero mean and unit variance. Then, 6 samples in the stationary phase are compared with 6 samples in the growth phase.

2.3 Statistical Methods

2.3.1 Application of pricipal comonent anaysis to proteome data

Suppose that we have proteome data x_{sp}, which is the normalized %emPAI of pth protein at sth sample ($s = 1, \ldots, S, p = 1, \ldots, P$). This data can be understood as two ways, i.e.,

Category 1 In total, there are regarded to be S kinds of samples, each of which is characterized by the set of amounts of P kinds of proteins; a set of P dimensional vectors, the number of which is S.

Category 2 In total, there are regarded to be P kinds of proteins, each of which is characterized by the amount of its expression at S kinds of samples; a set of S dimensional vectors, the number of which is P.

Principal component analysis (PCA) can be applied to both of two cases. If PCA is applied to the former (Category 1), the S kinds of samples are characterized with D_s principal components scores(PCSs) $y_s^i, (i = 1, \ldots, D_s)$,as

$$\mathbf{x}_s = (y_s^1, y_s^2, \ldots, y_s^{D_s})$$
$$y_s^i = \sum_p a_{ip} x_{sp}$$

instead of P kinds of proteins. Alternatively, if PCA is applied to the later (Category 2), the P kinds of proteins are characterized with D_p PCSs $y_p^i, (i = 1, \ldots, D_p)$,as

$$\mathbf{x}_p = (y_p^1, y_p^2, \ldots, y_p^{D_p})$$
$$y_p^i = \sum_s a_{js} x_{sp}$$

instead of S kinds of samples.

2.3.2 Selection of representative proteins

In some cases, PCA can be used to select representative $P'(< P)$ proteins[3)–4)] as follows. At first, each protein is embedded into $D'_p(< D_p)$ dimensional space (typically, D'_p is taken to be 2) by category 2 PCA. Then, the set S_p of top P' proteins which are far from origin are decided, i.e.,

$$S_p \equiv \left\{ p \mid \mathrm{rank}_p \left[\sum_{i=1}^{D'_p} \left(y_p^i\right)^2 \right] \leq P' \right\}$$

where $\mathrm{rank}_p[f_p]$ is the descent rank order of the element f_p.

P' is decided to take minimum number such that $y_s^i, (i = 1, \ldots, D'_s < D_s)$, where typically D'_s is taken to be 2, computed only with the selected P' proteins does not differ very much from the original y_s^i computed with all proteins.

This procedure is repeated after removing P' proteins, i.e., PCA is applied to the remaining $P - P'$ proteins. Then we get additional set $S_{p'}$ of P"$(< P - P')$ proteins to express new PCSs obtained by $P - P'$ proteins.

2.3.3 P-values to describe the difference of transcriptome between the growth phase and the static phase

Using the two sided t-test, we get P-values to check if expression of each phase differ from each other and the obtained P-values are attributed to each gene. After that, 1643 genes have significant P-values ($P < 0.05$) even after the application of FDR correction based upon BH criterion, among 1798 genes to which Spy-IDs are attributed.

3. Results

3.1 Overview of proteome with PCA analysis

Figure 1A shows two dimensional embedding of samples using category 1 PCA. Then $P' = 23$ proteins (Table 1) are selected based upon the two dimensional embedding (not shown here) of proteins obtained by category 2 PCA. Hereafter we call this as round one selection. After that, all of samples are re-embedded into two dimensional space (Fig. 1B) by category 1 PCA. Since Fig. 1B is almost identical with Fig. 1A, configuration seen in Fig. 1A turns out to be dependent upon the selected P' proteins only.

Above these procedures are repeated again for the remaining $P - P'$ proteins and we have successfully selected round two representative proteins $P" = 30$. (Figure 2 and Table 2).

Vol.2012-BIO-28 No.16
2012/3/29

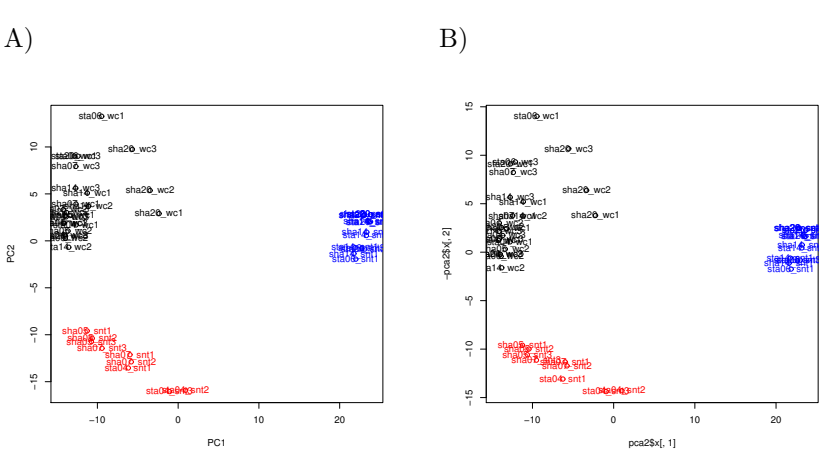

Fig. 1 A) Two dimensional embeddings of samples by Category 1 PCA. Black: the whole cell experiments (wc experiments), Red: the early phase extracellular proteomes (sha05_snt, sha07_snt, and sta04_snt experimets), and Blue: the late phase extracellular proteomes (sha14_snt, sha20_snt, sta06_snt, sta14_snt, and sta20_snt experiments) B) The same as A) but using only the selected $P' = 23$ proteins shown in Table 1.

Table 1 Round one representative proteins. Ribosomal proteins are underlined. The proteins in italic letter are mentioned in the text.

SPy1489:hlpA	*SPy2039:speB*	SPy1073:rplL	SPy2005	*SPy2018:emm1*
SPy0059:rpmC	*SPy0611:tufA*	*SPy0274:plr*	SPy0062:rplX	*SPy2043:mf*
SPy0613:tpi	*SPy2079:AhpC*	SPy1831:rpsF	SPy2160:rpmG	SPy1373:ptsH
SPy0731:eno	*SPy1371:gapN*	*SPy1881:pgk*	*SPy0711:speC*	SPy0071:rpmD
SPy2070:groEL	SPy0019	*SPy0712:mf2*		

Table 2 Round two representative proteins. Notations are the same as in Table 1.

SPy0076:rpmJ	SPy1888:rpmB	SPy0063:rplE	SPy0717:rpmE	*SPy1429:gpmA*
SPy0822:rpmA	*SPy0273:fus*	SPy2092:rpsB	SPy0051:rplW	*SPy1282:pyk*
SPy0055:rplV	SPy1835:trx	*SPy1889:fba*	SPy1294	SPy1544:arcB
SPy0857:mur1.2	SPy0460:rplK	SPy0069:rpsE	SPy0272:rpsG	SPy1932:rplM
SPy1261	SPy1547:sagP	SPy1801:isp2	SPy1262	*SPy1436:mf3*
SPy1234:rpsT	SPy0052:rplB	*SPy2072:groES*	SPy0913	SPy1613

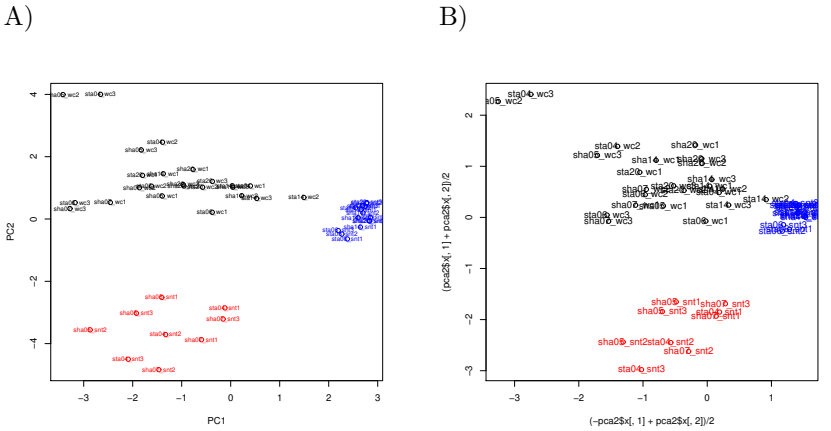

Fig. 2 A) Two dimensional embeddings of samples by Category 1 PCA, after the exclusion of P' proteins in Table 1. B) The same as A) but using only the selected $P'' = 30$ proteins shown in Table 2.

The proteomes of *S. pyogenes* SF370, that grown under shaking or static culture condition, were clustered into three groups (Figures 1 and 2): the whole cellular proteome (all whole cell experiments in Figures 1 and 2), the early phase extracellular proteome (sha05_snt, sha07_snt, and sta04_snt experiments in Figures 1 and 2), and the late phase extracellular proteome (sha14_snt, sha20_snt, sta06_snt, sta14_snt, and sta20_snt experiments in Figure 1 and 2), respectively. These results indicate that the proteomic phenotype of *S. pyogenes* were divided into the two growth stages, the early growth phase that consists of the states at 5 and 7 hours under the shaking condition and the state at 4 hours under the static condition, and the late growth phase that consist of the states at the 14 and 20 hours under the shaking condition and the states at the 6, 14, and 20 hours under the static condition. It is suggested that the proteomic phenotype grown under the static condition might rapidly grown from the early growth stage to the late growth stage compared with the shaking culture condition. Since the cell density (OD_{660}) at 5 hour under the shaking condition and the cell density at 4 hour under the static condition are the same value ($OD_{660} = 0.4$) and the cell density at 7 hour under the shaking condition and the cell density at 6 hour under the

static condition are the same value ($OD_{660} = 0.8$), the proteome is dependent upon the cellular fraction (whole cell or extracellular) or the time development rather than the culture condition.

3.2 Biological meanings of representative proteins

In Tables 1 and 2, we have shown representative proteins for round one and two. Figure 3 shows expressions of the below mentioned proteins among those.

In this study, there are four designed experimental groups characterized by the combination of two criterion: two cellular fractions (the whole cell component or the supernatant components) and two culture conditions (incubation with or without shaking). Several proteins are group specific and are picked up by PCA. For example, peroxiredoxin reductase (SPy2079:AhpC), which is estimated to be involved in oxygen metabolism and hydrogen peroxide decomposition, is found in shaking culture condition rather than static condition. It seems reasonable that the increasing amount of AhpC in shaking condition because the shaking condition induces the higher oxygen stress. On the other hand, twenty out of the fifty-three representative proteins picked up with PCA are riobosomal subunit proteins (the proteins underlined in Tables 1 and 2). This number is as many as a half of ribosomal proteins identified in this study, while total number of ribosomal proteins annotated in SF370 genome is fifty-three. These twenty ribosomal proteins were picked up with PCA due to the abundance in the cellular fraction (not shown here). The reason why several ribosomal proteins were also found in extracellular fraction is possibly because of the leakage during cell division (see below).

Besides, many virulence associated proteins, pyogenic exotoxin B (SpeB; SPy2039), pyogenic exotoxin C (SpeC; SPy0711), mitogenic factors (Mf; SPy2043, Mf2; SPy0712, and Mf3; SPy1436), and M protein (Emm; SPy2018), are picked up by PCA analysis. These virulence-associated proteins have their own combination of the spatial and temporal distributions. SpeB increases monotonically in time, in both shaking and static culture condition. On the other hand, both Mf2 and SpeC increase under the shaking condition, but decrease under the static condition. The amount of both M protein and Mf increase and that of Mf3 decrease in shaking condition, although their amount keep constant value under the static incubation condition. The common distribution patterns are shared by the several abundant enzymes concerning the protein biosynthesis: such as an elongation factor EF-2 (Fus, SPy0273), an elongation factor Tu (TufA, SPy0611), a chaperonin (GroEL, SPy2070), and a co-chaperonin (GroES, SPy2072). The other common fashion of the protein distribution is also observed in enzymes involved in glycolysis: glyceraldehyde-3-phosphate dehydrogenase (Plr, SPy0274), phosphopyruvate hydratase (Eno, SPy0731), pyruvate kinase (Pyk, SPy1282), NADP-dependent glyceraldehyde-3-phosphate dehydrogenase (GapN, SPy1371), phosphoglyceromutase (GpmA, SPy1429), phosphoglycerate kinase (Pgk, SPy1881), and fructose-bisphosphate aldolase (Fba, SPy1889). Each protain is also observed by not small amount in the extracellular fraction at the early growth stage (sha05_snt, sha07_snt and sta04_snt, which are demonstrated by the red color in Fig. 3). They keep constant values throughout all sampling points in the whole cell fraction. None of these proteins possessed signal sequence for secretion. Moreover, they are estimated to be intracellular enzymes such as the proteins involved in protein synthesis or glycolysis. It is confirmed the signal sequence-less proteins are always observed in the extracellular fraction of several bacterial species[10)–11)]. Most bacterial species that belong to firmicutes use autolytic enzymes, such as peptidoglycan hydrolase (Mur1.2, SPy0857), during the cell division processes[12)–14)]. Mur1.2 is also observed in early growth stage. It is supposed that these proteins are leaked from cytoplasm during cell division, especially in early growth stage.

In conclusion, we have successfully selected biologically important proteins.

3.3 Comparison with transcriptome analysis

Although there are no transcriptomic analyses performed to investigate the difference between the shaking and static incubation conditions, there is a research where the transcriptome is compared between the stationary phase and the exponential phase[9)]. We also analysed this public domain data sets (see Materials and Methods) and tried to investigate if the gene coding the proteins picked up with PCA in this study show the significant difference between transcriptome between the static and exponential phases. In order to compare transcriptome between stationary and exponential phase, P-values, the rejection probability for the difference between the static and exponential phases, are attributed to transcruptome which correspond to representative proteins. These P-values are

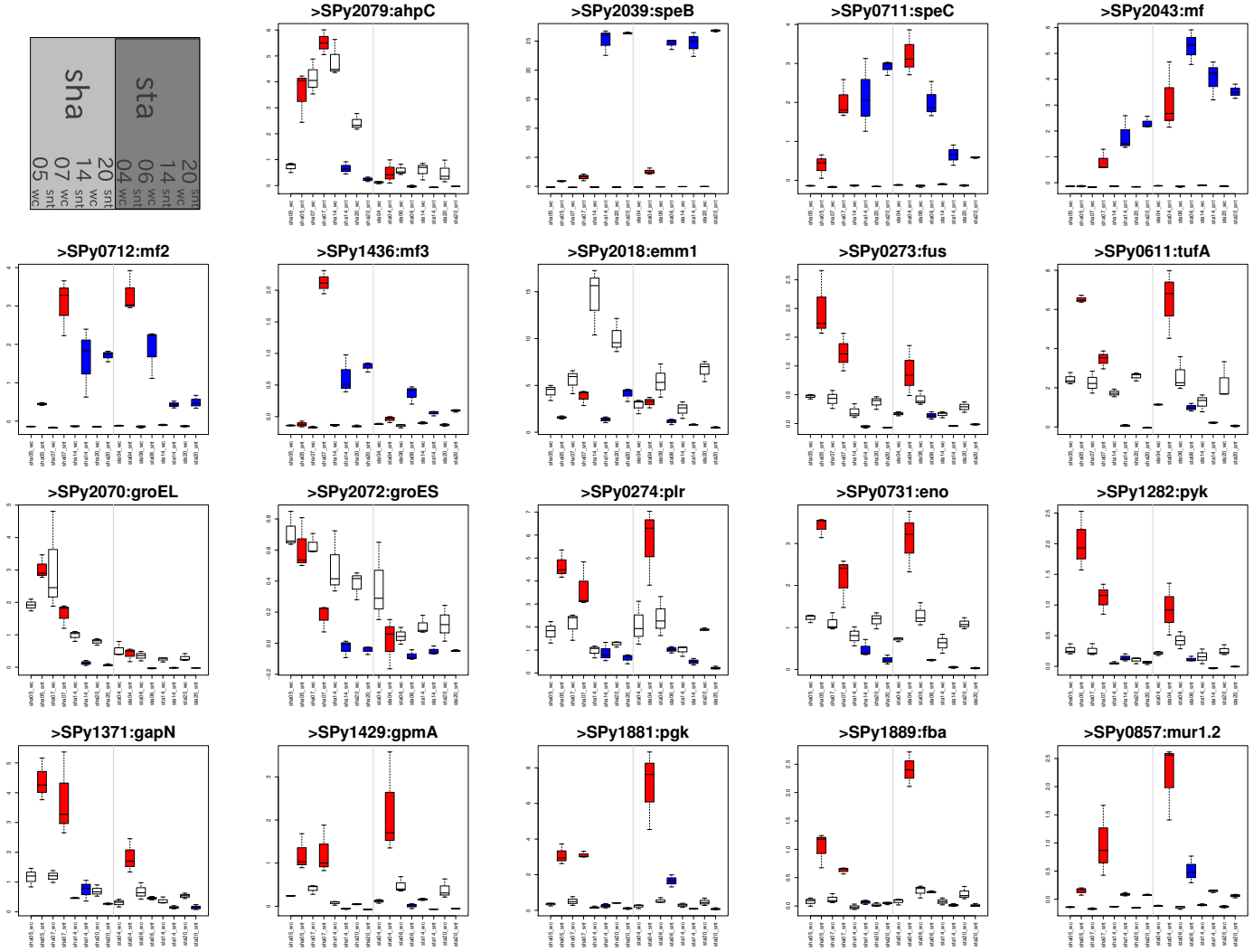

Fig. 3 Expression of representative proteins menthioned in the text. Colors (black, red, and blue) correspond to the colors in Figs. 1 and 2. The top-left panel: Schematic explanation of each panel.

compared with P-values for other proteins than representatives. Then P-values to depict the significant difference between two sets of P-values is obtained. Both of P-values attributed to each of round one and two are less than 1×10^{-3} (Wilcoxon test). This means, proteins whose expression differs between two incubation conditions are also significantly different with each other in transcriptome levels between exponential-phase and stationary-phase. Since the difference between two incubation conditions is supposed to be the difference of time scale as mentioned above, our selection of representative proteins based upon proteome data turns out to be coincident with transcriptome analysis.

4. Conclusion

In this paper, we have performed proteome analysis of *Streptococcus pyogenes*, under two distinct incubation conditions; stationary and shuffle. Representative proteins are selected by iterative applications of PCA in two ways. These proteins turn out to be biologically informative and their trasctiptome expression also differs significantly between stationary and grow phases.

Acknowledgement

This work was supported by KAKENHI (23300357)

References

1) Beyer-Sehlmeyer, G., Kreikemeyer, B., Hörster, A. and Podbielski, A.: Analysis of the growth phase-associated transcriptome of Streptococcus pyogenes, *International Journal of Medical Microbiology*, Vol.295, No.3, pp.161 – 177 (online), DOI:10.1016/j.ijmm.2005.02.010 (2005).
2) Rao, P.K. and Li, Q.: Principal Component Analysis of Proteome Dynamics in Iron-starved Mycobacterium Tuberculosis, *J Proteomics Bioinform*, Vol.2, pp.19–31 (2009).
3) Okamoto, A. and Taguchi, Y.-h.: Principal Component Analysis for Bacterial Proteomic Analysis, *IPSJ SIG Technical Report*, Vol.2011-BIO-26, pp.1–6 (2011).
4) Taguchi, Y.-h. and Okamoto, A.: Principal Component Analysis for Bacterial Proteomic Analysis, *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops*, pp.961–963 (online), DOI:10.1109/BIBMW.2011.6112520 (2011).
5) Ferretti, J.J., McShan, W.M., Ajdic, D., Savic, D.J., Savic, G., Lyon, K., Primeaux, C., Sezate, S., Suvorov, A.N., Kenton, S., Lai, H.S., Lin, S.P., Qian, Y., Jia, H.G., Najar, F.Z., Ren, Q., Zhu, H., Song, L., White, J., Yuan, X., Clifton, S.W., Roe, B.A. and McLaughlin, R.: Complete genome sequence of an M1 strain of Streptococcus pyogenes, *Proc. Natl. Acad. Sci. U.S.A.*, Vol.98, pp.4658–4663 (2001).
6) Okamoto, A. and Yamada, K.: Proteome driven re-evaluation and functional annotation of the Streptococcus pyogenes SF370 genome, *BMC Microbiol.*, Vol.11, p. 249 (2011).
7) Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J. and Mann, M.: Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein, *Mol. Cell Proteomics*, Vol.4, pp.1265–1272 (2005).
8) Shinoda, K., Tomita, M. and Ishihama, Y.: emPAI Calc–for the estimation of protein abundance from large-scale identification data by liquid chromatography-tandem mass spectrometry, *Bioinformatics*, Vol.26, pp.576–577 (2010).
9) Barnett, T.C., Bugrysheva, J.V. and Scott, J.R.: Role of mRNA stability in growth phase regulation of gene expression in the group A streptococcus, *J. Bacteriol.*, Vol.189, pp.1866–1873 (2007).
10) Lei, B., Mackie, S., Lukomski, S. and Musser, J.M.: Identification and immunogenicity of group A Streptococcus culture supernatant proteins, *Infect. Immun.*, Vol.68, pp.6807–6818 (2000).
11) Len, A.C., Cordwell, S.J., Harty, D.W. and Jacques, N.A.: Cellular and extracellular proteome analysis of Streptococcus mutans grown in a chemostat, *Proteomics*, Vol.3, pp.627–646 (2003).
12) Oshida, T., Sugai, M., Komatsuzawa, H., Hong, Y.M., Suginaka, H. and Tomasz, A.: A Staphylococcus aureus autolysin that has an N-acetylmuramoyl-L-alanine amidase domain and an endo-beta-N-acetylglucosaminidase domain: cloning, sequence analysis, and characterization, *Proc. Natl. Acad. Sci. U.S.A.*, Vol.92, pp. 285–289 (1995).
13) Blackman, S.A., Smith, T.J. and Foster, S.J.: The role of autolysins during vegetative growth of Bacillus subtilis 168, *Microbiology (Reading, Engl.)*, Vol.144 (Pt 1), pp.73–82 (1998).
14) Mercier, C., Durrieu, C., Briandet, R., Domakova, E., Tremblay, J., Buist, G. and Kulakauskas, S.: Positive role of peptidoglycan breaks in lactococcal biofilm formation, *Mol. Microbiol.*, Vol.46, pp.235–243 (2002).