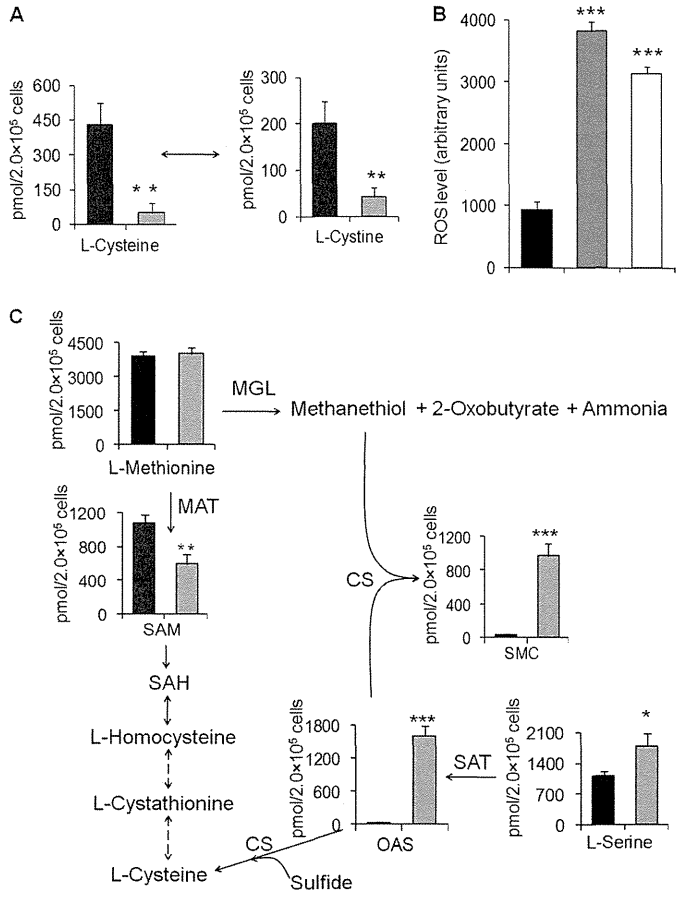


FIGURE 1. **Effects of L-cysteine depletion on the content of L-cysteine/L-cystine, reactive oxygen species, and metabolites in sulfur-containing amino acid metabolism in *E. histolytica*.** Trophozoites were cultured in normal (*black bars*) or cysteine-deprived medium (*gray bars*) for 48 h, or normal medium containing 2 mM paraquat (*white bars*) for 10 h. *Asterisks* (*, **, and ***) denote statistically significant differences with $p \leq 0.05$, $p \leq 0.01$, and $p \leq 0.001$, respectively, as determined by Student's *t* test, and all of the experiments were performed in triplicate. *A*, effects of L-cysteine depletion on intracellular L-cysteine and L-cystine concentrations. The average content (pmol) ± S.D (*error bars*) in 2×10^5 cells is shown. *B*, effects of L-cysteine depletion (72 h) and oxidative stress on the level of reactive oxygen species. The average level of 2′,7′-DCF-DA fluorescence (arbitrary units) ± S.D. (*error bars*) of 5×10^5 cells is shown. *C*, effects of L-cysteine depletion on the level of metabolites involved in sulfur-containing amino acid metabolism. The average content (pmol) ± S.D. (*error bars*) in 2×10^5 cells, performed in triplicate, is shown. *SAH*, *S*-adenosylhomocysteine; *MAT*, L-methionine adenosyltransferase; *MGL*, L-methionine γ-lyase.

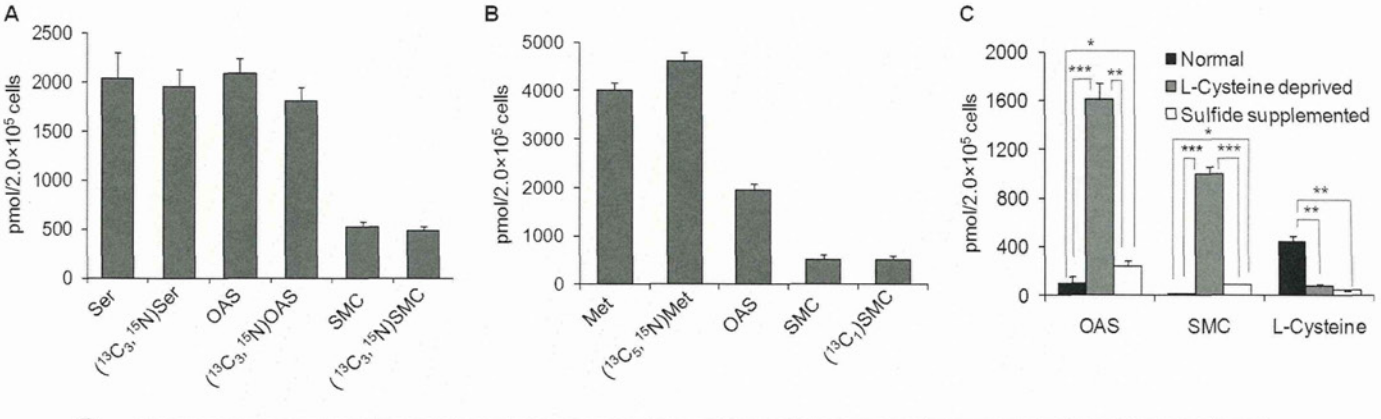
served a marked increase (nearly undetectable under normal conditions) in *S*-methylcysteine (SMC), which is suggested to be a storage compound for sulfide and methyl groups in plants (36). L-Cysteine deprivation also caused a 44 ± 6% decrement in the level of *S*-adenosylmethionine (SAM), whereas the level of L-methionine remained unchanged.

SMC can be formed by the methylation of L-cysteine using either SAM or *S*-methylmethionine as a methyl group donor or by the transfer of the alanyl moiety of OAS to methanethiol (CH_3SH) by CS (37). To differentiate between these possibilities, we performed metabolic labeling of *E. histolytica* trophozoites with stable isotope U-^{13}C,^{15}N-labeled L-serine and L-methionine in normal and L-cysteine-depleted media for 48 h. Upon the addition of [U-$^{13}C_3$,^{15}N]Ser to the L-cysteine-depleted culture medium, comparable levels of [$^{13}C_3$,^{15}N]SMC and unlabeled SMC, derived from [U-$^{13}C_3$,^{15}N]OAS and unlabeled OAS, respectively, were detected (Fig. 2*A*). Similarly, when trophozoites were cultured in the presence of [U-$^{13}C_5$,^{15}N]Met under the L-cysteine-deprived conditions, we also detected comparable levels of [$^{13}C_1$]SMC and unlabeled SMC (Fig. 2*B*). In contrast, under normal conditions, neither SMC nor OAS was detected after [U-$^{13}C_3$,^{15}N]Ser or [U-$^{13}C_5$,^{15}N]Met labeling (data not shown). Taken together, these data clearly indicate that SMC is not synthesized by SAM- or *S*-methylmethionine-dependent methylation of L-cysteine; rather, SMC is synthesized in *E. histolytica* from the backbone of Ser and thiomethyl group of methanethiol.

Surprisingly, [$^{13}C_3$,^{15}N]OAS was not incorporated into either L-cysteine or L-cystine (data not shown), or their levels were too low to be detected by CE-TOFMS. To determine whether the lack of OAS incorporation into L-cysteine was due to the low sulfide concentrations under the *in vitro* axenic growth conditions, we deprived trophozoites of L-cysteine for 45 h and then continued their culture in medium supplemented with 2 mM sulfide for a further 3 h. However, sulfide supplementation did not affect the level of L-cysteine, whereas the levels of SMC and OAS markedly decreased (90.6 ± 3.4 and 84.8 ± 7.7% decrement, respectively) compared with the unsupplemented medium (Fig. 2*C*). These data suggest that sulfide negatively regulates OAS and SMC synthesis and also imply that the pathway formally called the "L-cysteine biosynthetic pathway" is primarily involved in the synthesis of SMC, but not L-cysteine, at least under *in vitro* culture conditions.

In Vitro Examination of S-Methylcysteine Synthesis—To elucidate the enzyme(s) involved in the formation of SMC from OAS and methanethiol, we examined whether different CS isotypes could catalyze the synthesis of SMC. Among the three examined CS isotypes (EhCS1–3), two CS proteins (EhCS1 and EhCS2) are very similar (99% amino acid identity, with two conserved amino acid changes) (5, 6), whereas EhCS3 shares only 83% amino acid identity with the other two isotypes. Both recombinant EhCS1 and EhCS3 efficiently catalyzed the synthesis of SMC using OAS and methanethiol as substrates. As revealed from the kinetic parameters (Fig. 2*D*), EhCS1 and EhCS3 did not show any preference for either methanethiol or sulfide, because the K_m, V_{max}, k_{cat}, and k_{cat}/K_m values for both of these substrates were comparable.

L-Cysteine Depletion Repressed Glycolysis and Energy Generation—Similar to other anaerobic and microaerophilic parasitic protozoa, such as *Giardia lamblia* and *Trichomonas vaginalis*, *E. histolytica* lacks features of aerobic eukaryotic metabolism, including the TCA cycle and oxidative phosphorylation, and primarily generates energy by substrate level phosphorylation (10). The CE-TOFMS-based metabolomic analysis demonstrated that L-cysteine depletion affected the levels of the majority of metabolites involved in glycolysis and its associated pathways (Fig. 3). L-Cysteine-depleted amebae generally contained higher amounts of glycolytic intermediates, with the exception of acetyl CoA and ethanol, than cells cultured under normal conditions. The largest changes caused by L-cysteine depletion were the increment in the levels of glycerol-3-phosphate (2.18 ± 0.25-fold), *O*-phosphoserine (1.70 ± 0.22-fold), pyruvate (1.66 ± 0.26-fold), 3-phosphoglycerate (1.60 ± 0.17-fold), malate (1.50 ± 0.20-fold),



D

	EhCS1				EhCS3			
	Km (mM)	*Vmax* (μmol/min/mg)	*kcat* (s^{-1})	*Kcat/Km* ($mM^{-1}s^{-1}$)	*Km* (mM)	*Vmax* (μmol/min/mg)	*kcat* (s^{-1})	*Kcat/Km* ($mM^{-1}s^{-1}$)
OAS	2.34±0.32	36.20±5.43	22.07±3.26	9.43±2.12	2.54±0.28	31.96±.4.68	19.37±2.98	7.61±2.09
Sulfide	2.01±0.40	35.59±6.21	21.71±3.98	10.81±2.33	2.23±0.34	31.42±5.79	19.08±3.56	8.56±2.24
Methanethiolate	1.66±0.39	25.51±4.78	15.57±2.84	9.37±3.08	2.11±0.39	21.74±3.94	11.22±2.37	5.32±1.88

FIGURE 2. **Examination of *S*-methylcysteine biosynthesis in *E. histolytica*.** *A* and *B*, incorporation of labeled L-serine and L-methionine into *S*-methylcysteine. Trophozoites were cultured in the presence of 6 mM [U-$^{13}C_3$,^{15}N]L-serine (*A*) or 3 mM [U-$^{13}C_5$,^{15}N]L-methionine (*B*) in L-cysteine-deprived medium for 48 h. The average contents (pmol) ± S.D. (*error bars*) of unlabeled and labeled amino acids and their derivatives in 2×10^5 trophozoites, performed in triplicate, are shown. *C*, effects of the supplementation of the medium with sodium sulfide (2 mM) on the levels of OAS, SMC, and L-cysteine, under conditions of L-cysteine deprivation. *Asterisks* (*, **, and ***) denote statistically significant differences with $p \leq 0.05$, $p \leq 0.01$, and $p \leq 0.001$, respectively, as determined by Student's *t* test. *D*, kinetic parameters of recombinant cysteine synthase 1 (EhCS1) and 3 (EhCS3). All of the reactions were performed in triplicate as described under "Experimental Procedures," and the values are expressed as the means ± S.D.

and fumarate (1.60 ± 0.20-fold). Several other metabolites involved in glycolysis, including glucose 6-phosphate, glucose 1-phosphate, fructose 6-phosphate, and phosphoenolpyruvate also showed slightly elevated levels (1.2–1.5-fold), whereas the levels of fructose 1,6-bisphosphate and dihydroxyacetone-phosphate remained unchanged. In contrast to the significant increases in the glycolytic intermediates upstream of pyruvate in amebae cultured under L-cysteine-limited conditions, we observed reduced levels of acetyl CoA (29.4 ± 7.1%) and ethanol (40.7 ± 6.7%), suggesting a decrease in glycolytic flux and ATP generation by L-cysteine depletion. A number of other metabolites downstream of acetyl CoA, such as *N*-acetyl-glutamate, *N*-acetyl β-alanine, *N*-acetyl-leucine, and *N*-acetyl-phenylalanine, were also decreased (supplemental Fig. S1), supporting the premise that the glycolytic flux downstream of pyruvate was repressed.

Because glycolysis is the major source of energy generation in *E. histolytica*, a reduced glycolytic flux was thought to result in a decrement in the energy storage molecules of the trophozoites. As expected, the levels of the nucleotide triphosphates ATP, GTP, UTP, and CTP were significantly lower ($p \leq 0.05$) in the L-cysteine-depleted cells than in the trophozoites maintained under normal conditions (Fig. 3). We also observed slight deceases in the levels of ADP and GDP, whereas the levels of AMP and GMP were unchanged (Fig. 3).

L-Cysteine Depletion Altered Amino Acid Pools—Because amino acids are also used for energy production in *E. histolytica* (38), we examined the effects of L-cysteine deprivation on amino acid levels (supplemental Fig. S1). Next to L-cysteine and L-cystine, L-threonine and L-serine were the most highly modulated by L-cysteine depletion (1.63 ± 0.25- and 2.07 ± 0.29-fold increases, respectively) among the 20 amino acids. In *E. histolytica*, L-threonine and L-serine are catabolized by threonine dehydratase (39) to yield 2-oxobutyrate and pyruvate, respectively, which are in turn used by pyruvate:ferredoxin oxidoreductase for energy generation (40). L-Cysteine depletion also resulted in a slight increase in the intracellular concentration of L-alanine, which is synthesized from pyruvate by L-alanine:2-oxoglutarate aminotransferase (EHI_096750 (EAL50292.1) and EHI_159710 (EAL44861.1)). The levels of the remaining amino acids were not significantly affected by L-cysteine depletion.

L-Cysteine Depletion Caused Increases in Isopropanolamine, Aminoalcohol Phosphates, and Phosphatidylisopropanolamine—The metabolomic analysis of *E. histolytica* also revealed that L-cysteine depletion caused marked changes in amino alcohol metabolism (Fig. 4*A*). L-Cysteine depletion led to a dramatic increase in the levels of isopropanolamine (1-aminopropan-2-ol, Ispn) (5.44 ± 0.76-fold) and isopropanolamine phosphate (Ispn-P, undetected under normal conditions) (Fig. 4*A*). In addition, trophozoites cultured in L-cysteine-limited conditions showed 7.01 ± 1.38- and 2.8 ± 0.21-fold increases in ethanolamine phosphate (Etn-P) and choline phosphate (Cho-P) levels, respectively, whereas the levels of ethanolamine (Etn) and Cho were unchanged. Both Etn-P and Cho-P are intermediates in the Kennedy pathway,

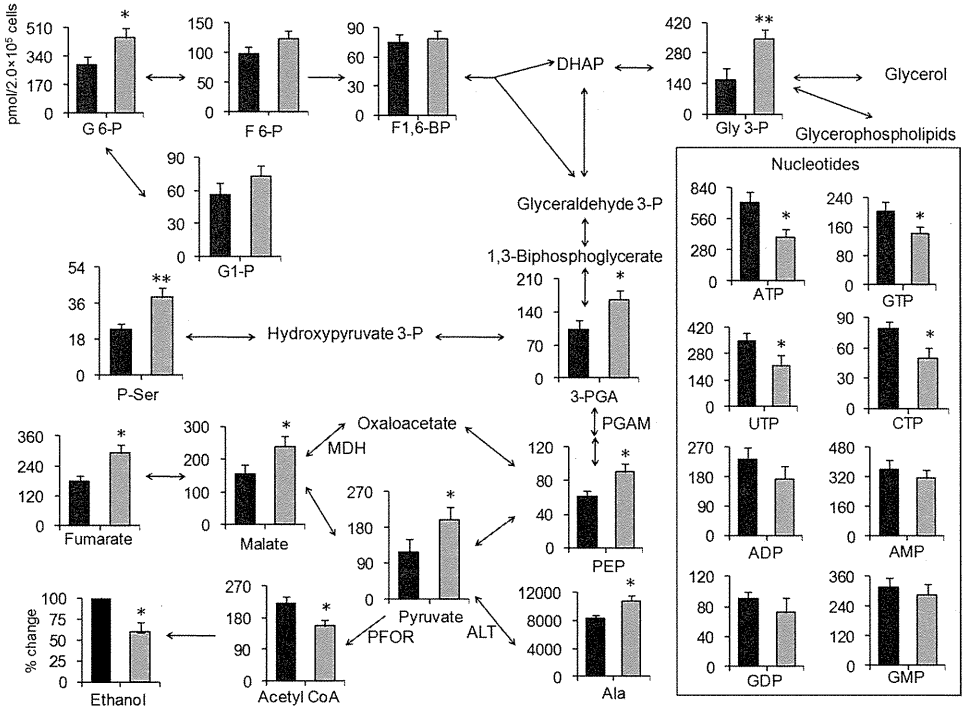


FIGURE 3. **Effects of L-cysteine depletion on the level of metabolites involved in central energy metabolism.** Trophozoites were cultured in normal (*black bars*) or cysteine-deprived medium for 48 h (*gray bars*), and the average contents (pmol) ± S.D. (*error bars*) of the indicated metabolites in 2×10^5 cells are shown (performed in triplicate). Nucleotide metabolites are *boxed*. Ethanol is shown as a percentage change. *Asterisks* (* and **) denote statistically significant differences with $p \leq 0.05$ and $p \leq 0.01$, respectively, as determined by Student's *t* test. *PFOR*, pyruvate:ferredoxin oxidoreductase; *ALT*, L-alanine: 2-oxoglutarate aminotransferase; *PGAM*, phosphoglycerate mutase; *MDH*, malate dehydrogenase; *G 6-P*, glucose 6-phosphate; *G1-P*, glucose 1-phosphate; *F 6-P*, fructose 6-phosphate; *F1,6-BP*, fructose 1,6-biphosphate; *Gly 3-P*, glycerol 3-phosphate; *3-PGA*, 3-phosphoglycerate; *PEP*, phosphoenolpyruvate; *P-Ser*, *O*-phosphoserine.

where phospholipids, including phosphatidylethanolamine and phosphatidylcholine, are produced.

Because L-cysteine limitation affected Ispn-P, Etn-P, and Cho-P concentrations, we next investigated whether L-cysteine deprivation influenced phospholipid synthesis by performing lipid profiling of amebic trophozoites cultured under L-cysteine-deprived or normal conditions using two-dimensional TLC (Fig. 4*B*). We found that in the absence of L-cysteine, *E. histolytica* synthesized an unconventional phospholipid that was verified to be phosphatidylisopropanolamine (PtdIspn) and was undetectable under normal conditions. Quantitation of individual lipids indicated that phosphatidylethanolamine (PtdEtn) decreased by 39.9 ± 6.9%, whereas other phospholipids, such as phosphatidylcholine (PtdCho), phosphatidylserine, phosphatidylinositol, and phosphatidic acid, were unchanged (Fig. 4*C*). These data are consistent with the premise that PtdIspn was formed in a competition for the formation of PtdEtn, the level of which decreased by approximately the identical amount that PtdIspn increased (Fig. 4*C*). To further demonstrate that PtdIspn was formed from Ispn, *E. histolytica* trophozoites were cultured in normal medium containing 5 mM Ispn for 24 h. Under this condition, trophozoites produced an appreciable amount of PtdIspn (Fig. 4*B*, *panel c*).

As described above, L-cysteine depletion increased the level of reactive oxygen species. We therefore examined whether oxidative stress caused the observed changes in amino alcohols and phospholipids. It was observed that the lipid profiling of *E. histolytica* trophozoites cultured with 2 mM paraquat in ambient air for 10 h did not increase PtdIspn (Fig. 4*B*, *panel e*). Furthermore, the addition of D-cysteine to the L-cysteine-lacking medium did not reverse the effects of L-cysteine deprivation on the phospholipid profiles (Fig. 4*B*, *panel d*). These results confirmed that the generation of PtdIspn caused by L-cysteine depletion was not a result of oxidative stress but represents a specific response to L-cysteine deprivation.

Examination of Isopropanolamine Biosynthesis in E. histolytica—Next, we investigated the synthesis route of Ispn in *E. histolytica*. From studies of *Escherichia coli*, it is known that Ispn is synthesized from 1-aminoacetone by the action of Ispn:NAD^+ oxidoreductase (41). 1-Aminoacetone is formed by the breakdown of L-threonine by L-threonine dehydrogenase (42) or is alternatively synthesized from methylglyoxal by monoamine oxidase, which catalyzes the interconversion of methylglyoxal and aminoacetone (43). Methylglyoxal is a by-product of several metabolic pathways, with glycolysis being the most important source (44). Methylglyoxal is synthesized either enzymatically or nonenzymatically from dihydroxyacetone phosphate or glyceraldehyde 3-phosphate (44).

To examine the Ispn synthesis pathway in *E. histolytica*, we cultured amebae in medium supplemented with either methylglyoxal, aminoacetone, or L-threonine and examined the resulting lipid profiles. We found that supplementation with 2 mM methylglyoxal or 4 mM aminoacetone, but not 50 mM L-threonine, led to the synthesis of PtdIspn (Fig. 5*A*). These results are consistent with the premise that *E. histolytica* is capable of Ispn synthesis from methylglyoxal and possesses the

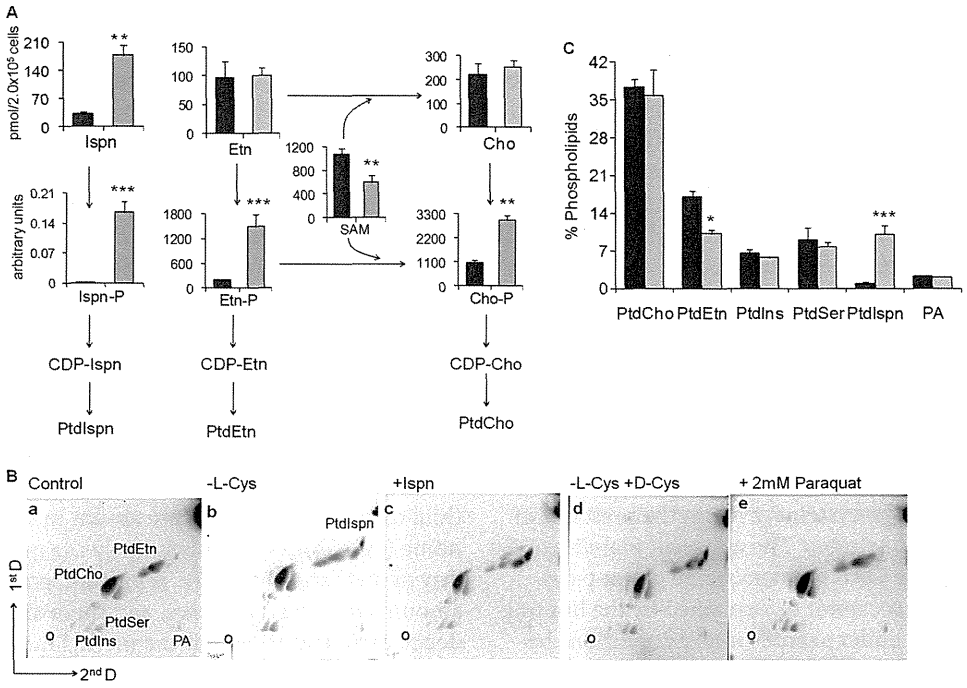


FIGURE 4. **L-Cysteine depletion affected phospholipid metabolism.** *A*, effects of L-cysteine depletion on the levels of metabolites involved in the Kennedy pathway of phospholipid metabolism. Trophozoites were cultured in normal (*black bars*) or cysteine-deprived medium for 48 h (*gray bars*). The average contents (pmol) ± S.D. (*error bars*) in 2×10^5 cells in triplicate are shown. *Asterisks* (*, **, and ***) denote statistically significant differences with $p \leq 0.05$, $p \leq 0.01$, and $p \leq 0.001$, respectively, as determined by Student's *t* test. *B*, profiles of phospholipids derived from trophozoites cultured under various conditions (*panels a–e*), analyzed by two-dimensional TLC. Trophozoites were cultured in normal culture medium (*panels a*, *c*, and *e*) or L-cysteine-deprived medium supplemented (*panels b* and *d*), without (*panels a* and *b*) or with 5 mM Ispn (*panels c*), or 8 mM D-cysteine (*panels d*) for 48 h or 2 mM paraquat for 10 h (*panels e*). *Circles* in the *bottom left corners* indicate the spots where the samples were applied. *C*, quantitation of the phospholipid species (percentage) derived from trophozoites cultured using normal (*black bars*) or L-cysteine-depleted (*gray bars*) medium. *Asterisks* (* and ***) denote statistically significant differences with $p \leq 0.05$ and $p \leq 0.001$, respectively, as determined by Student's *t* test. *PtdSer*, phosphatidylserine; *PtdIns*, phosphatidylinositol; *PA*, phosphatidic acid.

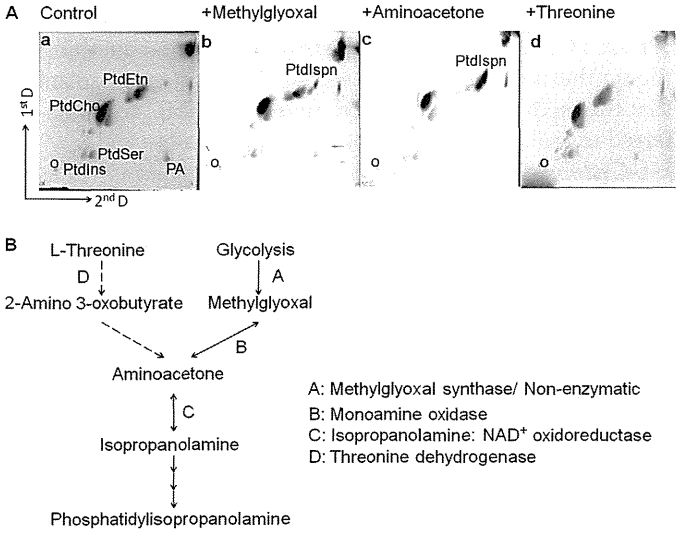


FIGURE 5. **Examination of isopropanolamine biosynthesis in *E. histolytica*.** *A*, effects of the supplementation of potential precursors to the culture medium on the synthesis of phosphatidylisopropanolamine. Trophozoites were cultured in normal culture medium supplemented without (*panel a*) or with 2 mM methylglyoxal (*panel b*), 5 mM aminoacetone (*panel c*), or 50 mM L-threonine for 24 h (*panel d*), and the lipids were then analyzed by two-dimensional TLC. *Circles* in the *bottom left corners* indicate the spots where the samples were applied. *B*, possible pathways of isopropanolamine biosynthesis. The reactions depicted by *solid arrows* were demonstrated in this study, whereas those indicated by *broken arrows* are considered to be absent in *E. histolytica*. *PtdSer*, phosphatidylserine; *PtdIns*, phosphatidylinositol; *PA*, phosphatidic acid.

enzymatic activities of monoamine oxidase and Ispn:NAD^+ oxidoreductase (Fig. 5*B*).

DISCUSSION

Identification of SMC and OAS as the Major Metabolites Increased upon L-Cysteine Deprivation—In the present study, using a CE-TOFMS-based approach (19–21), we identified novel metabolic changes caused by L-cysteine deprivation in the anaerobic/microaerophilic protozoan parasite *E. histolytica*. The major advantages of CE-MS analysis include its extremely high resolution and ability to simultaneously quantify charged low molecular weight compounds (19–21). We demonstrated that L-cysteine deprivation causes a dramatic accumulation of SMC and OAS (Fig. 1*C*). SMC is a sulfur-containing amino acid that has never been detected in protozoa but is widely present in relatively large amounts in several legumes, where it is considered to serve as a sulfur storage compound (36, 37). Using stable isotope-labeled L-serine and L-methionine, we showed that SMC is synthesized from these amino acids in *E. histolytica* via OAS and methanethiol, respectively, which is similar to the pathway reported in *A. thaliana* (36). Interestingly, the increase in both SMC and OAS was mitigated by supplementation of the culture medium with 2 mM sulfide. These results have solved one enigma concerning the biological roles of the sulfur assimilatory *de novo* L-cysteine biosynthetic pathway in *E. histolytica*.

Role of L-Cysteine Biosynthetic Pathway—Although *E. histolytica* is a unique organism that constitutively expresses high

levels of multiple cytosolic isotypes of CS and SAT, the physiological significance of the L-cysteine pathway and its redundancy are not well understood (8, 9, 18). *In vitro* cultivation of amebic trophozoites requires high concentrations of L-cysteine, which cannot be replaced by other thiols (16), indicating that the synthesis pathway may not be sufficient for the production of L-cysteine and might play an unknown role. Our metabolomic study using labeled L-serine did not support the hypothesis that L-cysteine is formed from L-serine and sulfide via OAS by the sequential action of SAT and CS, because labeled L-serine was not incorporated into L-cysteine (data not shown). We also demonstrated that amebic CS isotypes can catalyze the formation of SMC from OAS and methanethiol, unlike the CS from *T. vaginalis* (45), and also possess robust L-cysteine forming activity (Fig. 2*D*).

We also revealed that OAS is exclusively directed for the synthesis of SMC, but not L-cysteine, even in the presence of high concentrations of substrates. The apparent inability of *E. histolytica* to incorporate OAS into L-cysteine under L-cysteine-deprived conditions cannot be explained by the limiting concentration of sulfide under axenic culture conditions because the addition of sulfide did not increase L-cysteine levels, whereas the accumulation of OAS and SMC was immediately ceased by sulfide supplementation (Fig. 2*C*). In fact, the amebic trophozoites cultured under normal conditions contained appreciable concentrations of sulfide (134 μM) (46). The lack of OAS incorporation into L-cysteine is also not attributable to the low substrate specificity of CS toward sulfide, because the K_m values of CS isotypes for sulfide were comparable with those for methanethiolate (Fig. 2*D*). Thus, the preferred utilization of OAS by *E. histolytica* for SMC synthesis, but not for L-cysteine production, suggests that the L-cysteine biosynthetic pathway plays a major role in SMC production, whereas the apparent defect of L-cysteine production by this pathway *in vivo* remains puzzling.

Regulation of OAS Synthesis—The marked increase in OAS observed under L-cysteine deprivation is also worthy of attention. Similar to SMC, the level of OAS under normal culture conditions was nearly undetectable. Unlike other organisms, *E. histolytica* possesses three apparently functionally redundant, cytosolic SAT isozymes (SAT1–3) (9). Because these SAT isozymes have low to high sensitivity to feedback inhibition by L-cysteine, OAS and SMC were presumed to be formed even in the presence of high concentrations of L-cysteine, mainly by L-cysteine-insensitive SAT3. Thus, the fact that OAS and SMC were undetectable in the amebae cultured under normal conditions indicates that the activity of SAT, particularly SAT3, is repressed by unknown mechanisms. The fact that CS activity is a few orders of magnitude higher than SAT activity in the amebae may explain why OAS was not detected under the normal conditions, but it does not explain why SMC is not synthesized. The observed increase in OAS under L-cysteine depletion also indicates that L-cysteine-sensitive SAT1 and SAT2 are derepressed (*i.e.* L-cysteine-mediated feedback inhibition of SAT1/2 was reversed) under L-cysteine depletion.

In addition to the feedback inhibition, the cysteine biosynthetic pathway is also regulated by the bi-enzyme complex of SAT and CS (47). This complex is not involved in the metabolic channeling of OAS from SAT to CS, because OAS freely diffuses out of the complex. The formation of the SAT·CS complex (cysteine synthase complex) was shown to modulate the kinetic parameters of both enzymes. It was shown that the complex was dissociated by the elevated OAS levels (47). It is possible that OAS affects the formation and dissociation of the cysteine synthase complex in *E. histolytica* and the SMC but not L-cysteine forming activity of the complex. Although OAS also acts as an inducer of the L-cysteine regulon in bacteria (48), the gene expression of SAT and CS isotypes was not affected upon L-cysteine depletion in *E. histolytica* (data not shown).

Role and Fate of SMC—Because genes encoding other enzymes that utilize methanethiol as a substrate, such as *O*-acetylhomoserine sulfhydrylase (EC 2.5.1.49) and methanethiol oxidase (EC 1.8.3.4), are absent in the *E. histolytica* genome (10), SMC synthesis is likely the major salvaging pathway of methanethiol. SMC is present in relatively large amounts in several legumes, and there are a few lines of evidence demonstrating that the methyl moiety of SMC is incorporated into various metabolites (methionine, choline, and creatine) and proteins (pectin) (37, 49–51). In plants (*e.g. Brassica pekinensis*), SMC can also be demethylated to generate L-cysteine (52). Metabolic labeling with SMC revealed that L-methionine and L-cysteine are formed from SMC in *Neurospora crassa* grown in low sulfur medium (53). In addition, methionine and cystathionine-auxotrophic mutants of *N. crassa* were able to grow when supplemented with SMC (53). Despite evidence from the studies, the fate of SMC in *E. histolytica* remains to be established because neither labeled L-cysteine nor L-methionine was detected using isotope-labeled serine in our metabolomic analysis. The role of methionine γ-lyase in *E. histolytica* could be to generate methanethiol for the synthesis of SMC. It has been shown in other organisms that the methyl and thiomethyl moieties of SMC are transferred to unidentified metabolites or proteins (49–53).

L-Cysteine Deprivation Affected SAM and Amino Acid Concentrations, Glycolysis, and Energy Generation—L-Cysteine depletion results in reduced levels of SAM, a precursor for polyamine biosynthesis and the essential methyl donor for numerous transmethylation reactions, including DNA methylation. This decrement in the SAM level is likely caused by either the reduction of SAM production by methionine adenosyltransferase or increased utilization of SAM. Because the amount of polyamines, such as putrescine, spermidine (supplemental Fig. S1), and *N*-acetylputrescine (data not shown), remained unchanged, SAM-dependent methylation may have increased upon L-cysteine deprivation. Because the methionine adenosyltransferase activity from various organisms is inhibited by nitric oxide-mediated nitrosylation of the cysteine residues in its active site (54), it is conceivable that methionine adenosyltransferase activity is inhibited by L-cysteine deprivation. The observed increase in L-threonine and L-serine can be attributed to their increased uptake, which is supported by the fact that L-cysteine is a strong inhibitor of L-threonine and L-serine uptake in the BSC-1 epithelial cell line (55).

Response of E. histolytica to L-Cysteine Depletion

We also demonstrated that L-cysteine deprivation repressed glycolysis and energy generation. Upon L-cysteine depletion, pyruvate and other upstream glycolytic intermediates accumulated that appeared to be rerouted toward the associated pathways. For example, the metabolites linked to pyruvate and phosphoenolpyruvate (*i.e.* alanine, malate, and fumarate), 3-phosphoglycerate (*i.e. O*-phosphoserine), and dihydroxyacetone-phosphate (*i.e.* Gly 3-P) increased in response to L-cysteine depletion. In contrast, the level of acetyl-CoA, ethanol, and the major nucleotide triphosphates significantly decreased. In *E. histolytica*, pyruvate is utilized by pyruvate:ferredoxin oxidoreductase, a highly oxygen-sensitive iron-sulfur cluster-containing protein (56). Our data are consistent with the premise that L-cysteine depletion-mediated oxidative stress inactivates pyruvate:ferredoxin oxidoreductase and other redox-sensitive enzymes, which results in the overall reduction in the glycolytic flux and the accumulation of upstream glycolytic intermediates. It has been shown in *E. histolytica* that under oxidative or nitrosative stress, pyruvate, glucose 6-phosphate, and fructose 6-phosphate accumulate, whereas ethanol and ATP decrease (33, 56). Unlike nitrosative stress, L-cysteine depletion for 48 h did not induce apoptosis (data not shown), and the decrease in ATP content appears to be primarily a result of the reduced glycolytic flux. Whole genome microarray analysis has revealed that L-cysteine depletion does not affect the expression of most of the genes involved in energy metabolism, with the exception of phosphoglycerate mutase and malate dehydrogenase, which were slightly down-regulated by 1.6- and 2.1-fold, respectively.[7] Down-regulation of these two genes may also contribute, at least in part, to the overall reduction in the glycolytic flux.

Discovery of Isopropanolamine and PtdIspn Synthesis upon L-Cysteine Deprivation—We have demonstrated for the first time that the Kennedy pathway, the major pathway for phospholipid biosynthesis, is regulated by the level of L-cysteine in *E. histolytica*. L-Cysteine deprivation resulted in the accumulation of an unusual phospholipid, PtdIspn, and also affected the composition and ratio of the major phospholipids. Under L-cysteine-depleted conditions, the synthesis of Ispn, Ispn-P, Etn-P, and Cho-P was elevated, PtdEtn synthesis was down-regulated, and the levels of Etn, Cho, PtdCho, phosphatidylserine, phosphatidylinositol, and phosphatidic acid were unaffected (Fig. 4).

Based on the findings related to phospholipid biosynthesis, we propose the following scheme for the involvement of L-cysteine. When *E. histolytica* is cultured under L-cysteine-depleted conditions, Ispn synthesis is increased. Ispn-P, formed from Ispn and ATP in a reaction catalyzed by Etn/Cho kinase (EHI_148580 (EAL52090.1); EHI_152340 (EAL51511.1)), then competes with Etn-P for Etn-P cytidyltransferase (EHI_095120 (EAL44415.1); EHI_140590 (EAL48799.1)), which appears to be the rate-limiting enzyme for the production of CDP-Ispn. This competition leads to an accumulation of Etn-P and a decrease in PtdEtn. The fact that PtdCho level was not affected by either L-cysteine depletion or Ispn supplementation, whereas Cho-P level increased upon L-cysteine depletion (Fig. 4*A*), suggests that Ispn-P does not compete with Cho-P for Etn-P/Cho-P cytidyltransferase. This observation also indicates that the increase in Cho-P is a consequence of increased production from accumulated Etn-P by SAM-dependent methylation, which may also contribute to the decrement in the SAM level. Alternatively, Ispn-P may compete with Cho-P for Etn-P/Cho-P cytidyltransferase, but the contribution of *de novo* synthesized PtdCho is negligible compared with the PtdCho incorporated from the culture milieu (Fig. 4).

Significance of PtdIspn Production upon L-Cysteine Deprivation—One of the consequences of the L-cysteine-dependent increase in Ispn synthesis is the concomitant increment in Etn-P, which is a known scavenger of free radicals (57). However, the increment of PtdIspn by L-cysteine depletion is not associated with either oxidative stress or changes in the redox status, because D-cysteine did not alleviate the PtdIspn synthesis, and paraquat/air treatment did not increase PtdIspn synthesis.

Similar to PtdEtn, phospahtidyl-propanolamine, an analog of PtdIspn, is a nonbilayer or hexagonal phase-forming phospholipid; however, it is not known whether PtdIspn is also similar to PtdEtn and phospahtidylpropanolamine in its nonbilayer or hexagonal phase-forming nature (58). Hexagonal phase-forming phospholipids have been proposed to be important for membrane fluidity, protein translocation, and membrane fusion events (59, 60). In PtdEtn methylation-defective mutants of *Saccharomyces cerevisiae*, supplementation of the culture medium with propanolamine leads to phosphatidylpropanolamine production and thus presumably compensates for the role of PtdCho and its *N*-methylated phospholipid precursors (58). However, mitochondrial PtdEtn cannot be completely replaced by phosphatidylpropanolamine, suggesting that PtdEtn is essential for the structure and function of mitochondrial membranes (58). It has been shown that changes in the PtdCho/PtdEtn ratio affect membrane integrity of large unilamellar vesicles in mouse hepatocytes, and this ratio is inversely correlated with leakage across the membrane (61). Because L-cysteine depletion also increases the PtdCho/PtdEtn ratio, it is conceivable that this changes membrane integrity and fluidity and affects protein translocation across the plasma membrane. In addition, because PtdEtn also plays various metabolic roles in cells, a decrease in PtdEtn level may also affect other cellular processes, including the synthesis of GPI anchors and protein modification.

To date, this is the first report to show that PtdIspn synthesis is increased by changes in environmental conditions. PtdIspn and phosphatidylpropanolamine have been identified in various organisms, including yeast, protozoa, and animals, and are considered to be unnatural phospholipids synthesized only under conditions where Ispn or propanolamine are supplied in the culture medium (58, 62) or administered intraperitoneally (63). Recently, PtdIspn has been shown to be naturally synthesized in BHK cells through the decarboxylation of the rare phospholipid phosphatidylthreonine (64).

[7] A. Husain, D. Sato, G. Jeelani, M. Suematsu, T. Soga, and T. Nozaki, unpublished data.

In conclusion, we have demonstrated that L-cysteine regulates various metabolic pathways in *E. histolytica* and thus affects the concentrations of the amino acids, phospholipids, and intermediary metabolites involved in central energy metabolism. Further investigation on the physiological role and fate of SMC and PtdIspn will help to better understand sulfur-containing amino acid metabolism, which is considered an attractive drug target for the development of new chemotherapeutics against this pathogen (18, 65). Future research is also needed to understand the function of PtdIspn in the plasma membrane and membrane-bound organelles and in the regulation of phospholipid metabolism.

Acknowledgments—We thank Takako Hishiki (Keio University) for the initial acquisition and analysis of CE-MS data and helpful discussions, Masahiro Sugimoto and Akiyoshi Hirayama (Keio University) for the use of CE-MS data analysis software (MasterHands), and all of the members of our laboratory for technical assistance and valuable discussions.

REFERENCES

1. Beinert, H., Holm, R. H., and Münck, E. (1997) *Science* **277,** 653–659
2. Stanley, S. L., Jr. (2003) *Lancet* **361,** 1025–1034
3. Weinbach, E. C., and Diamond, L. S. (1974) *Exp. Parasitol.* **35,** 232–243
4. Mehlotra, R. K. (1996) *Crit. Rev. Microbiol.* **22,** 295–314
5. Nozaki, T., Asai, T., Kobayashi, S., Ikegami, F., Noji, M., Saito, K., and Takeuchi, T. (1998) *Mol. Biochem. Parasitol.* **97,** 33–44
6. Clark, C. G., Alsmark, U. C., Tazreiter, M., Saito-Nakano, Y., Ali, V., Marion, S., Weber, C., Mukherjee, C., Bruchhaus, I., Tannich, E., Leippe, M., Sicheritz-Ponten, T., Foster, P. G., Samuelson, J., Noël, C. J., Hirt, R. P., Embley, T. M., Gilchrist, C. A., Mann, B. J., Singh, U., Ackers, J. P., Bhattacharya, S., Bhattacharya, A., Lohia, A., Guillén, N., Duchêne, M., Nozaki, T., and Hall, N. (2007) *Adv. Parasitol.* **65,** 51–190
7. Nozaki, T., Asai, T., Sanchez, L. B., Kobayashi, S., Nakazawa, M., and Takeuchi, T. (1999) *J. Biol. Chem.* **274,** 32445–32452
8. Nozaki, T., Ali, V., and Tokoro, M. (2005) *Adv. Parasitol.* **60,** 1–99
9. Hussain, S., Ali, V., Jeelani, G., and Nozaki, T. (2009) *Mol. Biochem. Parasitol.* **163,** 39–47
10. Loftus, B., Anderson, I., Davies, R., Alsmark, U. C., Samuelson, J., Amedeo, P., Roncaglia, P., Berriman, M., Hirt, R. P., Mann, B. J., Nozaki, T., Suh, B., Pop, M., Duchene, M., Ackers, J., Tannich, E., Leippe, M., Hofer, M., Bruchhaus, I., Willhoeft, U., Bhattacharya, A., Chillingworth, T., Churcher, C., Hance, Z., Harris, B., Harris, D., Jagels, K., Moule, S., Mungall, K., Ormond, D., Squares, R., Whitehead, S., Quail, M. A., Rabbinowitsch, E., Norbertczak, H., Price, C., Wang, Z., Guillén, N., Gilchrist, C., Stroup, S. E., Bhattacharya, S., Lohia, A., Foster, P. G., Sicheritz-Ponten, T., Weber, C., Singh, U., Mukherjee, C., El-Sayed, N. M., Petri, W. A., Jr., Clark, C. G., Embley, T. M., Barrell, B., Fraser, C. M., and Hall, N. (2005) *Nature* **433,** 865–868
11. Tokoro, M., Asai, T., Kobayashi, S., Takeuchi, T., and Nozaki, T. (2003) *J. Biol. Chem.* **278,** 42717–42727
12. Sato, D., Yamagata, W., Harada, S., and Nozaki, T. (2008) *FEBS J.* **275,** 548–560
13. Fahey, R. C., Newton, G. L., Arrick, B., Overdank-Bogart, T., and Aley, S. B. (1984) *Science* **224,** 70–72
14. Gillin, F. D., and Diamond, L. S. (1980) *J. Protozool.* **27,** 474–478
15. Gillin, F. D., and Diamond, L. S. (1981) *Exp. Parasitol.* **52,** 9–17
16. Gillin, F. D., and Diamond, L. S. (1981) *Exp. Parasitol.* **51,** 382–391
17. Jeelani, G., Husain, A., Sato, D., Ali, V., Suematsu, M., Soga, T., and Nozaki, T. (2010) *J. Biol. Chem.* **285,** 26889–26899
18. Ali, V., and Nozaki, T. (2007) *Clin. Microbiol. Rev.* **20,** 164–187
19. Soga, T., Baran, R., Suematsu, M., Ueno, Y., Ikeda, S., Sakurakawa, T., Kakazu, Y., Ishikawa, T., Robert, M., Nishioka, T., and Tomita, M. (2006) *J. Biol. Chem.* **281,** 16768–16776
20. Sato, S., Soga, T., Nishioka, T., and Tomita, M. (2004) *Plant J.* **40,** 151–163
21. Soga, T., Ohashi, Y., Ueno, Y., Naraoka, H., Tomita, M., and Nishioka, T. (2003) *J. Proteome Res.* **2,** 488–494
22. Diamond, L. S., Harlow, D. R., and Cunnick, C. C. (1978) *Trans. R. Soc. Trop. Med. Hyg.* **72,** 431–432
23. Clark, C. G., and Diamond, L. S. (2002) *Clin. Microbiol. Rev.* **15,** 329–341
24. Ohashi, Y., Hirayama, A., Ishikawa, T., Nakamura, S., Shimizu, K., Ueno, Y., Tomita, M., and Soga, T. (2008) *Mol. Biosyst.* **4,** 135–147
25. Soga, T., and Heiger, D. N. (2000) *Anal. Chem.* **72,** 1236–1241
26. Soga, T., Ueno, Y., Naraoka, H., Ohashi, Y., Tomita, M., and Nishioka, T. (2002) *Anal. Chem.* **74,** 2233–2239
27. Soga, T., Igarashi, K., Ito, C., Mizobuchi, K., Zimmermann, H. P., and Tomita, M. (2009) *Anal. Chem.* **81,** 6165–6174
28. Sugimoto, M., Wong, D. T., Hirayama, A., Soga, T., and Tomita, M. (2010) *Metabolomics* **6,** 78–95
29. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006) *Anal. Chem.* **78,** 779–787
30. Baran, R., Kochi, H., Saito, N., Suematsu, M., Soga, T., Nishioka, T., Robert, M., and Tomita, M. (2006) *BMC Bioinformatics* **7,** 530
31. Bai, J., Rodriguez, A. M., Melendez, J. A., and Cederbaum, A. I. (1999) *J. Biol. Chem.* **274,** 26217–26224
32. Aceti, D. J., and Ferry, J. G. (1988) *J. Biol. Chem.* **263,** 15444–15448
33. Ramos-Martínez, E., Olivos-García, A., Saavedra, E., Nequiz, M., Sánchez, E. C., Tello, E., El-Hafidi, M., Saralegui, A., Pineda, E., Delgado, J., Montfort, I., and Pérez-Tamayo, R. (2009) *Int. J. Parasitol.* **39,** 693–702
34. Bligh, E. G., and Dyer, W. J. (1959) *Can. J. Biochem. Physiol.* **37,** 911–917
35. Zhou, X., and Arthur, G. (1992) *J. Lipid Res.* **33,** 1233–1236
36. Rébeillé, F., Jabrin, S., Bligny, R., Loizeau, K., Gambonnet, B., Van Wilder, V., Douce, R., and Ravanel, S. (2006) *Proc. Natl. Acad. Sci. U.S.A.* **103,** 15687–15692
37. Giovanelli, J., Mudd, S. H., and Datko, A. H. (1980) in *The Biochemistry of Plants* (Miflin B. J., eds.), Vol. 5, pp. 453–487, Academic Press, New York
38. Zuo, X., and Coombs, G. H. (1995) *FEMS Microbiol. Lett.* **130,** 253–258
39. Husain, A., Jeelani, G., Sato, D., Ali, V., and Nozaki, T. (2010) *Mol. Biochem. Parasitol.* **170,** 100–104
40. Anderson, I. J., and Loftus, B. J. (2005) *Exp. Parasitol.* **110,** 173–177
41. Kelley, J. J., and Dekker, E. E. (1984) *J. Biol. Chem.* **259,** 2124–2129
42. Green, M. L., and Elliott, W. H. (1964) *Biochem. J.* **92,** 537–549
43. Ray, M., and Ray, S. (1987) *J. Biol. Chem.* **262,** 5974–5977
44. Inoue, Y., and Kimura, A. (1995) *Adv. Microb. Physiol.* **37,** 177–227
45. Westrop, G. D., Goodall, G., Mottram, J. C., and Coombs, G. H. (2006) *J. Biol. Chem.* **281,** 25062–25075
46. Ariyanayagam, M. R., and Fairlamb, A. H. (1999) *Mol. Biochem. Parasitol.* **103,** 61–69
47. Wirtz, M., Birke, H., Heeg, C., Mueller, C., Hosp, F., Throm, C., Koenig, S., Feldman-Salit, A., Rippe, K., Petersen, G., Wade, R. C., Rybin, V., Scheffzek, K., and Hell, R. (2010) *J. Biol. Chem.* **285,** 32810–32817
48. Kredich, N. M. (1992) *Mol. Microbiol.* **6,** 2747–2753
49. Horner, W. H., and Kuchinskas, E. J. (1959) *J. Biol. Chem.* **234,** 2935–2937
50. Ronald, C. D., and John, F. T. (1971) *Phytochemistry* **10,** 1745–1750
51. Mae, T., and Ohira, K. (1976) *Plant Cell Physiol.* **17,** 459–465
52. Mae, T., Ohira, K., and Fujiwara, A. (1971) *Plant Cell Physiol.* **12,** 881–887
53. Wiebers, J. L., and Garner, H. R. (1964) *J. Bacteriol.* **88,** 1798–1804
54. Pérez-Mato, I., Castro, C., Ruiz, F. A., Corrales, F. J., and Mato, J. M. (1999) *J. Biol. Chem.* **274,** 17075–17089
55. Kuhlmann, M. K., and Vadgama, J. V.(1991) *J. Biol. Chem.* **266,** 15042–15047
56. Ramos, E., Olivos-García, A., Nequiz, M., Saavedra, E., Tello, E., Saralegui, A., Montfort, I., and Pérez Tamayo, R. (2007) *Exp. Parasitol.* **116,** 257–265
57. Gordon, L. I., Weiss, D., Prachand, S., and Weitzman, S. A. (1991) *Free Radic. Res. Commun.* **15,** 65–71
58. Choi, J. Y., Martin, W. E., Murphy, R. C., and Voelker, D. R. (2004)

J. Biol. Chem. **279,** 42321–42330
59. Yeagle, P. L. (1989) *FASEB J.* **3,** 1833–1842
60. Cullis, P. R., Fenske, D. B., and Hope, M. J. (1996) in *Biochemistry of Lipids, Lipoproteins and Membranes* (Vance, D. E., and Vance, J., eds) Vol. 31, pp. 1–33, Elsevier, Paris
61. Li, Z., Agellon, L. B., Allen, T. M., Umeda, M., Jewell, L., Mason, A., and Vance, D. E. (2006) *Cell Metab.* **3,** 321–331
62. Smith, J. D., and Barrows, L. J. (1988) *Biochem. J.* **254,** 301–302
63. Meyer, W., Wahl, R., and Gercken, G. (1979) *Biochim. Biophys. Acta* **575,** 463–466
64. Heikinheimo, L., and Somerharju, P. (2002) *Traffic* **3,** 367–377
65. Sato, D., Kobayashi, S., Yasui, H., Shibata, N., Toru, T., Yamamoto, M., Tokoro, G., Ali, V., Soga, T., Takeuchi, T., Suematsu, M., and Nozaki, T. (2010) *Int. J. Antimicrob. Agents* **35,** 56–61

Masahiro Sugimoto[1,2]
Akiyoshi Hirayama[1]
Martin Robert[1,2]
Shinobu Abe[1]
Tomoyoshi Soga[1,2]
Masaru Tomita[1,2]

[1]Institute for Advanced Biosciences, Keio University, Yamagata, Japan
[2]Systems Biology Program, Graduate School of Media and Governance, Keio University, Kanagawa, Japan

Received September 28, 2009
Revised December 3, 2009
Accepted December 19, 2009

Research Article

Prediction of metabolite identity from accurate mass, migration time prediction and isotopic pattern information in CE-TOFMS data

CE-TOFMS is a powerful method for profiling charged metabolites. However, the limited availability of metabolite standards hinders the process of identifying compounds from detected features in CE-TOFMS data sets. To overcome this problem, we developed a method to identify unknown peaks based on the predicted migration time (t_m) and accurate m/z values. We developed a predictive model using 375 standard cationic metabolites and support vector regression. The model yielded good correlations between the predicted and measured t_m ($R = 0.952$ and 0.905 using complete and cross-validation data sets, respectively). Using the trained model, we subsequently predicted the t_m for 2938 metabolites available from the public databases and assigned tentative identities to noise-filtered features in human urine samples. While 38.9% of the peaks were assigned metabolite names by matching with the standard library alone, the proportion increased to 52.2%. The proposed methodology increases the value of metabolomic data sets obtained from CE-TOFMS profiling.

Keywords:
CE-TOFMS / Metabolite identification / Metabolome / Non-target analysis
DOI 10.1002/elps.200900584

1 Introduction

Advances in metabolomics, the acquisition of knowledge regarding the quantitative metabolic response of complex living systems to various stimuli, are attributable to concomitant innovations in high-throughput profiling technologies. CE-MS, a powerful platform to simultaneously quantify numerous charged metabolites, is now widely used in metabolomic research [1]. TOFMS is commonly used as a detector for CE-MS, because it provides high mass accuracy and an excellent resolution for the simultaneous measurement of metabolites in a wide continuous range (*e.g.* 50–1000 m/z). CE-TOFMS is a powerful method for the collection of comprehensive metabolite profiles, *i.e.* non-targeted profiling.

The lack of commercially available metabolites limits the identification of features detected by CE-MS [2]. A library search strategy based on mass fragmentation patterns using MS/MS is best suited to identifying available molecules for which MS/MS spectra have been collected in databases. Peptides are often neglected in metabolomic analyses, but their repeated amino acid structure and relatively well-characterized fragmentation patterns, combined with the availability of sequence databases, renders their identification much simpler [3–6]. Often, only very limited structural information can be obtained from MS/MS spectra of singly charged small molecules. Thus, systematic and high-throughput identification by MS/MS continues to be difficult. Therefore, although most reports based on CE-MS provide more targeted metabolite analysis, few present this information in the context of comprehensive or non-targeted metabolome analysis [7].

Despite the good mass accuracy of TOFMS, MS information alone (accurate m/z and isotope distribution) is often insufficient for the metabolite identification [8, 9]. At best, it yields a list of potential candidates fitting specific molecular formulae compatible with the observed spectra. Thus, utilization of the t_m or electrophoretic mobility of individual

Correspondence: Dr. Masahiro Sugimoto, Institute for Advanced Biosciences, Keio University, 246-2 Mizukami, Kakuganji, Tsuruoka, Yamagata 997-0017, Japan
E-mail: msugi@sfc.keio.ac.jp; mshrsgmt@gmail.com
Fax: +81-235-29-0574

Abbreviations: CFSS, correlation-based feature subset selection; **CV,** cross-validation; **HMDB,** human metabolome database; **MLR,** multiple linear regression; **RFAE,** relief for attribute estimation; **SVM,** support vector machine; **SVR,** support vector regression

peaks in CE may considerably limit the number of metabolite candidates. Mathematical techniques to calculate the electrophoretic migration of peptides in CE and CE-MS are available to optimize CE measurement conditions or to identify peptides [10–14]. Because several models were developed with respect to electrophoretic mobility and involved a limited subset of small molecules of similar structure, *e.g.* monoamines [15–18], their prediction accuracy was limited and more versatile/accurate methods are needed for the prediction of most types of metabolites. Prediction techniques to assess electrophoretic mobility from molecular descriptors using artificial neural networks, support vector regression (SVR) [11, 13, 15–17, 19], and computational simulations of the electrophoretic mobility of cationic amino acids in CE-MS [18, 20] have also been developed.

The aim of this study was to facilitate metabolite identification from CE-MS profiles by combining t_m prediction modeling with MS spectral data. To address this issue, we developed a mathematical model based on SVR for migration time prediction of CE-MS peaks. Using our model, we tentatively annotated peaks from CE-MS analysis of human urine samples.

2 Materials and methods

2.1 Sample collection and preparation

Urine samples were collected from a healthy volunteer. Aliquots (20, 40 and 100 μL) were concentrated to dryness using a centrifugal concentrator. The residue was then dissolved in 20 μL of MilliQ water containing 200 μM each of methionine sulfone and 3-aminopyrrolidine as the internal standards and subjected to a CE-TOFMS analysis. The 40 and 100 μL samples were duplicated and a total of five samples was analyzed in duplicate.

2.2 Metabolite standards, instrumentation and CE-TOFMS conditions

The metabolite standards for small molecules, the instrumentation and the CE-TOFMS conditions have been described elsewhere [21]. All chemical standards were of analytical or reagent grade and were obtained from commercial sources and dissolved in MilliQ water (Millipore, Bedford, MA, USA), 0.1 N HCl or 0.1 N NaOH to obtain 1, 10 or 100 mM stock solutions. The working solution was prepared before use by diluting with MilliQ water to the appropriate concentration.

All CE-MS experiments were performed using an Agilent CE capillary electrophoresis system (Agilent Technologies, Waldbronn, Germany), an Agilent G3250AA LC/MSD TOF system (Agilent Technologies, Palo Alto, CA, USA), an Agilent 1100 series binary HPLC pump, the G1603A Agilent CE-MS adapter and G1607A Agilent CE-ESI-MS sprayer kit. The Agilent 1100 series pump equipped with a 1:100 splitter was used to deliver the sheath liquid. System control and data acquisition were performed using G2201AA Agilent Chemstation software for CE and Analyst QS software for TOFMS (ver. 1.1).

Samples were separated in fused-silica capillaries (50 μm id × 100 cm total length) filled with 1 M formic acid (pH 1.8) as the BGE. The sample solutions were injected at 50 mbar for 3 s and a voltage of 30 kV was applied. The capillary temperature was maintained at 20°C and the temperature of the sample tray was kept below 5°C using the external coolant system. The sheath liquid, comprising methanol/water (50% v/v) and 0.1 μM hexakis-(2,2-difluorothoxy)-phosphazene (Hexakis), was delivered at 10 μL/min. ESI-TOFMS was conducted in the positive ion mode. The capillary voltage was set at 4 kV and the nitrogen gas flow rate (300°C) was 10 L/min. In TOFMS, the fragmentor, skimmer and OCT RFV voltage were set at 75, 50 and 125 V, respectively. An automatic recalibration of each acquired spectrum was performed using the reference masses of reference standards ([^{13}C isotopic ion of protonated methanol dimer (2MeOH+H)]$^+$, m/z 66.063706) and ([protonated Hexakis (M+H)]$^+$, m/z 622.028963). Mass spectra were acquired at a rate of 1.5 cycles/s over a 50–1000 m/z range.

In this study, 245 nonpeptide metabolites and 130 peptides were used as the standard library. Their m/z range was from 3-aminopropionitrile (71.0604 m/z) to 3,3′,5-triiodothyronine (651.7931 m/z) and from Gly–Gly (133.0608 m/z) to Lys–Typ–Lys (461.2871 m/z), respectively.

2.3 Processing of CE-TOFMS data

Raw data obtained by CE-TOFMS were processed with our proprietary software, MasterHands [22]. The data analysis workflow started with the raw data and included noise filtering, baseline correction, peak detection, deconvolution and integration of the peak area from sliced electropherograms with a 0.02 m/z width. The accurate m/z value for each detected peak was calculated with Gaussian curve fitting to the peak along the mass spectrum axis.

The schematic metabolite identification strategy is depicted in Fig. 1. First, we removed the noisy and atypical part of the CE-MS data, *e.g.* neutral compounds and salt ions related to Na^+ and K^+ observed under our measurement conditions. Second, redundant features such as fragments, adducts, isotopes, dimers, trimers and their combinations, *e.g.* the adduct ion of a dimer, were eliminated based on established m/z differences [23]. Spike noise, CE-specific noise with small and narrow peaks and low-reliability signals (no peak-like shape) were also eliminated. We ignored multivalent peaks, which can be estimated based on the m/z difference between a peak and its isotope peaks. Third, metabolite names were assigned to peaks based on the matched m/z values and the normalized t_m of the standard compounds. Finally, the other peaks were

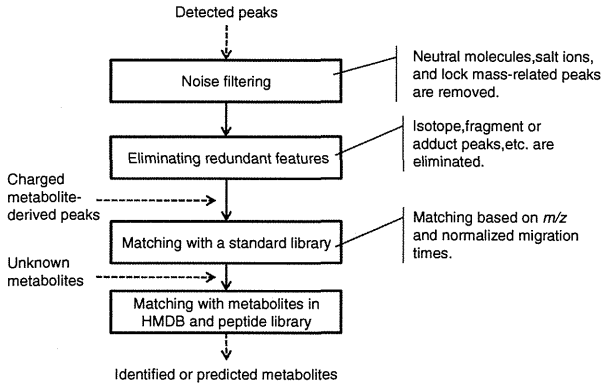


Figure 1. Flow chart of data processing after peak detection and assignment of metabolite names. The first three steps select only the charged metabolites. The last two steps show the procedure for assigning metabolite names to the selected peaks. Metabolites are first matched with the standard library and HMDB and subsequently peptide names were assigned. In these procedures, we used *m/z* and normalization migration measured or predicted by SVR.

annotated based on the *m/z* value and normalized t_m as predicted by the SVR model. Candidates for small nonpeptide metabolites and short peptides were obtained from the Human Metabolome Database (HMDB) [24] and the ARM project web site (http://www.metabolome.jp/), respectively. The composition formulae were calculated from the observed accurate mass and the isotopic distribution using MassHunter (ver. B.01.03, Agilent Technologies).

2.4 Normalization of run-to-run errors in t_m in CE-MS data

The electrophoretic mobility of the charged molecules is proportional to the net charge and is inversely proportional to the product of buffer viscosity and the effective ion radius of the compound expressed by

$$\mu_{ef}=\frac{q}{6\pi\eta r} \tag{1}$$

where μ_{ef} is the electrophoretic mobility, q the net charge, η the solution viscosity and r the effective ion radius (also known as the Stoke radius) of the compound [25]. The electrophoretic mobility for charged compounds is described as

$$\mu = a\frac{q}{M^b} \tag{2}$$

where a is the parameter fitted to experimentally observed data sets and b is the constant coefficient (*e.g.* 1/2 or 2/3) or several available mathematical functions (reviewed in [13]).

Although electrophoretic mobility can be accurately calculated using these mathematical equations, there are two complications in CE-MS. The shape of the marker peak for measuring EOF is not always clear because of the co-migration of a number of neutral molecules with the EOF; this decreases the accuracy of the t_{eof} measurements. Moreover, the fluctuation of the electrical conditions during a measurement might cause local shifts in electrolytes in CE [22], which prevents accurate mobility measurements. Therefore, we directly assessed the t_m rather than electrophoretic mobility. Here, we used two internal standards as a reference to normalize the t_m of subject samples [26]:

$$\alpha = \frac{\frac{1}{t_1}-\frac{1}{t_2}}{\frac{1}{t_{1R}}-\frac{1}{t_{2R}}} \tag{3}$$

$$\beta = \frac{\frac{1}{t_1}+\frac{1}{t_2}}{\alpha} - \left(\frac{1}{t_{1R}}+\frac{1}{t_{2R}}\right) \tag{4}$$

$$t_{normalized} = \frac{1}{\left(\frac{1}{\alpha t_{exp}}-\frac{\beta}{2}\right)} \tag{5}$$

where t_1 and t_2 are the t_m of the internal standards in the subject samples, t_{1R} and t_{2R} the internal standards in the standard library and t_{exp} and $t_{normalized}$ the observed and normalized t_m values, respectively, in the subject samples [26]. For the internal standards, we used 3-aminopyrrolidine (*m/z* 87.0917 $[M+H]^+$, $pK_a = 10.53 \pm 0.4$ and 7.06 ± 0.2) and methionine sulfone (*m/z* 182.0482 $[M+H]^+$, $pK_a = 8.73 \pm 0.16$ and 1.2 ± 0.1).

2.5 Preparation and selection of molecular descriptors relevant to electrophoretic migration

The development of a mathematical model for predicting the t_m starts with the collection of the molecular structure, optimizing the structure, calculating molecular descriptors, variable selection and training the predictive models with selected variables using experimental data [14]. To calculate the molecular descriptors, we used Molecular Operating Environment 2008.10 (Chemical Computing Group, 1225 University Street, Montreal, QC, Canada) to yield more than 100 molecular descriptors. To calculate the net charge of the nonpeptide metabolites in the standard library and HMDB, we used Henderson–Hasselbach equations [27] and acid dissociation constant pK_a values predicted by pK_a DB (ACD/pK_a DB, version 7.06, Advanced Chemistry Development, Toronto ON, Canada; www.acdlabs.com, 2003). Based on the pK_a values and the buffer pH, the net charge q for nonpeptide metabolites was calculated by

$$q = \sum_i^m \frac{1}{1+10^{(pH-pK_{ai})}} - \sum_j^n \frac{1}{1+10^{(pK_{aj}-pH)}} \tag{6}$$

where m and n are the numbers of acidic groups and basic groups, respectively [10, 19]. However, the collection of pK_a from molecular structures is time consuming. Thus, the net charge of a peptide was alternatively calculated as the sum of the net charges of the individual charged amino acid residues and the N- and C-terminals using a previously described method [13].

Before the development of predictive models, feature selection is commonly performed to eliminate redundant

variables; this enhances the versatility and interpretability of the developed model [14]. Here, we used the stepwise forward and backward feature selection method, which appends new features at $p<0.05$ and eliminates features at $p>0.05$ (JMP Version 7; SAS Institute, Cary, NC, 1989–2007; http://www.jmp.com/software/jmp.shtml). We also applied feature selection methods that can deal with continuous outputs; correlation-based feature subset selection (CFSS) and relief for attribute estimation (RFAE) implemented in Weka (version 3.6.0, The University of Waikato, Hamilton, New Zealand) [28]. CFSS evaluates subsets of features by considering the individual predictive ability and feature redundancy. RFAE evaluates features by resampling instances and considering the value of the given features for the nearest instance of the class attributes [28, 29]. For feature selection, we used the BestFirst search for CFSS, starting with an empty set of variables and finding the best variable subset by adding variables and the Ranker search for RFAE with default settings. This approach produced a ranked variable list based on the significance of the variables [28].

2.6 Prediction of t_m using SVR

Support vector machine (SVM), originally devised by Vapanik, is a method to classify given data sets in high-dimensional spaces [30]. SVM offers the advantage of reducing the possibility of overfitting to a specific problem and of being robust with relatively small data sets [31]. SVR is an extended version of SVM to accommodate regression problems [32]. As the detailed theory and the SVR algorithm are available elsewhere [15, 17, 32], we have given a brief description related to t_m prediction in CE-MS. SVR solves regression problems by mapping the given data sets to high-dimensional spaces through the user-defined kernel function. To address the regression problem of continuous variables, the radius basis function kernel is commonly used

$$\exp\left(-\gamma * |u - v|^2\right) \quad (7)$$

where u and v are the independent parameters and γ a constant coefficient for the control of the amplification of the Gauss function. The prediction accuracy of SVR predominantly depends on parameter γ in kernel functions and two other parameters implemented in SVR itself; C defines the trade-off between training errors and margins and ε controls the tube of the regression pipe. These parameters should be tuned depending on the data sets used [15, 17]. In this study, we performed tenfold cross-validation (CV) to optimize the parameter set yielding the best correlation coefficient between predicted and actual values. Pearson correlation coefficients were determined to evaluate the prediction accuracy as the association between observed and predicted normalized t_m.

All variables, including molecular descriptors, net charge and t_m, were transformed by a linear normalization into values between 0.1 and 0.9, as previously described [17]. A brute-force search was conducted in the range 0.01, 0.05, 0.1, 0.5, 1.0, 5.0 and 10.0 for C, 0.1, 0.2,... and 1.0 for γ and 0.00, 0.02,..., 0.18 for ε.

3 Results

3.1 Results from feature selection methods

The selected variables obtained by the three feature selection methods are listed in Supporting Information Table S1. The matrix of the correlation coefficients among the selected variables is summarized in Supporting Information Table S2. Stepwise feature selection (Supporting Information Table S2A) and CFSS (Supporting Information Table S2B) provided 39 and 8 variables, respectively. In contrast, RFAE did not select the subset of variables but yielded ranks for all the variables based on their significance. The a_base and vsa_base, ranked as 15th and 17th by RFAE, showed a high correlation coefficient ($R = 0.952$), only variables ranked within the top 16 were used for subsequent analyses (Supporting Information Table S2C). Stepwise feature selection did not eliminate strongly correlated variables (Supporting Information Table S2A). Since the 8th-ranked b_1rotR and the 9th-ranked b_rotR showed a high correlation coefficient ($R = 0.97$), we used only variables within the top eight, including net charge, number of oxygen atoms (a_nO), partial charge descriptors (PEOE_VSA+4 and PEOE_VSA-2), log of the aqueous solubility, absolute value of the difference of the water-accessible surface area of all atoms between the positive and negative partial charge (DASA), contact distance of the hydrophobic volume (vsurf_DD23) and mass density (Dens).

Unsurprisingly, with respect to the selected parameters (Supporting Information Table S1), the net charge was selected as the factor with the highest significance in all three methods; this finding was consistent with that of a previous study [19]. To calculate the net charge of the peptide, we determined the net charge of each charged amino acid residue, the N- and C-terminal at pH 1.8, using the Henderson–Hasselbach equation and pK_a values obtained from pK_a DB. Arginine (R), histidine (H), lysine (K) and the N-terminal contribute a charge of +1 [13]. Aspartic acid (D) (pK_a 2.28), glutamic acid (E) (pK_a 2.17) and the C-terminal (pK_a 3.2) [13] contribute charges of −0.0079, −0.0017 and −0.0383, respectively. The net charge is then calculated as the sum of these contributions [13]. For example, the net charge of the dipeptide HD is calculated as 1.9538 = +1 (N-terminal) +1 (H) −0.0079 (D) −0.0383 (C-terminal). DADA (conformation-dependent charge descriptors), PEOE_VSA (partial charge descriptors), vsurf (surface area, volume and shape descriptors) and the number of specific atoms (atom and bond counts) were commonly selected by the three methods. The variable Dens, only selected by CFSS, the molecular weight divided by the van der Waals volume, was consistent with the predictive models for the electrophoretic mobility of the peptide based

on Offord's model or its extended models using the molecular weight [10–13, 16, 33]. Except for net charge, different subsets were found by the individual feature selection methods. Thus, it is difficult to justify other variables.

3.2 Prediction of t_m by SVR and MLR

The scatter plots in Fig. 2 depict the predicted and actual normalized t_m values using the parameter set $\gamma = 0.3$, $C = 10.0$ and $\varepsilon = 0.02$ with RFAE, which showed the best correlation coefficient in the CV procedure in all cases and the correlation coefficients $R_t = 0.952$ (Fig. 2(A)) and $R_{cv} = 0.905$ (Fig. 2(B)), obtained using the complete data sets and CV, respectively, when the degree of freedom was 14 and the compound number was 375. The SD of the error between predicted and measured t_m was $E_t = 0.937$ min using complete data sets and $E_{cv} = 1.30$ min using the CV data set. We obtained a good predictive accuracy for the peptide t_m ($R_t = 0.957$, $E_t = 0.613$ min and $R_{cv} = 0.950$, $E_{cv} = 0.788$ min); they were generally better than those of other nonpeptide molecules ($R_t = 0.931$, $E_t = 1.07$ min and $R_{cv} = 0.897$, $E_{cv} = 1.50$ min). Accuracy was worse for nonpeptide molecules in late t_m, *i.e.* those with a slow electrophoretic mobility.

Figure 3 shows the quantitative association between the SVR parameters (γ, C and ε) and the correlation coefficient obtained in CV. As a general trend for all feature selection methods shown in Fig. 3(A), increase in ε decreased the correlation coefficients, and CFSS showed slightly lower sensitivity for ε compared with the other methods. Overall, the inclusion of no features or stepwise feature selection with eight variables showed lower correlation coefficients than the other methods; as ε increased, the correlation coefficients decreased. When ε is more than 0.10, CFSS showed better correlation coefficients for stepwise selection with 39 variables and RFAE showed even better results. Although the R_{cv} among CFSS, RFAE and stepwise selection with 39 variables were not significantly different,

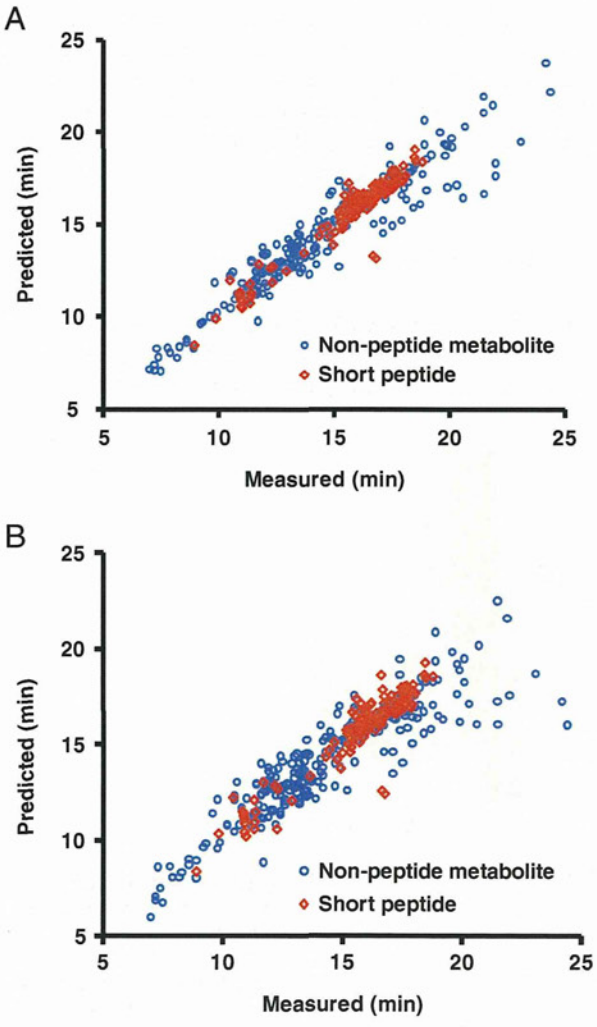


Figure 2. Correlation between predicted and measured normalized t_m values. The plots in (A) were obtained using the complete data sets and plots in (B) were obtained with the CV procedure. The red and blue signals depict the results for peptides and nonpeptide small molecules, respectively.

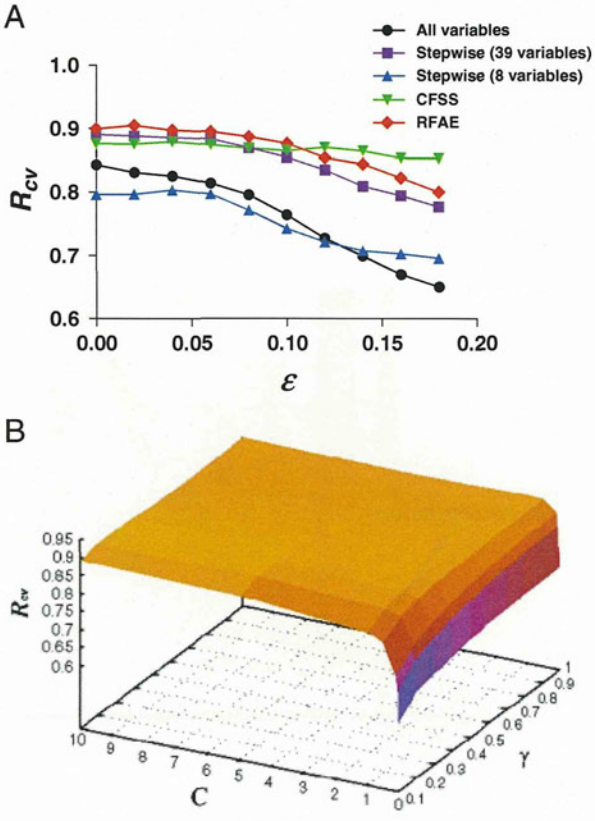


Figure 3. Quantitative relationship between correlation coefficients obtained with the CV procedure and the parameters used to develop SVR models. (A) Shows the correlation coefficient *R* and parameter ε. The other parameters, γ and *C*, were selected to generate the best *R* values for individual cases. (B) Shows the relationship among *R*, γ and *C* at RFAE and $\varepsilon = 0.02$, which yielded the best *R* value for all cases.

regardless of the ε value, when ε was 0.02, RFAE showed the strongest correlation coefficient ($R_{cv} = 0.905$). Figure 3(B) shows the correlation coefficients and the parameters γ and C for the same condition. Despite the γ values, R_{cv} remained relatively constant, except when C was close to 0. The R_{cv} value showed the best performance at around $C = 10$ and $\gamma = 0.3$ and was relatively constant, implying that the developed model was robust against parameter variation around the optimum point.

As an alternative solution for regression problems besides SVR, multiple linear regression (MLR) is the most common technique to capture linear relationships. The comparative experiments with MLR are summarized in Supporting Information Table S3. Although MLR was established with several feature selection models, there was a marked deterioration in the CV procedure. CFSS generally showed lower accuracy while RFAE with SVR showed the highest correlation in the CV procedures.

3.3 Metabolite annotation of peaks in biological samples

In the five measurement runs, an average of 1551 peaks (±182 peaks, SD) was detected in the human urine samples. Of the detected peaks, 460 were selected as nonredundant signals exhibiting metabolite-like features, such as isotopic and fragment peaks. An example total ion electropherogram and a two-dimensional map (the *X*- and *Y*-axes are migration time and *m/z*, respectively) is shown in Fig. 4.

The SD of the t_m for individual peaks between the five runs decreased from 0.260 to 0.00886 min by the alignment procedure (time normalization and peak matching) using Eqs. (3)–(5). The associations between t_m and the calculated SDs is shown in Fig. 5. Although the t_m variation deviation near the internal standard at 7.27 (3-aminopyrrolidine) and 19.31 min (methionine sulfone) decreased after the alignment step, the peaks far from the internal standard along the time axis, particularly the peaks observed at late t_m, were less accurately aligned.

For masses ranging between 50 and 500 *m/z*, 618 nonpeptide small molecules and 2926 peptides were extracted from the HMDB and the metabolome.jp database. After eliminating metabolites present in the standard library and nonpositively charged compounds (net charge < 0.25 at pH = 1.8), 80 nonpeptide molecules and 2858 peptides remained. Based on the annotation using *m/z* and normalized t_m values, a total of 179 metabolites in our standard library were assigned to the 460 peaks in the urine sample with an average difference between biological samples and the standard *m/z* of 4.63 ± 4.72 ppm and 0.19 ± 0.23 min, respectively. Using the difference tolerance of 15 ppm in *m/z* and 1.5 min in normalized t_m, 13 metabolites in HMDB and 48 peptides, not present in our standards (Supporting Information Table S4), were further assigned to the remaining 281 peaks (5.31 ± 5.24 ppm and 0.79 ± 0.47 min, respectively).

4 Discussion

To facilitate metabolite identification, we developed a mathematical model to predict the t_m. The principal factors affecting the t_m prediction are the quality of time normalization of the

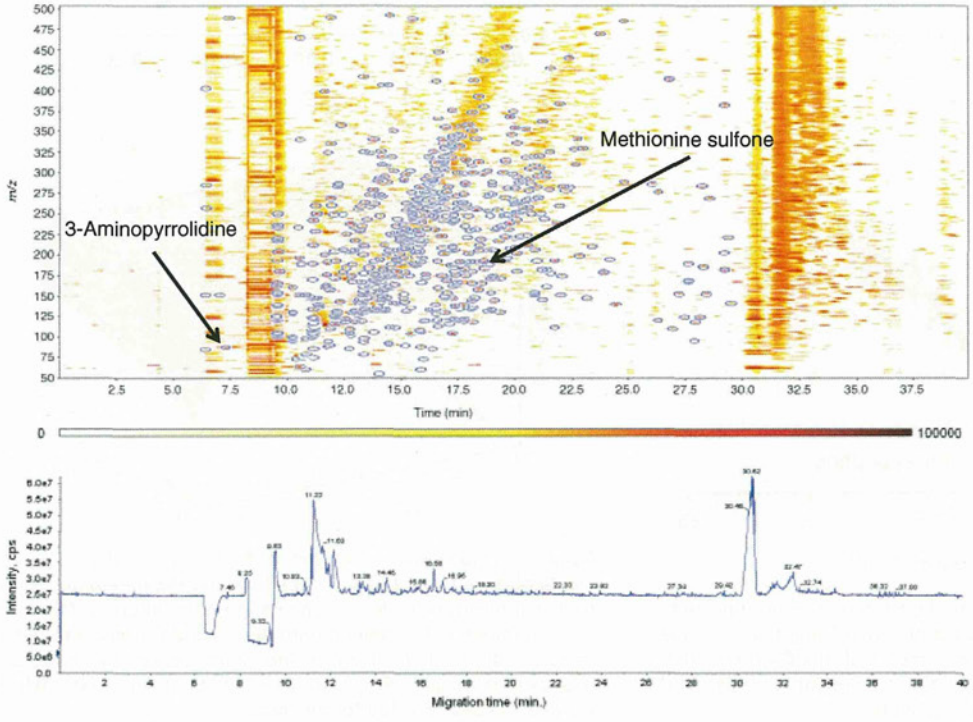


Figure 4. 2-D map (migration time and *m/z*) showing the location of the internal standards, methionine sulfone and 3-aminopyrrolidine and a total ion electropherogram obtained from a human urine sample. The orange and white colors in the 2-D map show the high and low intensities, respectively.

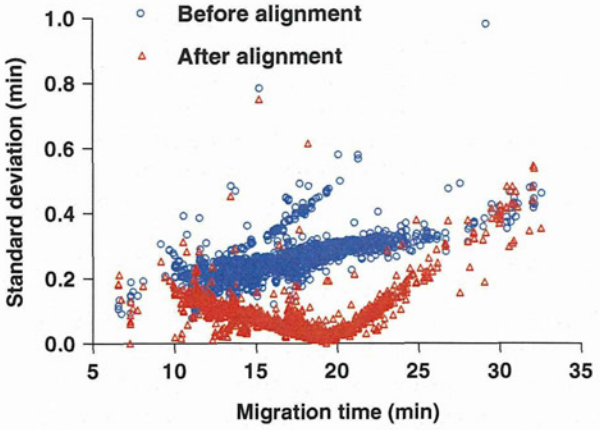


Figure 5. The t_m variations for matched peaks in the standard solution and five urine samples before and after time normalization. The *X* and *Y*-axes indicate the normalized t_m and the SD of the matched peaks, respectively. The average SD for t_m before and after alignment was 0.260 and 0.00866 min, respectively.

run-to-run time shift and errors in calculating molecular descriptors. As shown in Fig. 4, the peaks whose t_m was far from the internal standards were normalized less accurately. This may degrade the prediction accuracy for late t_m, as shown in Fig. 2. One possible simple solution to compensate for this was to append more internal standards for use in time normalization. The t_m, *i.e.* the peak top time, was used and fluctuated according to various peak shapes attributable to stacking conditions when electric power was applied. To reduce the noise derived from the peak top fluctuation, we used lower sheath liquid flow to reduce the skewness of the peak shapes. Regarding the molecular descriptors, the net charge was the most important factor and was commonly ranked first and selected in all feature selection methods. Therefore, the accuracy of the pK_a values and the net charge was essential for the quality of the t_m prediction. In particular, accurate pK_a values are required for molecules yielding a pK_a close to pH 1.8 because their net charge is more sensitive to these values.

To provide versatile uses for the predictive model, we applied feature selection and CV, techniques that are commonly used to quantitatively model the structure–retention relationship in GC and LC [34]. However, there is a risk of overfitting; thus, the established model consists of parameters optimistically adjusted only for the given data sets, which does not guarantee the prediction accuracy of the external molecules [35]. Further improvements can be achieved by: (i) the structural and functional diversity of the metabolites used for training data sets should be sustained; (ii) rigorous validation, *e.g.* separating data sets into three groups and training, validation and the test examination is conducted to prevent local parameter adjustment [36] and (iii) a group of predictors is alternatively used to enhance the robustness against noise [19]. The model presented in this manuscript is only the normalized t_m and could not eliminate run-to-run systematic t_m shifts. Thus, t_m prediction accuracy strongly depends on the success of the normalization procedure.

In comparison with a mathematical model that uses the net charge and molecular mass (*e.g.* Offord's model), the method presented here is limited because of the difficulty in interpretation of the developed model. On the other hand our model is advantageous because it can differentiate the migration behavior of differently ordered amino acids. Among the nine variables selected by RFAE, only two, DASA and DCASA, descriptors for the charge status of the surface area, showed different values that contributed to the differentiation of the t_m difference between peptides with the same composition. However, accurate t_m prediction is required for their differentiation because the average difference in t_m of these differently ordered peptides was small (0.19 min) compared with the average difference between the predicted and measured t_m in the complete data set (0.61 min). Of the standard peptides examined, 54 (27 peptide sets) potentially showed different orders, *e.g.* AP and PA, and the order in the electropherogram axis of only 13 and 11 peptide sets was correct based on the normalized t_m predicted with the complete data sets and the CV data sets, respectively. Therefore, for more accurate prediction, there is still potential to optimize the measurement conditions, such as buffer pH, which affects the resolution of the peaks and the nebulizer flow rate, which may affect the peak shapes, in addition to improving the mathematical prediction techniques.

For biological sample analysis, the composition formula of the assigned metabolites based on the *m/z* and predicted t_m was compared with formulae based on the *m/z* and isotope distribution alone. Only 66% of the assigned metabolites were consistent in both methods. Using fivefold concentrated urinary sample data, we tested the confidence of the composition formulae calculated from the *m/z* and isotope distribution using the metabolites in the standard library. As shown in Supporting Information Table S5, while the composition formulae at low *m/z* were correctly estimated, those for larger metabolites were not. For example, an incorrect candidate was more highly ranked for adenosine (*m/z* 268.104), and no candidates were listed for ophthalmate (290.135 *m/z*) and reduced glutathione (*m/z* 308.091). This may be due to the existence of a larger number of candidates when we consider larger *m/z* values with a certain error tolerance and sometimes unclear small isotope peaks. Thus, the accuracy of the estimated composition formula based on the isotope distribution in TOFMS data alone cannot satisfy the confidence criteria, as cited in [9] and it is not enough to identify metabolites. Therefore, integrating several types of data, including the predicted t_m, is useful for metabolite annotations.

Although thousands of molecules were used for annotation, the annotation coverage, assuming this estimation is correct, was merely raised from 38.9 to 52.2%. Approximately, half of the peaks that might be derived from charged molecules remain unknown. The addition of known exogenously derived metabolites to the model may be beneficial because urine may include several exogenous molecules and metabolites derived from tobacco, food, beverages and toxicological factors, for example. Many of the remaining

features may also have an m/z >500 or originate from molecules that are not in the training data set (HMDB). Although we only focused on peaks under 500 m/z in this study, the identification of larger molecules requires greater accuracy because more metabolites become candidates. In addition, longer peptides may show different trends in terms of electrophoretic mobility compared with shorter peptides because of the variable secondary structure of the peptides [10]. The model is currently limited to the prediction of low-molecular-weight metabolites, but further analyses with longer peptides may improve the versatility of this method. Methods for predicting t_m, as described here, can facilitate metabolite identification or significantly reduce the number of candidate metabolites for a given m/z and isotope pattern. This approach results in reasonably confident metabolite identification without requiring additional systematic and time-consuming MS/MS or NMR analyses of detected features. Therefore, our approach can considerably enhance the value of non-targeted CE-MS-based metabolomic analysis.

This study was supported by a grant from the Global COE Program entitled, "Human Metabolomic Systems Biology" and by a Grant-in-Aid for Scientific Research on Priority Areas "Systems Genomes" and on "Lifesurveyor" from the Ministry of Education, Culture, Sports, Science and Technology of Japan as well as research funds from the Yamagata prefectural government and the City of Tsuruoka.

The authors have declared no conflict of interest.

5 References

[1] Soga, T., Baran, R., Suematsu, M., Ueno, Y., Ikeda, S., Sakurakawa, T., Kakazu, Y., Ishikawa, T., Robert, M., Nishioka, T., Tomita, M., *J. Biol. Chem.* 2006, *281*, 16768–16776.

[2] Soga, T., Ohashi, Y., Ueno, Y., Naraoka, H., Tomita, M., Nishioka, T., *J. Proteome Res.* 2003, *2*, 488–494.

[3] Hu, A., Chen, C. T., Tsai, P. J., Ho, Y. P., *Anal. Chem.* 2006, *78*, 5124–5133.

[4] Bailon-Perez, M. I., Garcia-Campana, A. M., del Olmo Iruela, M., Cruces-Blanco, C., Gamiz Gracia, L., *Electrophoresis* 2009, *30*, 1708–1717.

[5] Cho, S. H., Lee, J., Choi, M. H., Lee, W. Y., Chung, B. C., *Biomed. Chromatogr.* 2009, *23*, 426–433.

[6] Zurbig, P., Renfrow, M. B., Schiffer, E., Novak, J., Walden, M., Wittke, S., Just, I., Pelzing, M., Neususs, C., Theodorescu, D., Root, K. E., Ross, M. M., Mischak, H., *Electrophoresis* 2006, *27*, 2111–2125.

[7] Monton, M. R., Soga, T., *J. Chromatogr. A* 2007, *1168*, 237–246; discussion 236.

[8] Mungur, R., Glass, D. M., Goodenow, D. B., Lightfoot, D. A., *J. Biomed. Biotechnol.* 2005, *2*, 198–214.

[9] Kind, T., Fiehn, O., *BMC Bioinformatics* 2007, *8*, 105.

[10] Tessier, B., Blanchard, F., Vanderesse, R., Harscoat, C., Marc, I., *J. Chromatogr. A* 2004, *1024*, 255–266.

[11] Jalali-Heravi, M., Shen, Y., Hassanisadi, M., Khaledi, M. G., *J. Chromatogr. A* 2005, *1096*, 58–68.

[12] Benavente, F., Balaguer, E., Barbosa, J., Sanz-Nebot, V., *J. Chromatogr. A* 2006, *1117*, 94–102.

[13] Jalali-Heravi, M., Shen, Y., Hassanisadi, M., Khaledi, M. G., *Electrophoresis* 2005, *26*, 1874–1885.

[14] Liu, K. P., Xia, B. B., Zhang, X. Y., *J. Liq. Chromatogr. Relat. Technol.* 2008, *31*, 11–12.

[15] Xue, C. X., Zhang, R. S., Liu, M. C., Hu, Z. D., Fan, B. T., *J. Chem. Inf. Comput. Sci.* 2004, *44*, 950–957.

[16] Ma, W., Luan, F., Zhang, H., Zhang, X., Liu, M., Hu, Z., Fan, B., *Analyst* 2006, *131*, 1254–1260.

[17] Yu, K., Cheng, Y., *Talanta* 2007, *71*, 676–682.

[18] Lee, R., Ptolemy, A. S., Niewczas, L., Britz-McKibbin, P., *Anal. Chem.* 2007, *79*, 403–415.

[19] Sugimoto, M., Kikuchi, S., Arita, M., Soga, T., Nishioka, T., Tomita, M., *Anal. Chem.* 2005, *77*, 78–84.

[20] Horakova, J., Petr, J., Maier, V., Tesarova, E., Veis, L., Armstrong, D. W., Gas, B., Sevcik, J., *Electrophoresis* 2007, *28*, 1540–1547.

[21] Hirayama, A., Kami, K., Sugimoto, M., Sugawara, M., Toki, N., Onozuka, H., Kinoshita, T., Saito, N., Ochiai, A., Tomita, M., Esumi, H., Soga, T., *Cancer Res.* 2009, *69*, 4918–4925.

[22] Sugimoto, M., Wong, T. D., Hirayama, A., Soga, T., Tomita, M., *Metabolomics* 2009, DOI: 10.1007/s11306-11009-10178-y.

[23] Brown, M., Dunn, W. B., Dobson, P., Patel, Y., Winder, C. L., Francis-McIntyre, S., Begley, P., Carroll, K., Broadhurst, D., Tseng, A., Swainston, N., Spasic, I., Goodacre, R., Kell, D. B., *Analyst* 2009, *134*, 1322–1332.

[24] Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S. *et al. Nucleic Acids Res.* 2007, *35*, D521–D526.

[25] Adamson, N. J., Reynolds, E. C., *J. Chromatogr. B Biomed. Sci. Appl.* 1997, *699*, 133–147.

[26] Reijenga, J. C., Martens, J. H., Giuliani, A., Chiari, M., *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* 2002, *770*, 45–51.

[27] Skoog, B., Wichman, A., *Trends Anal. Chem.* 1986, *5*, 82–83.

[28] Witten, I. H., Frank, E., *Data Mining : Practical Machine Learning Tools and Techniques*, Morgan Kaufman, Amsterdam; Boston, MA 2005.

[29] Chiu, S. H., Chen, C. C., Lin, T. H., *Artif. Intell. Med.* 2008, *44*, 221–231.

[30] Vapnik, V. N., *The Nature of Statistical Learning Theory*, Springer, Berlin 1995.

[31] Dudek, A. Z., Arodz, T., Galvez, J., *Comb. Chem. High Throughput Screen.* 2006, *9*, 213–228.

[32] Smola, A. J., Borer, S., *Stat. Comput.* 2004, *14*, 199–222.

[33] Jing, J. H., Liang, G. Z., Mei, H., Xiao, S. Y., Xia, Z. N., Li, Z. L., Mol. Simul. 2009, *35*, 263–269.

[34] Heberger, K., *J. Chromatogr. A* 2007, *1158*, 273–305.

[35] Hawkins, D. M., *J. Chem. Inf. Comput. Sci.* 2004, *44*, 1–12.

[36] Shinoda, K., Sugimoto, M., Yachie, N., Sugiyama, N., Masuda, T., Robert, M., Soga, T., Tomita, M., *J. Proteome Res.* 2006, *5*, 3312–3317.

Metabolomics (2010) 6:27–41
DOI 10.1007/s11306-009-0175-1

ORIGINAL ARTICLE

Differential metabolomics software for capillary electrophoresis-mass spectrometry data analysis

Masahiro Sugimoto · Akiyoshi Hirayama · Takamasa Ishikawa · Martin Robert · Richard Baran · Keizo Uehara · Katsuya Kawai · Tomoyoshi Soga · Masaru Tomita

Received: 9 March 2009 / Accepted: 24 July 2009 / Published online: 26 September 2009

Abstract In metabolomics, the rapid identification of quantitative differences between multiple biological samples remains a major challenge. While capillary electrophoresis–mass spectrometry (CE–MS) is a powerful tool to simultaneously quantify charged metabolites, reliable and easy-to-use software that is well suited to analyze CE–MS metabolic profiles is still lacking. Optimized software tools for CE–MS are needed because of the sometimes large variation in migration time between runs and the wider variety of peak shapes in CE–MS data compared with LC–MS or GC–MS. Therefore, we implemented a standalone application named *JDAMP* (Java application for Differential Analysis of Metabolite Profiles), which allows users to identify the metabolites that vary between two groups. The main features include fast calculation modules and a file converter using an original compact file format, baseline subtraction, dataset normalization and alignment, visualization on 2D plots (*m/z* and time axis) with matching metabolite standards, and the detection of significant differences between metabolite profiles. Moreover, it features an easy-to-use graphical user interface that requires only a few mouse-actions to complete the analysis. The interface also enables the analyst to evaluate the semiautomatic processes and interactively tune options and parameters depending on the input datasets. The confirmation of findings is available as a list of overlaid electropherograms, which is ranked using a novel difference-evaluation function that accounts for peak size and distortion as well as statistical criteria for accurate difference-detection. Overall, the *JDAMP* software complements other metabolomics data processing tools and permits easy and rapid detection of significant differences between multiple complex CE–MS profiles.

Keywords Capillary electrophoresis–mass spectrometry · Metabolome · Data analysis · Software

Electronic supplementary material The online version of this article (doi:10.1007/s11306-009-0175-1) contains supplementary material, which is available to authorized users.

M. Sugimoto (✉) · A. Hirayama · M. Robert · R. Baran · K. Kawai · T. Soga · M. Tomita
Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017, Japan
e-mail: msugi@sfc.keio.ac.jp

M. Sugimoto · K. Uehara · K. Kawai
Department of Bioinformatics, Mitsubishi Space Software Co. Ltd, Amagasaki, Hyogo 661-0001, Japan

T. Ishikawa · T. Soga · M. Tomita
Human Metabolome Technologies Inc, Tsuruoka, Yamagata 997-0052, Japan

R. Baran
Life Sciences Division, MS: 84R0171, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

1 Introduction

The objective of metabolomics is to quantitatively analyze complete profiles of small molecules in biological samples, one of the most challenging tasks in systems biology (Nicholson and Wilson 2003). Most experiments involve the unbiased identification of biologically meaningful signal differences in the levels of a small number of metabolites, within a multitude of signals. In addition, biomarker discovery and the detection and association of significant sample differences and patterns that identify specific biological conditions are major tasks in metabolome analysis. Analytical platforms commonly used to collect metabolite

profiles include nuclear magnetic resonance (NMR) (Reo 2002), as well as gas chromatography (GC) (Fiehn et al. 2000), liquid chromatography (LC) (Plumb et al. 2003), and capillary electrophoresis (CE) combined with mass spectrometry (MS) (Soga et al. 2003). Typically, the data analysis workflow, starting with raw data, includes filtering or baseline correction, peak detection, alignment of peaks across multiple datasets, generation of a processed data matrix, and statistical analysis such as principal component analysis and partial least squares discriminant analysis to identify significant differences between datasets (Katajamaa and Oresic 2007). Although software packages for automatic processing are available, most of the existing tools were developed or optimized for NMR (Wang et al. 2009; Zhao et al. 2006), LC–MS, and GC–MS (Bellew et al. 2006; Bunk et al. 2006; Fischer et al. 2006; Katajamaa et al. 2006; Katajamaa and Oresic 2005; Smith et al. 2006; Styczynski et al. 2007; Tautenhahn et al. 2008), or for MS alone (Broeckling et al. 2006; Haimi et al. 2006; Karpievitch et al. 2007; Wong et al. 2005). There are currently relatively few tools optimized for CE–MS data analysis (Wittke et al. 2003).

CE–MS is a versatile system, which is well suited for metabolome studies that require high-resolution separation of metabolites and high-detection sensitivity for the analysis of numerous charged and low molecular weight molecules. CE allows for temporal separation of components based on their charge and size and, using MS, most compounds that co-migrate in CE can be resolved (Monton and Soga 2007). However, a major challenge in CE–MS is the variability in migration time. This run-to-run variability in electro-osmotic flow (EOF) is mainly due to changes in the capillary wall or electrolyte solution induced by the sample matrix that results in greater migration time variation compared with other separation methods such as GC or LC. On the other hand, even in a single run, fluctuations of capillary electric condition and run-to-run variability also cause migration time shifts. Although good reproducibility in electrophoretic mobilities was reported for amino acids in CE–MS (Lee et al. 2007), accurate and versatile migration time correction applicable to a large variety of metabolites is necessary. With regard to migration time, once it has been corrected, the actual electrophoretic mobility of molecules in CE can be highly reproducible. In addition, the peak shapes in CE–MS show more diversity and differences compared with those derived from chromatographic techniques such as LC–MS and GC–MS, making the peak detection problem particularly challenging. Thus, software that implements robust migration time alignments and efficient feature analyses is needed for CE–MS data processing. To address these issues, we previously developed MathDAMP, a collection of tools running as a Mathematica package (Baran et al. 2006; Baran et al. 2007). MathDAMP was instrumental in the discovery of metabolite biomarkers (Soga et al. 2006) and for elucidating enzyme and gene functions (Saito et al. 2006; Yoshida et al. 2007). However, the use of complex scripts with large datasets in a generic mathematical environment involves large computation overhead costs and, consequently, a relatively limited throughput. Specifically, the alignment procedures in the electrophoretic dimension are sensitive to measurement quality and require iterative quality control steps and manual optimization of multiple parameters to avoid incomplete alignment due to outlier peaks or large migration time-shifts between datasets. In addition, the datapoint-by-datapoint method for difference detection, as implemented in MathDAMP, which detects significant differences among groups without peak-selection, can yield a number of false-positive results.

The objective of this project was to develop a user-friendly and high-performance platform suitable for differential analysis of CE–MS metabolite profiles that is complementary to existing tools. Therefore, we developed *JDAMP* (Java application for Differential Analysis of Metabolome Profiles), which offers a graphical user interface (GUI) and is designed to facilitate iterative analyses with graphical confirmation of findings. It also uses a specific file converter that allows direct conversion of standard data formats such as NetCDF and CSV (text) file, or Agilent-specific CE-TOFMS raw data to the *JDAMP* original file format. The possibility of directly using Agilent-specific CE-TOFMS raw data has the added benefit of avoiding the large size of intermediate standard file formats based on text or XML. In addition, the newly developed difference detection algorithm allowed for a reduced number of false-positive peaks, which can accelerate discovery-oriented applications of CE–MS-based metabolomics.

2 Materials and methods

2.1 File conversion

The first step in the data processing workflow is file conversion from either standard or vendor-specific raw data files to the *JDAMP* input file. Because a large number of samples are usually analyzed simultaneously to identify statistically reliable differences, the huge file size of conventional standard file formats such as netCDF or mzXML (Hardy and Taylor 2007; Pedrioli et al. 2004) can constitute a significant barrier to large-scale and high-throughput analyses in terms of performance and data storage. Therefore, we implemented a separate program, *dotMZ*, to convert *wiff* data files generated by Analyst QS for Agilent TOF software (Applied Biosystems, CA, USA; MDS SCIEX, ON, Canada) and binary data files (called dot D

dataset) generated by the MassHunter software (Agilent, Santa Clara, CA), which controls the latest versions of Agilent TOF mass spectrometers. To support other vendor platforms as well as non-CE–MS data formats, ASCII-based comma-separated values (CSV) files formatted as generated by Analyst QS and MassHunter, Tab delimited files formatted as generated by MassLynx software (Waters Corporation, Milford, MA), and NetCDF format files that are generated from most types of instruments can also be used as the input. NetCDF, CSV and Tab delimited data files can be converted by *dotMZ* to a specifically designed binary file format (named *ciff* files). Using the application programming interfaces (APIs) of Analyst QS or MassHunter, the Agilent-supplied binary files (*wiff* file or D dataset) can also be directly converted to *ciff* files. In this case, because some Analyst QS or MassHunter libraries are required during conversion, the converter must be installed on a system hosting the Analyst QS or MassHunter software, which is usually provided to owners of Agilent TOF systems.

2.2 Data processing and analysis

The analytical workflow includes data preprocessing, normalization of time-shift (alignment) and signal intensities, and difference-detection, all of which are commonly used feature-detection steps in metabolomics processing of LC–MS and GC–MS data (reviewed in Katajamaa and Oresic 2007). The strategy for data analysis in *JDAMP* is shown in Fig. 1. Overall, it corresponds with the workflow of MathDAMP and its basic algorithm (Baran et al. 2006). Briefly, in the preprocessing step, raw datasets undergo primary binning along the *m/z* dimension to fine resolution (default 0.02 *m/z*) while subtracting the baseline from each electropherogram by polynomial curve-fitting using a nonlinear regression method (Ruckstuhl et al. 2001) and by fixing signals under a specified threshold to 0. Noise values are calculated from signals between 2 and 3 min, where metabolite signals are not usually found. Values obtained in the first minute are not usually used because of unstable signals. The resulting datasets are then further binned to 1 *m/z* unit resolution along the *m/z* axis (secondary binning). Directly binning electropherograms into 1 *m/z* units without primary narrow binning and background-subtraction and noise reduction will result in a low signal/noise ratio for small (narrow peaks (in *m/z* axis)) peaks. Therefore, a primary narrow binning step is preferred to facilitate and maintain the detection of these peaks for subsequent procedures. For the secondary binning process, MathDAMP used n $\pm$ 0.5 *m/z* (n; integer) as edges of binning electropherograms. By contrast, *JDAMP* uses n $-$ 0.3 to n $+$ 0.7 *m/z* to limit the possibility of separating isotopic peaks derived from a divalent peak into two different bins. Subsequently, migration time correction (optimized for CE–MS-specific variation) is performed by a dynamic time-warping method (Bylund et al. 2002). This step (1) executes peak selection using the Douglas-Peucker algorithm (Wallace et al. 2004) for each electropherogram (peaks detected at this step are called representative peaks), (2) matches the peaks across datasets by dynamic programming (DP), (3) changes the parameters of the time-normalization function with the optimization method, and (4) returns to (2) until the improvement of the objective function is reduced to a specific limit value. The score produced by DP is used to evaluate the two numerical parameters, α and γ, of the normalization Eq. 1 derived for CE migration (Reijenga et al. 2002), as previously described (Baran et al. 2006).

$$t_R = \frac{1}{(1/\alpha t) - (\gamma/2)}, \tag{1}$$

where t_R and t are the normalized and original migration times, respectively. Briefly, to enhance the robustness of the alignment, the optimization loop steps from (2) to (4) are performed twice using different gap penalties; a larger gap penalty is used to generate a primary normalization function for rough alignment and a smaller gap penalty is used for secondary fine-tuning of the function. The resulting function is then used to rescale the migration times of each dataset, thus eliminating the time shifts for each run. Signal intensities are adjusted to compensate for the compression or expansion of peaks during the normalization and thus conserve the same peak areas, as previously implemented in MathDAMP (Baran et al. 2006). Finally, differences are detected from complete, aligned datasets on a datapoint-by-datapoint basis using a novel difference-detection function that was not implemented in MathDAMP. The results are visualized as numerical values or statistical scores on overlaid electropherograms and 2D maps.

Except for the difference-detection phase, all of the steps include parameters that can be tuned by the user based on the input datasets. This is an important step that can involve considerable user time and input. Therefore, the GUI was designed to facilitate quality control and optimization of iterative parameters by the user. The GUI is implemented in Java language. The GUI is easy to use and allows interactive data processing with visualization. On the other hand, the calculation engines are written in C++ for rapid performance. Each process was implemented as a separate program to benefit programmers who want to write scripts to create directly executable files for routine analyses.

A datapoint-by-datapoint approach was originally implemented in MathDAMP to highlight differences between multiple datasets. This approach enables the

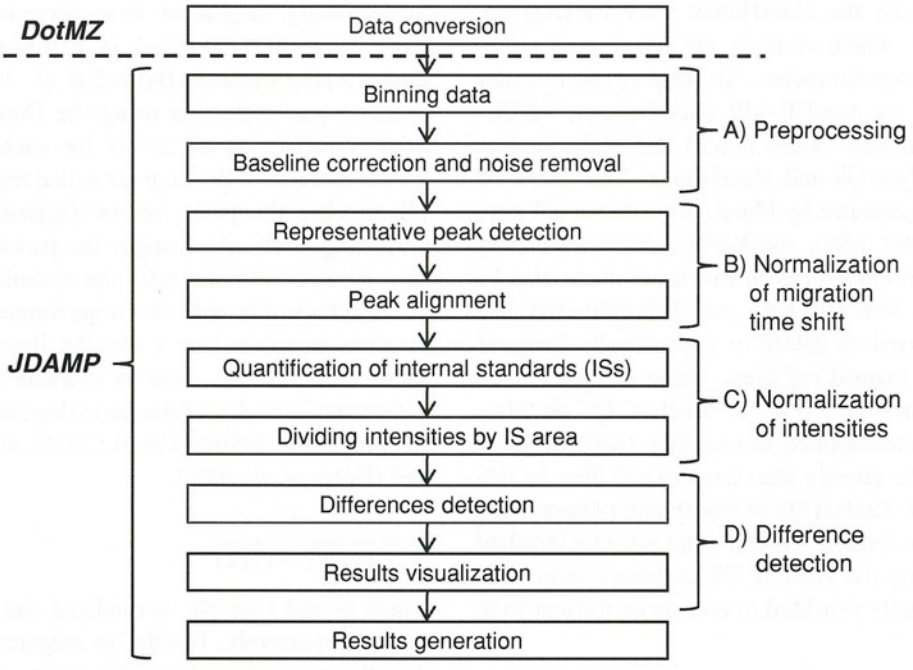


Fig. 1 Schematic representation of the analytical workflow performed by the *dotMZ* converter and *JDAMP*. In preprocessing (A), binning datasets, baseline correction for eliminating background drift and noise removal to delete small-intensity signals below a user specified S/N are performed. In the migration-time normalization procedure (B), representative peak detection, peak matching across datasets with dynamic programming, and correction of migration times are conducted. In the normalization of intensities step (C), internal standards selected by the user are used to normalize the intensities in the entire datasets. For users who do not use internal standards, this process can be omitted. In the difference detection step (D), significant differences are detected depending on multiple criteria and are visualized as overlaid electropherograms with 2D plots

identification of differences while avoiding the limitations of peak-selection for CE–MS electropherograms and the common resulting problem of missing values. However, this method yields a number of false-positives, e.g., a data point at the edge of a peak exhibiting a significant difference is recursively selected as a different result. In addition, the noise-related regions of the electropherograms are sometimes highlighted for reasons such as incomplete background-correction and noise removal. To eliminate such false-positive results, we defined an additional intuitively interpretable, simple evaluation function E;

$$E = \frac{\sum_{t\in\emptyset} I_t}{\mathrm{E}_{t\in\varepsilon}|I_t - G_t|} \times \frac{AR}{AR_{\max}} \times \frac{T}{T_{\max}}, \quad (2)$$

where AR and T represent the intensity differences that are significant in both absolute and relative terms (absolute × relative difference, named ABSRel) (Baran et al. 2006) and the t-score of intensities at the selected time point, respectively. AR_{max} and T_{max} are the maximum ABSRel and t-score values in the dataset, respectively. I_t is the signal intensity for the actual datapoint and G_t is the height of the Gaussian curve at time-point t. Because Φ and ε are the peak area and the Gaussian area along the time axis, respectively, the numerator and denominator of the first term become the peak area and the degree of distortion from the Gaussian curve. First, to determine the peak area, the electropherograms in a group in which the average intensities of the points of interest are larger than that of the others, are averaged. Second, both the leading and trailing peak edges are identified by moving away, in both directions, from the local maximal intensities. The peak edges are assigned to the first datapoints that are below the threshold (5% of the local maximal intensity). Third, a Gaussian curve is fitted to the peak shape using the simplex method and the differences between the curve and the peak are summed. Then, the datapoint-by-datapoint detection score, using function E, is used to increase the weight of the contribution of datapoints located in regions with larger and more statistically significant differences, and with better Gaussian peak shapes.

The performance of peak-selection based on the difference-detection function using the Douglas-Peucker algorithm (Wallace et al. 2004) was compared with the datapoint-by-datapoint method with and without evaluation using function E.

2.3 Test data

To test the utility of the software to detect differential features in complex datasets, we processed data collected by CE–MS analysis of mixtures of standards in which a few metabolites were spiked at different levels. The