http://dx.doi.org/10.5012/bkcs.2012.33.12.4041

Tracing Metabolite Footsteps of *Escherichia coli* Along the Time Course of Recombinant Protein Expression by Two-Dimensional NMR Spectroscopy

Young Kee Chae,[*] Seol Hyun Kim, James J. Ellinger,[†] and John L. Markley[†]

Department of Chemistry and Institute for Chemical Biology, Sejong University, Seoul 143-747, Korea
[*]E-mail: ykchae@sejong.ac.kr
[†]Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706, USA
Received August 14, 2012, Accepted September 13, 2012

The recombinant expression of proteins has been the method of choice to meet the demands from proteomics and structural genomics studies. Despite its successful production of many heterologous proteins, *Escherichia coli* failed to produce many other proteins in their native forms. This may be related to the fact that the stresses resulting from the overproduction interfere with cellular processes. To better understand the physiological change during the overproduction phase, we profiled the metabolites along the time course of the recombinant protein expression. We identified 32 metabolites collected from different time points in the protein production phase. The stress induced by protein production can be characterized by (A) the increased usage of aspartic acid, choline, glycerol, and *N*-acetyllysine; and (B) the accumulation of adenosine, alanine, oxidized glutathione, glycine, *N*-acetylputrescine, and uracil. We envision that this work can be used to create a strategy for the production of usable proteins in large quantities.

Key Words : Metabolite profiling, Recombinant protein, Overexpression, NMR

Introduction

Optimal expression of recombinant proteins is one of the major hurdles for many scientists in the modern era of biological science. Many recombinant proteins are expressed in a prokaryotic system such *Escherichia coli* due to its simplicity.[1] However, not all recombinant proteins are expressed equally well under all conditions. For example, many eukaryotic proteins could not be produced at all, or are produced only in a misfolded form in *E. coli* because of the difference in the cellular environment and/or the lack of proper post-translational modification machinery.[2] To overcome these obstacles, people have developed various methods such as strains supplemented with rare codons,[3] strains with disabled thioredoxin reductase and glutathioine reductase,[4] plasmids with a tunable promoter,[5] fusion tags,[6-11] low temperature expression,[12] transport to periplasmic space,[13] and coexpression with other proteins.[14-17] Despite all these efforts based on trial-and-error, however, there has not been any systematic or targeted approach to resolve the problems.

When the protein production is induced in the *E. coli* cell, cellular physiology is disrupted in such a way that the metabolic activities and gene expression patterns are reoriented to meet the burden, which is similar to the response to the environmental stress.[18] This stress response may be specific for each recombinant protein to be produced, which would explain the reason why it is hard to derive a generalized method for the successful recombinant protein production. If we could harness the metabolic information during the protein production, and sort out several key components, by supplementing or restricting such components, it would be possible to guide the cells along the pathway to the successful production. As a first step toward developing a protocol for systematic optimization of the production of recombinant proteins, we used nuclear magnetic resonance spectroscopy (NMR) to profile the metabolic footsteps of the *E. coli* cells before and after the induction of recombinant fusion protein.

Materials and Methods

***E. coli* Growth.** The stationary phase induction method was employed for protein expression.[19,20] The trial protein was ubiquitin which had been fused with the cytoplasmic domain of syndecan-2 (2 L).[21] BL21(DE3)RILP (Novagen, Madison, WI, USA) harboring the expression plasmid, pET28a/ubisacII/2L, was grown overnight at 37 °C in 1 L 70% rich medium (7% bactotrypton, 3.5% yeast extract, 7% NaCl) supplemented with 50 mg/mL kanamycin and 34 mg/mL chloramphenicol. The next morning, 3 g of bactotrypton, 1.5 g of yeast extract, and 3 g of NaCl was added to the fully grown culture. As soon as the solid materials dissolved, IPTG was added to the final concentration of 0.5 mM, and the culture was allowed to grow further. Aliquots of 100 mL culture were taken out and harvested at the time points of 0, 0.1, 0.5, 1.0, 1.5, 2.0, 3.0, and 18.0 h after the IPTG induction. 1 mL of each aliquot was reserved for SDS-PAGE analysis. The harvested cells were washed 3 times with 100 mL PBS buffer (10 mM sodium phosphate buffer at pH 7.4 with 150 mM NaCl) and frozen at −80 °C.

Metabolite Extraction and NMR Sample Preparation. Metabolites were extracted following a modified aqueous boiling method.[22,23] The frozen cells were first lyophilized. 15 mL of boiling distilled and deionized water was added to the dried cell pellet in a 50 mL conical tube. The mixture

Table 1. Summary of extraction result from 9 different time points

Time after induction (h)	Mass of dried extract (mg)
0.0	7
0.1	8.3
0.5	9
1.0	9.7
1.5	9.5
2.0	9.8
2.5	10.6
3.0	10.1
18.0	5.9

was incubated at 121 ºC for 20 minutes in an autoclave, then cooled down on ice, and then centrifuged at 4000 RPM at 4 ºC for 30 minutes. The supernatant was transferred to VIVASPIN 20 centrifugal filters (Sartorius Stedim, Bohemia, NY, USA), and centrifuged at 4000 RPM at 4 ºC until less than 0.5 mL of solution remained in the upper compartment. The centrifugal filters were washed with 10 mL water 3 times before use to remove glycerol. The final filtrate was frozen and lyophilized. The dried extract powder was weighed (Table 1) and dissolved in 200 μL of HEPES buffer (5 mM HEPES, 0.2 mM DSS, 0.5 mM NaN_3 in D_2O). The pH was adjusted to 7.35 ± 0.05 using deuterated acid or base.

NMR Data Collection and Data Processing. NMR data were collected at the National Magnetic Resonance Facility at Madison on a Bruker Avance III instrument operating at 600.133 MHz equipped with a triple-resonance (^{1}H, ^{13}C, ^{15}N, ^{2}H lock) 1.7 mm cryogenic probe and a SampleJet. All samples were centrifuged at maximum speed for 5 min prior to placing them into NMR tubes. After inserting each sample into the spectrometer, we waited 120 s to allow the sample to reach thermal equilibrium. The probe was tuned, matched, and locked to deuterium for the first sample; the lock signal was adjusted automatically as needed. All samples were collected using automated shimming and manual shimming as needed to obtain a 50% linewidth on DSS less than 1 Hz. The optimal 90° pulsewidth was determined manually for each sample and the receiver gain was automatically calculated. One-dimensional ^{1}H were collected with excitation sculpting using the zgesgp pulse sequence from the Bruker pulse sequence library; spectra were collected at 298 K with 2 steady state scans, 16 transients, 25250 complex points, a spectral width of 10.5 ppm and an interscan delay of 1.75 s. Two-dimensional ^{1}H-^{13}C HSQC (Heteronuclear Single Quantum Coherence) spectra at natural ^{13}C abundance were collected using the hsqcetgpsisp2.2 pulse sequence from the Bruker pulse sequence library. All spectra were collected at 298 K with 4 transients, 32 steady state scans, GARP (Globally-optimized Alternating-phase Rectangular Pulses) broadband decoupling and a 1.75 s interscan delay time; 1262 complex data points were collected in the direct dimension (^{1}H) with a spectral width of 10.5 ppm (6313 Hz); 256 complex data points were collected in the indirect dimension (^{13}C) with a spectral width of 99.5 ppm (15015 Hz). The resulting spectra were visualized and analyzed by rNMR.[24] The peaklist was sent to MMCD (http://mmcd.nmrfam.wisc.edu) to find candidate metabolites. With the help of the in-house software (S. H. Kim and Y. K. Chae, unpublished data), the peak intensity data were first normalized to one of the peaks of the internal standard HEPES, then adjusted to the sample mass, and processed to yield a proper output format for the multivariate analysis. Multivariate analysis was performed with the R statistics software package (http://www.r-project.org).

Results and Discussion

Sample Preparation. The stationary phase induction was chosen because it simplified the procedure without compromising the amount of the target protein compared to the conventional protocol where the cells were grown continuously until they reached mid-log phase and then the protein production was induced.[21] This method did not require a special medium as in the case of autoinduction protocol for the high throughput routine.[25] Since we used a normal growth temperature (37 ºC) and IPTG concentration (0.5 mM), we focused mostly on the first 3 h after the induction, which was when the *E. coli* culture was most vigorously producing recombinant proteins, and this was proven to be true from the SDS-PAGE analysis (see next section). 15 mL of boiling water was added as the first step in an attempt to denature endogenous enzymes that might otherwise change metabolites during the extraction procedure. The resuspended cells were heated in the autoclave machine, and only the small, soluble and heat-resistant metabolites survived for the next steps. Even with the hot extraction condition, we observed that only a fraction of the starting cell mass was lysed as evidenced by the fact that we could collect the cell pellet after autoclave by centrifugation only at 4000 g. If most of the cells had been lysed, we would have observed a cloudy supernatant. As shown in Table 1, most samples yielded more than 7 mg, but this amount was not enough to make the NMR sample volume with the buffer-to-extract ratio Lewis *et al*. suggested.[23] As an alternative, we used a fixed buffer volume of 200 μL which was manageable to adjust pH and more than enough to fill 1.7 mm NMR tubes. Since we equalized the volume, the difference in concentration of each sample had to be taken into account in the analysis stage. The average mass of the dried extract per g of dried cells was 110 ± 20 mg. The yield of the final dried extract from the initial dried yeast cells was around 11%, which was about 5 times higher than in the case of yeast.

Protein Expression Analysis. The target protein (ubiquitin-fused 2L) was produced well with the chosen expression method. The target protein band began to be discernible 30 min after the induction as shown in Figure 1. Each SDS-PAGE sample was prepared from a 1 mL aliquot at each time point studied, therefore it represents the total protein level at that specific moment. Figure 1 shows that the amounts of total protein and the target protein reached a maximum 2 h after induction (lane 7). Very large proteins (> 97 kDa)

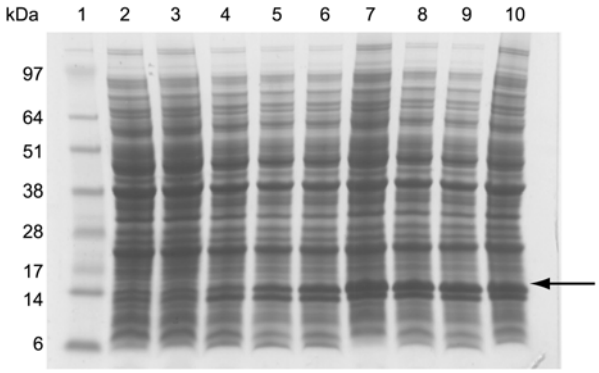

Figure 1. Expression of ubiquitin-2L as monitored by 4-12% SDS-PAGE. Lane 1, size marker; lane 2, whole cell lysate before induction; lane 3-10, whole cell lysate at 0.1, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 18.0 h after IPTG induction. The band corresponding to the target protein is marked with an arrow.

were expressed most when the cells experienced the stationary phase (lanes 2, 3, and 10). It is interesting to find that apparently the same very large proteins were expressed to a similar level 2 h after induction (lane 7) which coincides with the time when the target protein reached its maximum level. This may suggest that the protein production burdens cells in a similar way that the aging process does.

NMR Data Collection and Processing. Each 2D HSQC experiment took about 2 h. Since the sample volume was so small (45 mL), the sample temperature was influenced by the GARP decoupling during acquisition so that we had to wait for a longer time (1.75 s) for the next scan than usual (1 s). However, thanks to the small sample volume required, we were able to make a more concentrated sample. Although 2D NMR took a longer time than 1D NMR which has been employed as a major technique, it can be considered as a potent alternative for higher accuracy and robustness.[26-29] In fact, we could obtain a modest spectrum in 2 h, which was greatly facilitated by the cryoprobe and autosampler.

NMR Data Analysis. Figure 2 shows an upfield region of the NMR spectrum of a sample (2 h after induction) along with the assignments of identified metabolites. With the combination of rNMR and the MMCD, 32 metabolites were identified by observing multiple peaks. The variation in intensity was minimized by the internal standard (HEPES) which acted as an excellent calibration reference; every NMR sample contained the same concentration (5 mM) of HEPES. The error came not only from the NMR experiment, but also from the NMR sample itself: Each NMR sample contained a slightly different amount of NaCl resulting from the cell washing step with PBS, which influenced the sensitivity of the NMR experiment to a different, though small, degree. A representative peak of each identified metabolite is shown in Figure 3. The intensity of each peak was measured and the internal standard (HEPES) for quantification. Although Figure 3 is based on the raw data before the two adjustments, it offers an "at-a-glance" view of the metabolite changes was provided. First of all, the metabolite profiles of the two stationary states (at 0.0 and 18.0 h) were found different depending on whether the protein was produced: the total metabolite level was noticeably reduced after the protein production. This is reasonable because the cellular resources were used to assemble the recombinant protein, which is not limited to the already existing amino acids, but should also be extended to all the compounds needed to carry out the biosynthetic processes. Thus, the profile at 18 h would represent the "exhausted stage". The two most prominent metabolites regarding this observation were aspartic acid and *N*-acetyllysine whose signals appeared the strongest before induction, but disappeared 18 h after the induction. There is a strong possibility of cell death and DNA degradation at the 18 h time point which would have skewed the

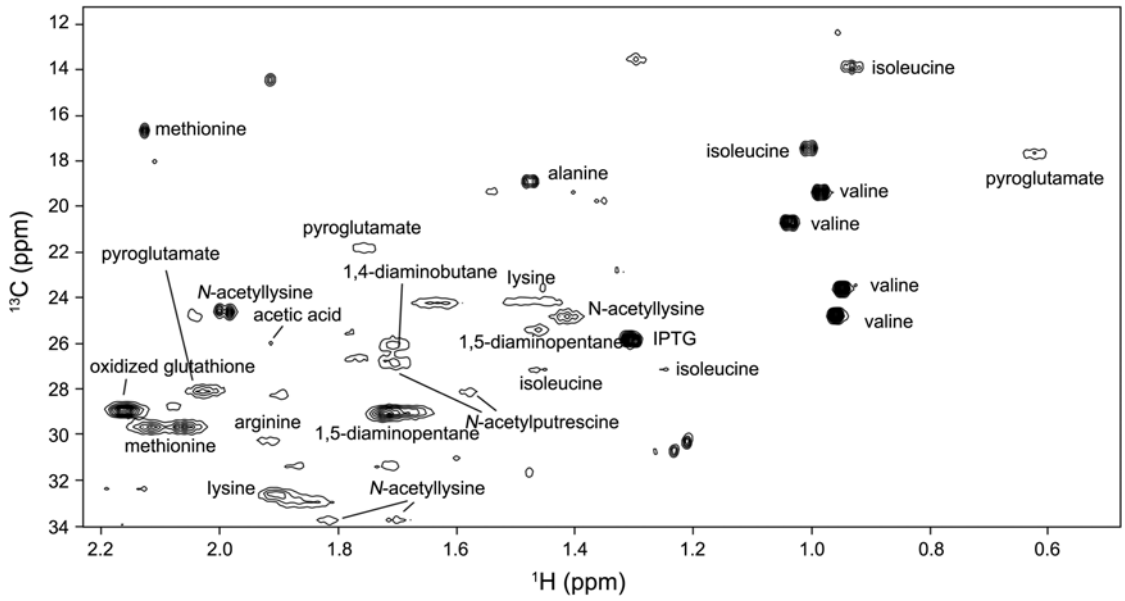

Figure 2. An upfield region of the two-dimensional ^{1}H-^{13}C HSQC spectrum of a sample harvested at 2.0 h after induction. The assigned resonances are labeled.

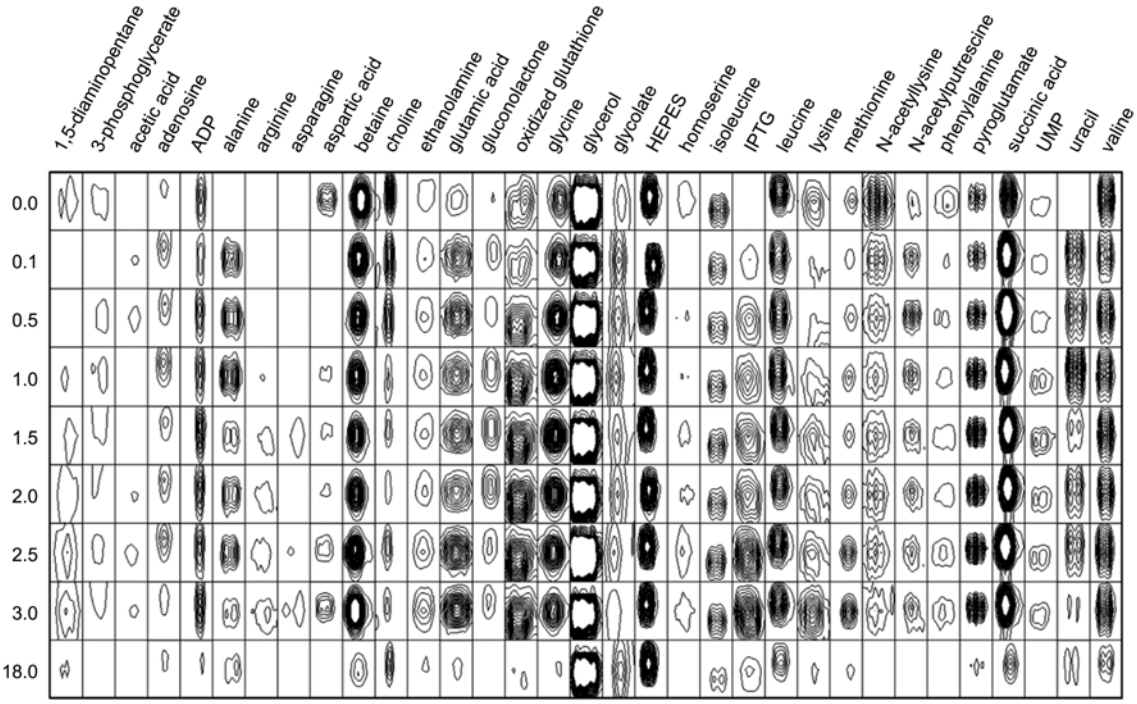

Figure 3. ROI view of the representative resonances of identified metabolites.

metabolite concentrations.

The levels of betaine and glycerol, two widely known osmo-protectants, decreased after the protein expression was induced, but did not change much during the protein production although betaine bounced back up at the later stage (Fig. 4(a)). However, betaine diminished to a much lower level 18 h after the induction while glycerol stayed almost the same (Fig. 3), and this may suggest glycerol as a housekeeping metabolite. It was curious to observe that the glycine level showed a convex shape as opposed to betaine's concave shape, which could be interpreted that glycine molecules were converted to betaine to relieve the overproduction stress. The aspartate and glutamate, which act as precursors of many other amino acids and metabolites, showed an opposite trend: after the induction, the former decreased while the latter increased (Figs. 4(a) and (b)). The increase in pyroglutamic acid would be a direct consequence of this increase of glutamate concentration. The increased level of *N*-acetylputrescine compared to the state just before the induction was interesting because putrescine resulted from the breakdown of the amino acids, which indicated the recycling of existing amino acids for other purposes (Fig. 4(b)). In addition, putrescine has been reported to act against the oxidative stress,[30] which is consistent with our result that the oxidized glutathione increased after the induction. It is worth noting that the intracellular concentration of the exogenous IPTG increased during the active protein production stage, but eventually diminished to a much lower level, indicating it was consumed by the cell or transported back to the medium (Fig. 3). We also noticed that the concentrations of adenosine and uracil were increased after the induction, which suggested that the cells produced these molecules actively to cope with mRNA consumption due to the recombinant protein production.

In summary, our results suggest that the stress induced by the production of the recombinant protein can be characterized by (A) the increased usage of aspartic acid, betaine, choline, glycerol, isoleucine, leucine, *N*-acetyllysine, and phenylalanine; and (B) the accumulation of adenosine, alanine, glutamic acid, oxidized glutathione, glycine, *N*-acetylputrescine, pyroglutamic acid, succinic acid, and uracil. This information could be used to formulate a medium for better production of the target protein by supplying metabolites in demand, for example, such as aspartic acid, choline, and lysine which showed the largest decrease during protein production in our case.

Multivariate Analysis. The multivariate analysis offers the holistic view of the differences between samples. That is, it overlooks the shape of the overall changes rather than the specific details of the differences. Principal component analysis (PCA) was chosen in our study in order to provide an unbiased, multivariate analysis of the changes in metabolites. The intensity data were preprocessed using in-house software to be fed into the R software package and the PCA was applied. The script for performing PCA was kindly written and provided by Ian Lewis (Princeton University, USA) and further modified in-house.

From Figures 5(a) and 5(b), we can see the footsteps of the physiological state of the *E. coli* producing the recombinant protein (dotted lines). The state before the induction clearly differed from that 18 h after. So, the state before the induction could be stationary, and that 18 h after, exhausted. It is interesting to observe that the profile pathway seems to be moving forward and backward between 0.5 h and 2.5 h

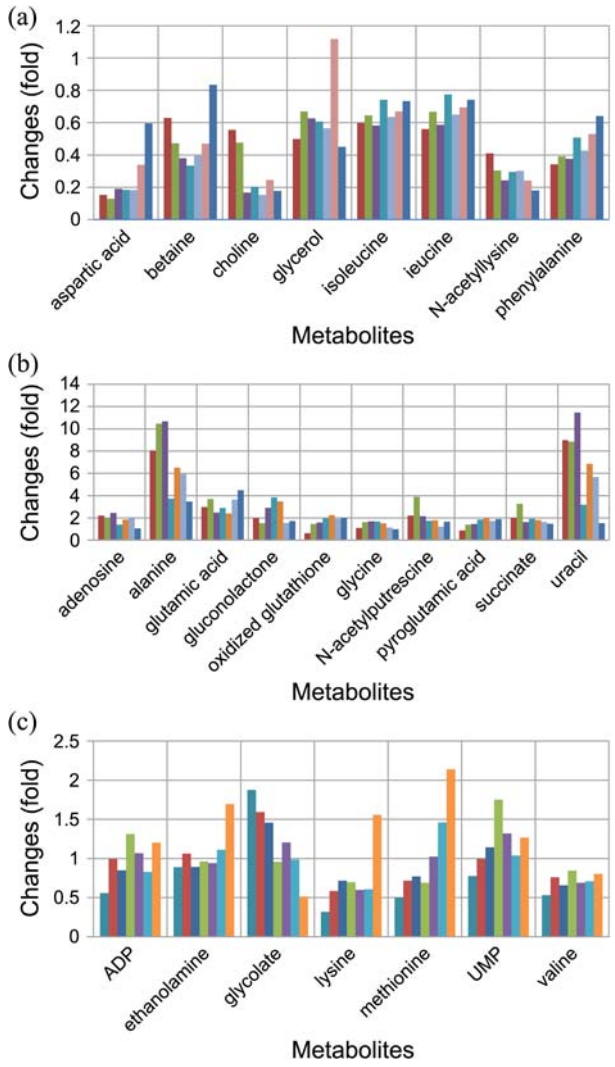

Figure 4. Changes of the intracellular metabolite concentrations at various time points as referenced to that of 0 h. The bar plots show metabolites that, on the average, (a) decreased more than 30%; (b) that increased more than 30%; (c) that increased or decreased less than 30%. The bars of each metabolite represent relative concentrations at 0.1, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, and 18 h after the induction, respectively.

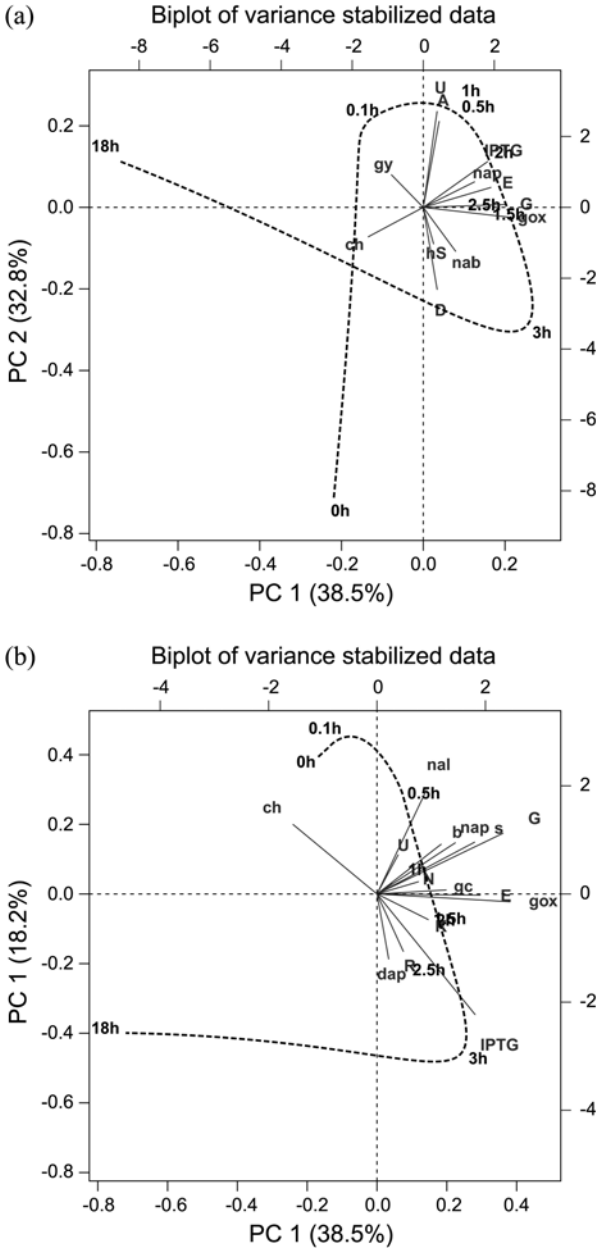

Figure 5. Biplots along the first and second principal component axes (a) and along the first and third axes (b). For better visual representation, only the following metabolites were selected to avoid crowdedness: A, alanine; b, betaine; ch, choline; D, aspartic acid; dap, 1,5-diaminopentane; E, glutamic acid; G, glycine; gc, gluconate; gox, oxidized glutathione; hS, homoserine; K, lysine; N, asparagine; nal, *N*-acetyllysine; nap, *N*-acetylputrescine; R, arginine; s, succinic acid; U, uracil. The dotted line represents the flow of time.

where the order of some time points are in reverse order along the dotted line. This may be the direct result from the stationary phase induction method where the cells in the stationary phase should change their physiology to adjust to both protein production (getting older) and a fresher medium (getting younger). Therefore, we think that the mixed effect was observed. We can also see which metabolites are responsible for positioning each time point in the principal component space. In the biplots (Figs. 5(a) and 5(b)), only the metabolites that contributed more were represented. Other metabolites were omitted for a simpler and clearer view: *N*-acetylputrescine, oxidized glutathione, and glutamic acid were indicative of the state when the protein production started to get more vigorous while IPTG, aspartic acid, and arginine when the protein production reached the end point. We can also observe that a series of metabolites were spread clockwise as time went by. For example, the major indicator of a given time point moves from choline at the beginning, to alanine, then to succinic acid, then to lysine, and finally to IPTG. The exhausted state showed decreased levels of most of the metabolites.

Speculations on Optimizing Recombinant Protein Production by Metabolite Manipulation. So far, many methods have been devised for better production of recombinant proteins in *E. coli* or other hosts. However, all those methods

are essentially based on trial-and-error, that is, one has to try one system just to see whether it works or not, and in many cases, if the first try fails, then it is not easy to find a suitable system that produces the target protein successfully because it is difficult to find a good reason for failure. Our report is intended to evade such a roadblock by providing a systematic search for the successful production of a target protein. In this report, we followed the metabolic footsteps during the protein production and found which metabolites were accumulated or exhausted. If we could supply or remove the appropriate metabolites by manipulating the medium or by applying stress conditions, we could drive the intracellular environment into a properly functioning protein factory. Protein production under various stress conditions is currently under investigation, and we hope this will plow the road to successful protein production.

Conclusions

In this study we used NMR to trace the metabolite profile during the production of the recombinant protein. The metabolite profile followed a trajectory out from the initial state just before the induction, passing through the intermediate time points when the protein was produced vigorously, and finally reaching the exhausted state that was positioned far from all others. Metabolites such as amino acids, nucleic acid components, and stress-related molecules were used to differentiate the physiological states during the protein production. We also observed that the *E. coli* cells utilized even the osmo-protectants like glycerol and betaine at the exhausted state. We hope this work might serve as a stepping stone to the successful production of the recombinant protein.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (2010-0007161). This study made use of the National Magnetic Resonance Facility at Madison, which is supported by NIH grants P41RR02301 and P41RR02301-25S1 (BRTP/NCRR) and P41GM103399 (NIGMS).

References

1. Baneyx, F. *Curr. Opin. Biotech.* **1999**, *10*, 411.
2. Oganesyan, N.; Ankoudinova, I.; Kim, S. H.; Kim, R. *Protein Expr. Purif.* **2007**, *52*, 280.
3. Tegel, H.; Steen, J.; Konrad, A.; Nikdin, H.; Pettersson, K.; Stenvall, M.; Tourle, S.; Wrethagen, U.; Xu, L.; Yderland, L.; Uhlen, M.; Hober, S.; Ottosson, J. *Biotechnol. J.* **2009**, *4*, 51.
4. Prinz, W. A.; Aslund, F.; Holmgren, A.; Beckwith, J. *J. Biol. Chem.* **1997**, *272*, 15661.
5. Guzman, L. M.; Belin, D.; Carson, M. J.; Beckwith, J. *J. Bacteriol.* **1995**, *177*, 4121.
6. Dyson, M. R.; Shadbolt, S. P.; Vincent, K. J.; Perera, R. L.; McCafferty, J. *BMC biotechnology* **2004**, *4*, 32.
7. Niiranen, L.; Espelid, S.; Karlsen, C. R.; Mustonen, M.; Paulsen, S. M.; Heikinheimo, P.; Willassen, N. P. *Protein Expr. Purif.* **2007**, *52*, 210.
8. Brown, B. L.; Hadley, M.; Page, R. *Protein Expr. Purif.* **2008**, *62*, 9.
9. De Marco, V.; Stier, G.; Blandin, S.; de Marco, A. *Biochem. Biophys. Res. Commun.* **2004**, *322*, 766.
10. Zhang, Y. B.; Howitt, J.; McCorkle, S.; Lawrence, P.; Springer, K.; Freimuth, P. *Protein Expr. Purif.* **2004**, *36*, 207.
11. Malakhov, M. P.; Mattern, M. R.; Malakhova, O. A.; Drinker, M.; Weeks, S. D.; Butt, T. R. *J. Struct. Funct. Genomics* **2004**, *5*, 75.
12. Volonte, F.; Marinelli, F.; Gastaldo, L.; Sacchi, S.; Pilone, M. S.; Pollegioni, L.; Molla, G. *Protein Expr. Purif.* **2008**, *61*, 131.
13. Bessette, P. H.; Aslund, F.; Beckwith, J.; Georgiou, G. *Proc. Natl. Acad Sci. USA* **1999**, *96*, 13703.
14. Brown, B. L.; Grigoriu, S.; Kim, Y.; Arruda, J. M.; Davenport, A.; Wood, T. K.; Peti, W.; Page, R. *PLoS Pathogens* **2009**, *5*, e1000706.
15. Wang, Y. H.; Ayrapetov, M. K.; Lin, X.; Sun, G. *Biochem. Biophys. Res. Commun.* **2006**, *346*, 606.
16. Chen, Y.; Song, J.; Sui, S. F.; Wang, D. N. *Protein Expr. Purif.* **2003**, *32*, 221.
17. Sahu, S. K.; Rajasekharan, A.; Gummadi, S. N. *Biotechnol. Lett.* **2009**, *31*, 1745.
18. Hoffmann, F.; Rinas, U. In *Physiological Stress Responses in Bioprocesses*; Springer Berlin/Heidelberg: 2004; Vol. 89, p 73.
19. Chae, Y. K.; Cho, K. S.; Chun, W.; Lee, K. *Protein Pept. Lett.* **2003**, *10*, 369.
20. Chae, Y. K.; Moon, W. J.; Cho, J. Y. *Protein Expr. Purif.* **2009**, *65*, 267.
21. Chae, Y. K.; Lee, W. *Bull. Korean Chem. Soc.* **2008**, *29*, 2449.
22. Chan, E. C.; Koh, P. K.; Mal, M.; Cheah, P. Y.; Eu, K. W.; Backshall, A.; Cavill, R.; Nicholson, J. K.; Keun, H. C. *J. Proteome Res.* **2009**, *8*, 352.
23. Lewis, I. A.; Schommer, S. C.; Hodis, B.; Robb, K. A.; Tonelli, M.; Westler, W. M.; Sussman, M. R.; Markley, J. L. *Anal. Chem.* **2007**, *79*, 9385.
24. Lewis, I. A.; Schommer, S. C.; Markley, J. L. *Magn. Reson. Chem.* **2009**, *47*(Suppl 1), S123.
25. Fox, B. G.; Blommel, P. G. In *Current Protocols in Protein Science*; John Wiley & Sons, Inc.: 2001, p Unit 5.13.
26. Robinette, S. L.; Ajredini, R.; Rasheed, H.; Zeinomar, A.; Schroeder, F. C.; Dossey, A. T.; Edison, A. S. *Anal. Chem.* **2011**, *83*, 1649.
27. Motta, A.; Paris, D.; Melck, D. *Anal. Chem.* **2010**, *82*, 2405.
28. Martineau, E.; Giraudeau, P.; Tea, I.; Akoka, S. *J. Pharm. Biomed. Anal.* **2011**, *54*, 252.
29. Lewis, I. A.; Schommer, S. C.; Hodis, B.; Robb, K. A.; Tonelli, M.; Westler, W. M.; Suissman, M. R.; Markley, J. L. *Anal. Chem.* **2007**, *79*, 9385.
30. Tkachenko, A.; Nesterova, L.; Pshenichnov, M. *Archives of Microbiology* **2001**, *176*, 155.