Large-scale analysis of post-translational modifications in *E. coli* under glucose-limiting conditions over 2 weeks

Viswanadham Sridhara[1], Colin Brown[2], Daniel R. Boutz[2,3], Maria Person[4], Jeffrey E. Barrick[1,2,3,5], Edward M. Marcotte[1,2,3,5], and Claus O. Wilke[1,2,3,6]

January 18, 2015

[1]Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, TX, USA
[2]Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, TX, USA
[3]Center for Systems and Synthetic Biology, The University of Texas at Austin, Austin, TX, USA
[4]College of Pharmacy, The University of Texas at Austin, Austin, TX, USA
[5]Department of Molecular Biosciences, The University of Texas at Austin, Austin, TX, USA
[6]Department of Integrative Biology, The University of Texas at Austin, Austin, TX, USA

Abstract

How do the post-translational modifications [PTMs] change over time during microbial growth under nutrient limiting conditions? We sought to answer this question using *E. coli* grown under glucose starvation conditions. We generated mass-spectrometry based proteomics data at 9 different time points ranging from exponential to long stationary phases i.e., upto 2 weeks. We then ran MODa on this data for an unrestricted search of PTMs. There were several interesting observations in this MODa analysis. First, we show that the amount of modified protein seems to be constant, occuring approx. 30%, at all the 9 time points. Second, we show that acetylations, carboxylations and phosphorylations increase, while nitrosylations seem to be constant from exponential to stationary phases.Third, we show that sulfoxide reductases fix MetSO to Met during stationary phase, protecting *E. coli* from oxidative damage. Finally, we found some novel post-translational modifications in this data, a frequent one being phospho-gluconylation on serine in R6 ribosomal protein.

1 Introduction

Mass-spectrometry based proteomics [16] offers a unique way to characterize different post-translational modifications [PTMs] associated with proteins. PTMs are key in determining protein function, localization and regulation. Hundreds of PTMs are known *in vivo*, however most of the current

studies focus on identifying only few PTMs from mass-spec data. This is both because of computational limitations as well as the inefficiency of enrichment techniques to characterize or identify all PTMs. However with the improvement of search algorithms for an unrestricted search of PTMs, identifying multiple PTMs in *tandem* has become a possibility. In this work, we used MODa [20], a naive based spectral alignment algorithm to identify different kinds of PTMs in *E. coli* REL606 strain.

Mass-spectrometry data is generated for *E. coli* grown under glucose limiting conditions. *E. coli* grown under these nutrient limiting conditions generally behave differently at different phases of their growth i.e., during exponential and stationary phases. During the stationary phase, stress response proteins act because of the conditions casued by limiting nutrients, change in pH value, accumulation of toxic substances in the flasks etc. In an earlier study under similar conditions [29], the number of phosphorylation sites as well as their occupancy levels seem to increase during stationary phase as a response to stress. In the same study, protein abundance of SspA seemed to be constant, while the phosphorylation of the same protein increased suggesting PTM level quantitation as key compared to protein quantitation. In another study, in acinetobacter baumannii by the same group [28], the oxidative stress has been shown to be compensated by the E. coli machinery at the later stages of growth i.e., stationary phase compared to the early exponential phase. This itself suggests that PTM level analysis is necessary to understand the response of organisms to environmental and other disturbances.

Few other studies of PTMs focussed on multiple PTMs, but they looked at only one "snapshot" of interest i.e., a single time point during the growth. Since it has already been shown that during different phases of growth, different proteins will be expressed, a "snapshot" of PTM analysis does not provide a clearer picture on the significance or the function associated by these PTMs. Our current work of using multiple time points during the growth curve or looking at multiple "snapshots" along with an unrestricted search of PTMs provides a holistic view of proteome function.

2 Results

2.1 Running MODa on *E. coli* proteome

Post-translational modifications, in most cases, determine the specific function of the protein. Identifying all the PTM combinations or the set of PTMs associated with all the proteins (i.e., entire proteome) is limited by experimental enrichment techniques for PTMs as well as computational limitations. The current analysis of *E. coli* dataset is not enriched for any PTMs. This means that this dataset is suitable for hunting for different kinds of PTMs, instead of focussing on particular PTMs. Most of the current studies focus on a single PTM coverage, such as generally done in phosphoproteomics studies, glycosylation studies or acetylation studies, where the sample is enriched for a particular PTM.

We then used MODa, a sequence search algorithm to search for peptides and PTMs associated with it in this mass-spectrometry dataset. MODa is a naive based search algorithm that uses an unrestricted approach to find the post-translational modifications in the data. We have mass-spectrometry data at 9 different points of the OD600 curve. So, we have 9 sets of peptide lists, that represent proteins+PTMs from early exponential phase (3,4,5,6 hours of growth) to long-stationary phases (8,24,48, 168 and 336 hours) of *E. coli* growth. OD600 at these 9 time points is given in

supplementary information of this article (see Suppl Figure S1).

The parameters used with MODa is described in detail in the methods section of the paper. In brief, MODa uses multiple short sequence tag information in the spectra along with a dynamic programming approach to identify the peptide hits along with their protein identifiers. The use of multiple tags reduces both the database size and the false positives. Once we have the MODa search results, we asked the following questions: (a) How much of the proteome is modified? (b) How do these modifications change over time? (c) Is there any biological relevance associated with the temporal change of a particular PTM? (c) Can we find any novel modifications in this *E. coli* data?

2.2 Modified *E. coli* proteome

First, we are interested in identifying the amount of proteome modified. Since we have data at each of the 9 time points i.e., from exponential to long-stationary phases, we plotted the total frequency of all PTMs at each of these time points. For this, we simply calculated the percentage of the total number of modified peptide spectral matches (PSMs) at each time point. We used 1 standard error to plot the variation within 3 biological replicates.

Figure 1 shows the total percent of the peptide-spectral matches on the y-axis and the time from OD600 plot on the x-axis. It is clear that the amount of spectra that is modified stays approx. at 30% during all the phases of the growth. ***Add total number of PSMs probably at 3 hours, 24 hours and 336 hours to have an idea***. This itself is an interesting result. The 30% of protein getting modified is in agreement with some of the previous studies ***(Cite papers which show the same numbers)***. However these previous studies looked at a "snapshot" of the proteome, instead of looking at different phases during the growth. We next looked at the distribution of the mass-shifts or PTMs at each time point of the growth curve i.e., what PTMs are present at a given time point and how does it change over time?

2.3 Mapping mass-shifts to post-translational modifications

MODa outputs mass-shifts in integer masses in the peptide hits. A mass-shift indicates either a PTM or a substitution compared to the reference database. The mass-shifts are then converted to PTMs using UNIMOD database, as described in methods section. For example, we know that +16 Dalton shift on any amino acid is most likely the oxidation. When we looked at the UNIMOD PTM list, we did see that this is oxidation, generally shown to occur on few amino acids such as cysteine, methionine etc. We did not limit our analysis to frequently occuring mass-shifts, but also focussed on well known PTMs, and novel mass-shifts that were not characterized earlier. However, we did not investigate into mutations and this will be a topic we will look in the future as this is a possibility with MODa output.

Even though the total amount of modified proteome is mostly constant 30%, the content of PTMs at each time point vary during different phases of growth i.e., a particular modification can either stay constant, decrease or increase from expoential to long-stationary phases. To look at this in detail, we first plotted the frequency of different mass-shifts from the MODa output at 2 different time points. Figure 2 shows this mass-shift frequency at 3 hrs and 24 hrs. Figure 2 (A) shows the mass-shift frequency at 3 hours, while (B) shows the frequency at 24 hours. We did not plot the error bars as this makes the figure a bit clumsy, with these bars appearing at each bin. The height

of the bins is the percentage of the mass-shifts in the entire peptide-spectral match list. Since we used 3 biological replicates, the actual % is the mean of the 3 replicates. It is clear that +1 Da peak is the most frequent one. However this +1 Da modification seems to be 13C peak-picking as it seems to be randomly occuring on all the amino acids, as shown in the profile (Suppl Figure S5). At both 3 and 24 hours, next frequent modification seems to be the bin at +16 Dalton. 16 Dalton (shown as +16 Da in figure) indicates oxidation. 2 frequently oxidized amino acids are the cysteine and methionine. However, the sample is treated in a way that caps cysteines and hence we don't expect to see Cys oxidized. A look at profile (Suppl Figure S6) shows majority of oxidations on methionine which is expected. Next thing, we analyzed are the temporal changes of few interesting PTMs, in acetylation, oxidation, phosphorylation, carboxylation and nitrosylation.

2.4 N-term protein acetylations are frequent in *E. coli*

Acetylation in *E. coli* has been studied well in the past [3] [12]. Acetylation occurs in 2 forms i.e., N-alpha acetylation and N-sigma acetylation. It is known that acetylation occurs co-translationally on n-term (N-alpha) and is not reversible, while post-translationally it occurs as a reversible modification, mostly on lysine (cite this paper) *The diversity of lysine-acetylated proteins in Escherichia coli.*. In the current work, we focus on n-terminal acetylation. It has been known that almost 2/3rds of the proteins are substrates for n-terminal methionine excision. Following this methionine excision, n-terminal processing, especially n-terminal acetylation [7] has been shown in the past to be one of the most frequently occuring PTMs. Previous data suggests that this modification is more frequent in eukaryotes than bacteria. However our results indicate that there are quite a few n-term protein acetylations identified. We summarize our results below.

N-terminal acetyltransferases are responsible for the n-terminal acetylation [31]. Irrespective of the type of acetylation, since MODa allowed us to look at the global acetylation level i.e., all possible shifts of 42Da, we plotted the acetylation frequencies at different time points as a function of time (see Figure 3). The total number of acetylations seem to go up from exponential to stationary phases. In our analysis, most of the acetylations seem to be protein n-term acetylations. Either this acetylation occurs on the 1st methionine amino acid or it happens on the 2nd amino acid after the excision of the 1st methionine. Figure 3 also shows the total n-term acetylations and the total serine acetylations. Later, we show that the serine acetylations seem to happen ¿99% of the times at n-term.

E. coli grown on glucose have shown to accumulate acetate (Cite proper references). Our results might suggest the same given the increase in the number of acetylated proteins from exponential to stationary phases. Table 1 shows the acetylated proteins found under these 2 different phases under E. coli growth under glucose limiting conditions.

Colin Brown's analysis could go here

Out of 3919 total acetylations at all 9 time points combined, 3373 (86%) of them seem to occur at n-term of the protein. A look-up of the amino acid position for n-term acetylations on ID'ed peptides showed that these were the 2nd amino acid in most (quote?) of the cases. May be add about the signal peptides here. Also, most of the acetylations seem to occur on serine (2358 out of 3919). Among n-term protein acetylations, 2350 (70%) happened on serine, 483 (14.3%) happened on alanine and 260 (7.7%) were on threonine. These were the top 3 frequent n-term acetylations. A literature search also showed us the same frequently occuring n-term acetylated amino acids in E. coli ***cite***.

2.5 *E. coli* Phosphorylations, carboxylations, and nitrosylations

Few PTMs in *E. coli* have been shown to be key in previous studies (Cite those studies). In particular, we looked at phosphorylation, carboxylation, nitrosylation and oxidation. First, we looked at the phosphorylations in *E. coli*. Most of the studies that looked at phosphoproteome involved enriching the sample for this PTM using techniques such as IMAC, TiO2 and then running mass-spectrometry on the sample. Figure 4 shows the phosphorylations at different points of the growth curve. We considered only the phosphorylations on serine, threonine and tyrosine, as these are well known amino acids that get phosphorylated in most species including *E. coli* [18]. The number of phosphorylations seem to be small. This is expected, as the sample is not enriched for phosphorylations. An interesting observation is that there is an increase of the number of phosphorylations from exponential to stationary phases. Such pattern is also previously shown in phosphoproteomic studies [29] probably pointing to a role of phosphorylation in later stages of the growth cycle.

We then looked at carboxylations (See Suppl Figure S2)and nitrosylations (See Suppl Figure S3). Like, phosphorylations, carboxylations seem to increase during the later stages of the growth indicating a likely role of this PTM in late stationary phase. ***(There are studies linking cellular growth rate to carboxylases, may be related to what is shown here?) A previous study in S-nitrosylation in E.coli [27] (read this to see if the pattern or the results hold).***

2.6 Oxidative damage and repair in *E. coli*

Next, we looked into oxidation of proteins. This is an important PTM as it has been linked to various diseases, along with a primary cause in aging. Oxidation frequently occurs on cysteine and methionine. However in our sample, since cysteine is capped to avoid disulphide bond formation, we expect +16 Dalton oxidation modification to occur primarily on methionone (see Suppl Figure 6).

E. coli grown aerobically react with oxygen in the atmosphere producing reactive oxygen species such as H2O2 CiteGonzalez-Flecha B, Demple B (1995) Metabolic sources of hydrogen peroxide in aerobically growing Escherichia coli. J Biol Chem 270: 13681–13687. Targets for these ROS are generally proteins and lipids, that get oxidized. This oxidized state affects the protein structure and hence results in changes in its function leading to disturbances in the metabolism. Bacteria have genetic systems that responds to this oxidative stress through oxyR, SoxXY etc or through reductases that reduce protein to its original non-oxidized state. Here, we investigated the levels of oxidation at different phases of growth.

Figure 5 (A) shows the global oxidation levels occuring on all amino acids. The oxidations seem to go down from exponential to stationary to long stationary phases. Figure 5 (B) shows the oxidation occuring on methionine only. The pattern seems to be same as the global oxidation level i.e., goes down from exponential to stationary phases. This result seemed interesting as we expected oxidative stress to increase under these glucose-starvation conditions in *E. coli* at later stages of the growth. May be the bacterial genetic systems that respond to oxidative stress are acting to bring back the oxidation levels down to protect *E. coli* against this oxidative damage. To see if this is the case, we looked into literature and found a class of reductases that fix methionine sulfoxide (MetSO) back to methionine (Met). In *E. coli* REL606 strain, these are MsrA and yeaA. Since MODa also outputs the peptide and hence the protein level information, we plotted the

protein abundances of these reductases in Figure 6. There seems to be a sudden increase in these levels during stationary and long-stationary phases that might explain the lower levels of methionine oxidation levels. To double-check our result, we plotted the previous Sequest search results that were used to plot the protein abundances in our preceding paper using this dataset. We could not use the same search results for our multiple PTM analysis here, as the only variable modification used in that study is oxidation methionine which is typical. The Sequest result is shown as a fig:SequestMODaFig (see Suppl Figure 7).

2.7 Other PTMs and artifacts

We then investigated few frequently occuring modifications on a particular site in 1 or few proteins. Our analysis revealed 2 such modifications in succinylation and pyroglutamate formation. Succinylation modification seem to frequently occur at different phases of *E. coli* growth, interestingly all on the succinyl-transferase. However, we were not able to find any substrates for this transferase. This was also previously identified in another study, however in that study the authors enriched the sample for this PTM. Here we would like to point out that this large-scale study to analyze different PTMs without any enrichment was able to identify and characterize not only widely occuring PTMs, but other low-abundant PTMs, such as succinylation. Next, we looked at pyroglutamate formation from n-glutamine. This is a well known PTM that stabilizes the protein to avoid any further n-terminal processing. Interestingly, the PTM levels of this conversion seems to be more or less same at different phases of the growth (see Suppl Figure 8).

One of the artifacts caused by mass-spectrometry instrumentation is the Na and K adduct formation. The frequency of this mass-shift seems to be 3-4%, combined. To validate if these are really adducts, or if it is the PTM on any particular amino acid, we plotted the profiles of Na and K adducts (see Suppl Figure S4). This mass-shift seems not to occur on H, K and R but seemed to occur randomly on all the other amino aicds. Histidine, lysine and arginine are amino acids that have side chains that are basic and carry a small positive charge at physiological pH. This explains that this mass-shift is indeed adducts from Na and K, as these adducts don't form on basic amino acids at physiological pH. Since including these Na and K adducts seem to increase the sensitivity of the search, we give a general recommendation to include these adducts as variable modifications in search algorithms that limit the number of PTMs as variable modifications.

2.8 Novel phosphoserinegluconylation on ribosomal protein S6

Finally, we did an analysis where we changed the mass-range to include large PTMs upto 300Da, compared to 200Da in the earlier run. The trend of the PTMs remained same (not shown), however including a larger range provided information on other larger PTMs. For example, a +258 Da shift on serine is found consistently at all time points of the growth curve. When we looked into into literature, this seemed to be a combination of both phosphorylation and gluconylation on Serine. However it seems to form on the ribosomal protein S6 very frequently.

To summarize, MODa is useful not only to understand known PTMs very well, but can unravel new PTMs that have biological relevance.

3 Discussion

(Add how MODa uses score and other features to calculate the probability?), (Also add how MODa can distribute the hits with the charge state?)

We looked at temporal changes of the post-translational modifications in *E. coli* REL606 strain. For this, we used mass-spectrometry based proteomics data. This data is collected under glucose limiting conditions for different phases of the growth i.e., exponetial to long-stationary phases upto 2 weeks under these conditions. For the 9 time points analyzed (3 hours to 2 weeks), several interesting insights into the frequency and the temporal changes of post-translational modifications seen in this analysis. First, we show that consistently 30% of the proteome is modified at all the 9 time points analyzed. Second, we show that n-term protein acetylation goes up from exponential to stationary phases pointing to a likely role of this modification during late stages of the growth. Similar results were found for phosphorylation, carboxylation too. However, oxidative stress seem to go down from exponetial to stationary phases. Nutrient limitation, accumulation of toxic substances and change in pH were supposed to increase the oxidiative stress, however *E. coli* has a mechanism to protect the cell from oxidative damage using the sulfoxide reductases fixing MetSO back to Met. Finally, the computational techniques used in this study can be used with any mass-spectrometry based proteomics data to identify or characterize novel post-translational modifications.

Recently, the focus has shifted from identifying peptides/proteins to characterizing PTMs as shown in this recent review article [22]. However identifying all the PTMs present in the sample is limited both by experimental enrichment techniques that cannot be applied to all kinds of PTMs at once and also from the computational limitations, although there has been an improvement in the latter limitation lately. Some of the future applications of current methodology of identifying PTMs at once is to help understand the PTM crosstalk under each condition analyzed. This in turn helps us to understand the protein regulation or function specifically generally caused by many PTMs acting in concert, as earlier witnessed [23]. So, analysis of PTMs in *tandem* is a necessity to get accurate reflections of the underlying interactions of the PTMs. Previous work in this area used the PTMs deposited in the database, however care must be taken, as these PTMs were identified at different growth phases under differing conditions [23].

In a typical mass-spectrometry based proteomics experiment, proteins are first digested into peptides and then the sample is enriched for any PTMs of interest. This sample is then analyzed by mass-spectrometry. Peptide search algorithms are then run on this mass-spectrometry data to computationally identify peptides and PTMs associated with it. These search algorithms generall fall into 3 categories: (1) Sequence or database search algorithms, (2) de novo sequence search and (3) Hybrid approach that is a mixture of partial de novo search followed by a database search. Most of the current search algorithms such as Mascot [24], Sequest [9], OMSSA [11], X!Tandem [5] search for only few variable modifications, i.e., oxidation of methionine etc or a small targeted list of PTMs. This technique of restricted search reduces the size of the database to search and hence alleviates lot of computational limitations [19]. However this limited search identifies only PTMs that were used in the target list, limiting the analysis to either the previously known PTMs or target PTMs. However, recently, software for unrestricted search of PTMs is starting to be available. MODa [20] is one such unrestricted search engine, along with others such as TagRecon [6] and Byonic [1]. We used MODa, a naive based multi-blind spectral alignment algorithm, to look for PTMs in our *E. coli* dataset. A few studies that used MODa in the past looked at hand-ful of

interesting PTMs in depth such as [14], or looked at wider coverage of PTMs such as [17]. Our interest is the wider coverage of PTMs and look in depth at both the frequently occuring mass-shifts and the well known PTMs mapped from known mass-shifts.

Most of the peptide identification search algorithms require a list of PTMs to search for and generally this list is limited to 6. However there are hundreds of PTMs known to date, that are identified *in vivo* and well documented in databases like UNIMOD, RESID. Since MODa outputs mass-shifts instead of the PTMs in the peptide-spectral matches, we used UNIMOD to manually map the mass-shift to the appropriate PTM that matches the mass-shift. However we did not try and match all the mass-shifts, but investigated in detail the frequent and well known mass-shifts identified by MODa. This resulted in analysis of +1 Da mass-shift (C13 peak detection), oxidation, acetylation, Na and K adducts, along with some widely studied PTMs such as phosphorylation, carboxylation and nitrosylation. This program is previously used for similar large-scale analysis with urinary proteomics and they identified novel PTMs. However the study used 2 programs and considered the overlap of PTMs as highly confident. Instead here, our focus is not to identify highly-confident PTMs that can be used later as biomarkers, but to get a wider coverage at a lower FDR of 1% and look at the time course or evolution of these PTMs during the entire 2 weeks of *E. coli* growth.

N-terminal processing of proteins is a well-known PTM in diverse kinds of species ranging from eukaryotes to bacteria [15]. N-terminal acetylation has been shown in the past to provide insights into the nature of the n-terminal of the proteome [13]. In eukaryotes, n-terminal acetyl-transferases are well characterized and studied, for example, as shown in these studies [25], [26]. In *E. coli*, NATs are primarily characterized into mainly rim types [32], [33]. There is lot of work done on lysine acetylation in *E. coli*, however there are not many n-terminal acetylation sites identified or investigated in detail in the past. Lysine acetylation E. coli, acetylation in exponential and stationary phases ***Colin's discussion***.

Protein oxidation by reactive oxygen species (ROS) is important and linked to various diseases including aging [30]. ROS generally seem to accumulate at later stages of growth, because of possible accumulation of toxic substances, nutrient limiting conditions, changes in pH etc. *E. coli's* internal machinery to tackle this oxidative stress comes in many forms at different levels and in this analysis, we were interested at the protein level. *E. coli* has a class of sulfoxide reductases [2] [34] that fix methionine sulfoxide back to methionine. MsrA and yeaA are the proteins found in this reductase class of enzymes. These reductases has been shown to protect *E. coli* from oxidative damage [10]. Here, we looked at the abundances of these proteins during these late stationary phases, where oxidative stress seems to be more. Our earlier work has already shown the induction of these stress proteins at stationary phases using the same mass-spec data set ***(CiteRef John's paper).***

One of the PTM networks that is well studied in the past is the phosphorylation signaling cascade [21] (also look for phosphorylation papers that you used in the past at NCBI). These are well known signalling mechansims and have been shown to respond quickly to stress during the late-stationary phases [29]. Our analysis also supports this behavior and would like to emphasize that while the previous paper investigated only 1 PTM, our analysis investigated most of the PTMs in tandem.

Look for PTMs on SspA

There are some limitations with respect to how we did MODa searches. One of the limitations in the searches performed in the current work is that we used the default mass-range search between

-200 and 200 Da (and one other search with -100 and 300 Da range). So, we miss out on larger PTMs i.e., polyubiquitination tails etc. Another computational limitation is that we looked for only 1 possible modification on each of the peptides identified. This is because MODa was shown in the past to generate many false positives if searched for multiple PTMs on the same peptide. After the peptides are ID'ed, generally the PTMs are validated by a 2nd round by using programs like Ascore etc. However here, we did not do any 2nd round of validating peptides, as MODa is shown to ID high-confidence PTM identifications in its search. Another, probably important limitation is that we did not do any PTM level quantitation or try to understand the specific PTM stoichiometry, as these require sophisticated experimental instrumentation, good enrichment protocol and the algorithms to characterize the PTMs associated with proteins. However MODa has been a software suite that has been recently used in both in bacterial proteomics as well as biomedical applications such as urinary proteomics to find biomarkers [17].

Current whole-cell models [4] integrate diverse kinds of OMICS data i.e., transcriptomics, proteomics not only to refine the existing models, but also make reasonable predictions on which genes/enzymes are key for example, for a particular metabolite production etc. Here, we argue that including the modification information (i.e., number of modified proteins to that of the unmodified version) would improve the protein level abundances and thus will improve the computational predictions.

http://www.nature.com/ncomms/2014/140725/ncomms5405/full/ncomms5405.html

4 Conclusions

The modified protein seems to be 30% during exponential as well as the long stationary phases. Acetylation, in particular n-term acetylation seems to go up from exponential to stationary phases. Surprisingly oxidation seemed to go down from exponential to stationary phases, owing to sulfoxide reductases playing a role in protecting *E. coli* from oxidative damage by fixing methionine sulfoxide back to methionine. Phosphorylations and carboxylations also seem to increase indicating a likely role of these PTMs in late-stationary phases of the growth curve. A novel phosphoserinegluconylation on ribosomal protein seem to happen frequently. Finally, we would like to conclude that unrestricted search engines can be used to identify frequently occuring PTMs, which can then be used with restricted search algorithms to improve the sensitivity of the ID'ed peptides.

5 Materials and Methods

5.1 *E. coli* growth

The details of *E. coli* growth is provided in our manuscript that described the initial analysis of this data ***(See citeRefHouserJRetal2015)***. Details on the mass-spectrometry is provided in the same paper as well. In short, trypsin was used to digest the proteins and then the sample is analyzed using liquid chromatography mass spectrometry (LC/MS) on a LTQ-Orbitrap (Thermo Fisher). For each time point, there were 3 biological replicates that were analyzed.

5.2 Post-translational modification identification and analysis

Mass-spectrometry raw data was then converted into mzXML files to input into MODa [20]. MODa is a naive based spectral alignment algorithm that identifies peptides and their associated PTMs from the input spectral files. Difference between MODa and most of other search engines is that MODa outputs mass-shifts instead of directly outputting the post-translational modifications. So, to convert these mass-shifts to known PTMs, we used UNIMOD database. We did a manual mapping of the known PTMs, i.e., if we see +16Da, from UNIMOD we know that it is oxidation. For some mass-shifts, we looked at the amino-acid profiles i.e., which amino acid has the mass-shift frequently to confirm the UNIMOD PTM list i.e., from oxidation profile, it is clear that Met is the most frequent and from UNIMOD, it is obvious that this corresponds to methionine oxidation. We also looked at well known PTMs that were observed in previous studies in *E. coli*. For example, even though carboxylations and nitrosylations seemed rare (from frequencies of mass-shifts), since we know the mass-shift and the expected amino acids on which this modification happens, we were able to plot the temporal changes. Some frequently occuring mass-shifts did not map to PTMs, but they did map to Na and K adducts, as evident from UNIMOD. In this analysis, we did not consider any mutations. In future, we plan to look at these separately.

We ran separate MODa searches for each of the 9 time points. Since there were 3 biological replicates, this resulted in total 27 MODa searches. To speed up the searches, we used UT TACC computing resources. The enzyme used in the searches is trypsin with fully-tryptic and no proline rule. The missed cleavages allowed are 2. Since the fragmentation technique used is CID, we looked for b/y ions. The mass-tolerance used for the precursor ion is 10 ppm, while the mass-tolerance used for the product ion is 0.5 Da. We set carbamidomethylation of cysteine as a static or fixed modification. As mentioned earlier, MODa requires a mass range to search for variable modifications, so we run MODa searches for 2 scenarios: (1) mass range between -200 to 200Da and (2) second search with mass range between -100 to 300Da. Since we were interested in PTMs, there was no need to include negative mass-shifts, but we included these as mentioned in the MODa protocol. We used REL606 NCBI *E. coli* sequence library.

Once we have the peptide hits from MODa output, we used target-decoy approach [8] to identify high-confidence hits. In this approach, we reverse the original REL sequences and concatenate these to the original sequence database to form a target-decoy database. This database is twice the size of the original sequence database. The idea is that there are as many false positive hits to that of the original database as that of the decoy database. We used a 1% FDR that is a general norm in mass-spectrometry based proteomics searches.

5.3 Raw data and analysis scripts

All raw data and analysis scripts are available online in the form of a git repository at `https://github.com/wilkelab/Ecoli_PTMs`.

6 Author Contributions

Conceived and designed the experiments: V.S, J.E.B, E.M.M, and C.O.W. Performed the experiments: V.S. Analyzed the data: V.S, C.W.B, M.D.P, J.E.B, E.M.M. and C.O.W. Wrote the paper:

V.S, C.W.B, D.R.B, C.B, M.D.P, J.E.B, E.M.M and C.O.W.

7 Acknowledgments

This project was funded by ARO Grant W911NF-12-1-0390. We thank John Houser and Kevin Drew for useful discussions. We thank the Bioinformatics Consulting Group (BCG) and the Texas Advanced Computing Center (TACC) at UT for high-performance computing resources.

References

1. M. Bern, Y. J. Kil, and C. Becker. Byonic: advanced peptide and protein identification software. *Curr Protoc Bioinformatics*, Chapter 13:Unit13 20, 2012.

2. N. Brot, L. Weissbach, J. Werth, and H. Weissbach. Enzymatic reduction of protein-bound methionine sulfoxide. *Proc Natl Acad Sci U S A*, 78(4):2155–8, 1981.

3. E. Charbaut, V. Redeker, J. Rossier, and A. Sobel. N-terminal acetylation of ectopic recombinant proteins in escherichia coli. *FEBS Lett*, 529(2-3):341–5, 2002.

4. M. W. Covert, N. Xiao, T. J. Chen, and J. R. Karr. Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics*, 24:2044–50, 2008.

5. R. Craig and R. C. Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–7, 2004.

6. S. Dasari, M. C. Chambers, R. J. Slebos, L. J. Zimmerman, A. J. Ham, and D. L. Tabb. Tagrecon: high-throughput mutation identification through sequence tagging. *J Proteome Res*, 9(4):1716–26, 2010.

7. H. P. Driessen, W. W. de Jong, G. I. Tesser, and H. Bloemendal. The mechanism of n-terminal acetylation of proteins. *CRC Crit Rev Biochem*, 18(4):281–325, 1985.

8. J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4(3):207–14, 2007.

9. J. K. Eng, A. L. McCormack, and J. R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom*, 5(11):976–89, 1994.

10. B. Ezraty, R. Grimaud, M. El Hassouni, D. Moinier, and F. Barras. Methionine sulfoxide reductases protect ffh from oxidative damages in escherichia coli. *EMBO J*, 23(8):1868–77, 2004.

11. L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant. Open mass spectrometry search algorithm. *J Proteome Res*, 3(5):958–64, 2004.

12. Y. Gordiyenko, S. Deroo, M. Zhou, H. Videler, and C. V. Robinson. Acetylation of l12 increases interactions in the escherichia coli ribosomal stalk complex. *J Mol Biol*, 380(2):404–14, 2008.

13. A. O. Helbig, S. Gauci, R. Raijmakers, B. van Breukelen, M. Slijper, S. Mohammed, and A. J. Heck. Profiling of n-acetylated protein termini provides in-depth insights into the n-terminal nature of the proteome. *Mol Cell Proteomics*, 9(5):928–39, 2010.

14. H. J. Kim, S. Ha, H. Y. Lee, and K. J. Lee. Rosics: Chemistry and proteomics of cysteine modifications in redox biology. *Mass Spectrom Rev*, 2014.

15. Y. Kimura, Y. Saeki, H. Yokosawa, B. Polevoda, F. Sherman, and H. Hirano. N-terminal modifications of the 19s regulatory particle subunits of the yeast proteasome. *Arch Biochem Biophys*, 409(2):341–8, 2003.

16. A. I. Lamond, M. Uhlen, S. Horning, A. Makarov, C. V. Robinson, L. Serrano, F. U. Hartl, W. Baumeister, A. K. Werenskiold, J. S. Andersen, O. Vorm, M. Linial, R. Aebersold, and M. Mann. Advancing cell biology through proteomics in space and time (prospects). *Mol Cell Proteomics*, 11(3):O112 017731, 2012.

17. L. Liu, X. Liu, W. Sun, M. Li, and Y. Gao. Unrestrictive identification of post-translational modifications in the urine proteome without enrichment. *Proteome Sci*, 11(1):1, 2013.

18. B. Macek, F. Gnad, B. Soufi, C. Kumar, J. V. Olsen, I. Mijakovic, and M. Mann. Phosphoproteome analysis of e. coli reveals evolutionary conservation of bacterial ser/thr/tyr phosphorylation. *Mol Cell Proteomics*, 7(2):299–307, 2008.

19. L. McHugh and J. W. Arthur. Computational methods for protein identification from mass spectrometry data. *PLoS Comput Biol*, 4(2):e12, 2008.

20. S. Na, N. Bandeira, and E. Paek. Fast multi-blind modification search through tandem mass spectrometry. *Mol Cell Proteomics*, 11(4):M111 010199, 2012.

21. J. V. Olsen, B. Blagoev, F. Gnad, B. Macek, C. Kumar, P. Mortensen, and M. Mann. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, 127(3):635–48, 2006.

22. J. V. Olsen and M. Mann. Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol Cell Proteomics*, 12(12):3444–52, 2013.

23. M. Peng, A. Scholten, A. J. Heck, and B. van Breukelen. Identification of enriched ptm crosstalk motifs from large-scale experimental data sets. *J Proteome Res*, 13(1):249–59, 2014.

24. D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–67, 1999.

25. B. Polevoda and F. Sherman. Composition and function of the eukaryotic n-terminal acetyltransferase subunits. *Biochem Biophys Res Commun*, 308(1):1–11, 2003.

26. B. Polevoda and F. Sherman. N-terminal acetyltransferases and sequence requirements for n-terminal acetylation of eukaryotic proteins. *J Mol Biol*, 325(4):595–622, 2003.

27. D. Seth, A. Hausladen, Y. J. Wang, and J. S. Stamler. Endogenous protein s-nitrosylation in e. coli: regulation by oxyr. *Science*, 336(6080):470–3, 2012.

28. N. C. Soares, M. P. Cabral, C. Gayoso, S. Mallo, P. Rodriguez-Velo, E. Fernandez-Moreira, and G. Bou. Associating growth-phase-related changes in the proteome of acinetobacter baumannii with increased resistance to oxidative stress. *J Proteome Res*, 9(4):1951–64, 2010.

29. N. C. Soares, P. Spat, K. Krug, and B. Macek. Global dynamics of the escherichia coli proteome and phosphoproteome during growth in minimal medium. *J Proteome Res*, 12(6):2611–21, 2013.

30. E. R. Stadtman. Protein oxidation and aging. *Science*, 257(5074):1220–4, 1992.

31. K. K. Starheim, K. Gevaert, and T. Arnesen. Protein n-terminal acetyltransferases: when the start matters. *Trends Biochem Sci*, 37(4):152–61, 2012.

32. S. Tanaka, Y. Matsushita, A. Yoshikawa, and K. Isono. Cloning and molecular characterization of the gene riml which encodes an enzyme acetylating ribosomal protein l12 of escherichia coli k12. *Mol Gen Genet*, 217(2-3):289–93, 1989.

33. A. Yoshikawa, S. Isono, A. Sheback, and K. Isono. Cloning and nucleotide sequencing of the genes rimi and rimj which encode enzymes acetylating ribosomal proteins s18 and s5 of escherichia coli k12. *Mol Gen Genet*, 209(3):481–8, 1987.

34. X. H. Zhang and H. Weissbach. Origin and evolution of the protein-repairing enzymes methionine sulphoxide reductases. *Biol Rev Camb Philos Soc*, 83(3):249–57, 2008.

Figures

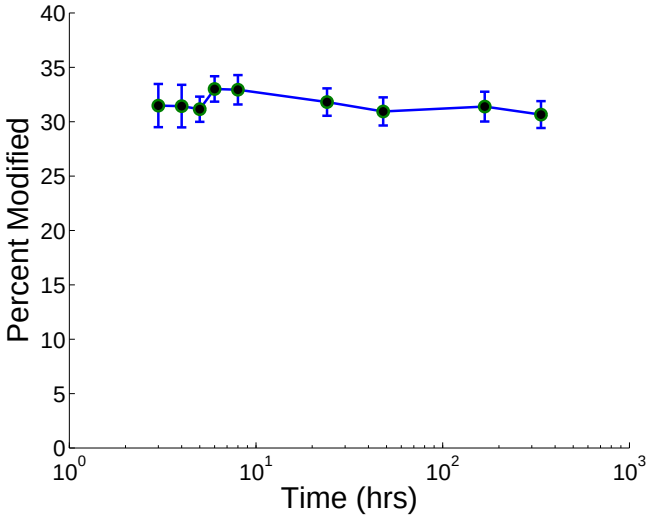

Figure 1. *E. coli* modified proteome. The total number of modified peptide-spectral matches seem to be constant at 30% for all 9 time points. .

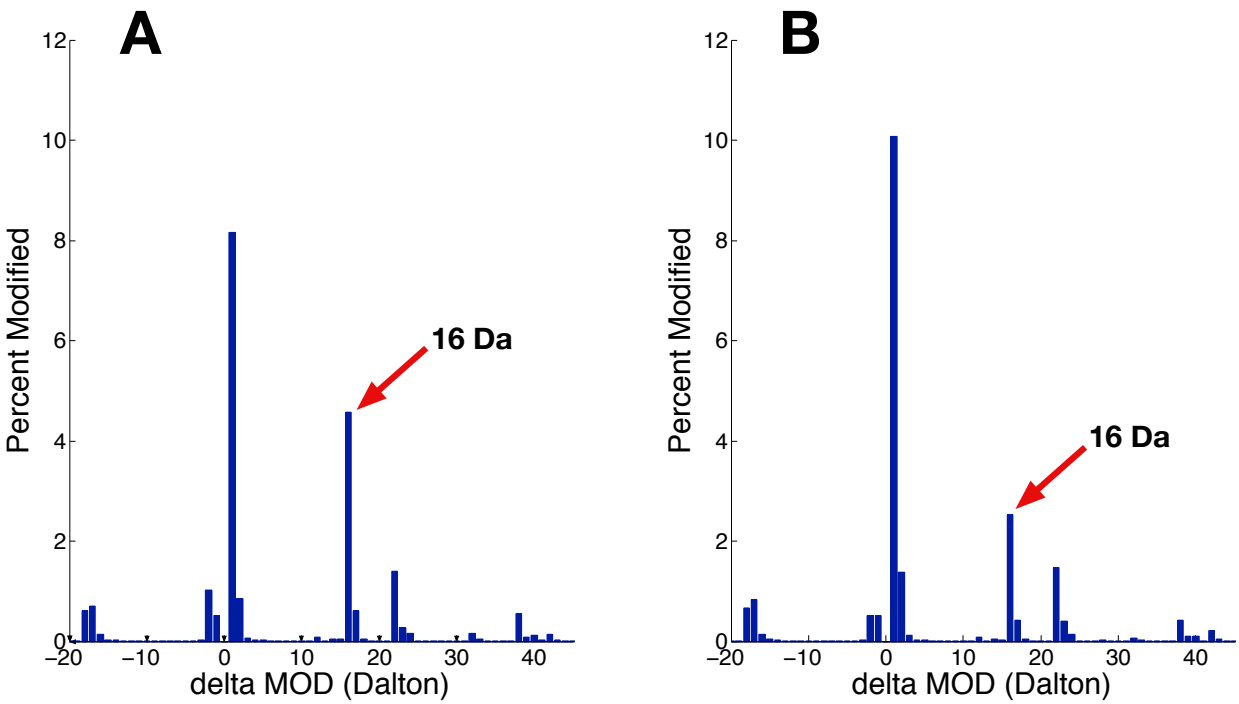

Figure 2. MODa outputs mass-shifts. A naïve based algorithm like MODa can alleviate the requirement of guessing PTMs beforehand. However MODa outputs mass-shifts on the amino acids. We can then use PTM databases like UNIMOD to map the mass-shift to the most probable PTM. (A) and (B) are the frequencies of the mass-shifts observed at 3 hours and 2 weeks of the *E. coli* growth..

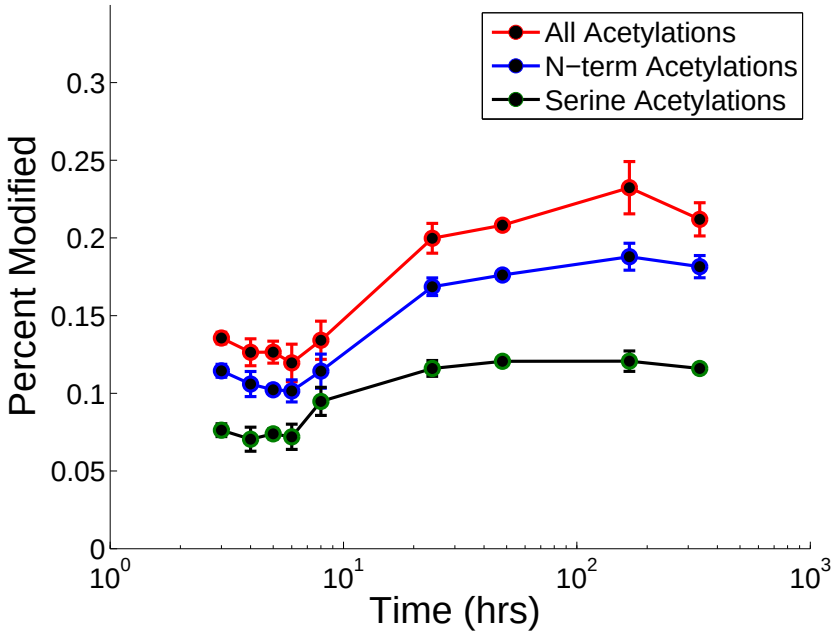

Figure 3. Protein n-term acetylations are dominant. Total number of acetylations as well as the n-term/serine acetylations seem to go up over 2 weeks. *E. coli* grown on glucose generally tend to accumulate acetate, perhaps this resulted in increase in acetylations..

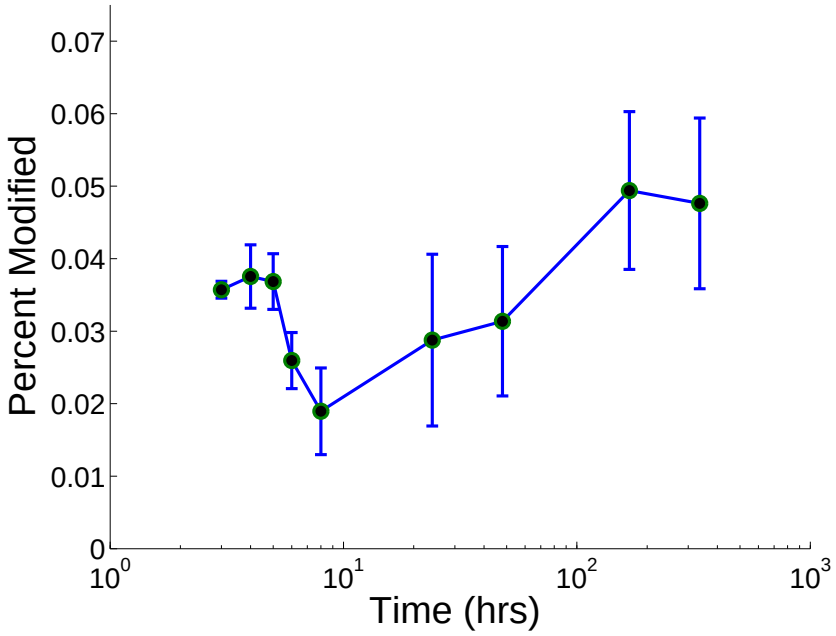

Figure 4. Phosphorylations are rare. Phosphorylations seem to be low and tend to increase during last week of growth. 2 frequently phosphorylated proteins in MODa search output are phosphoglucomutase and elongation factor Tu.

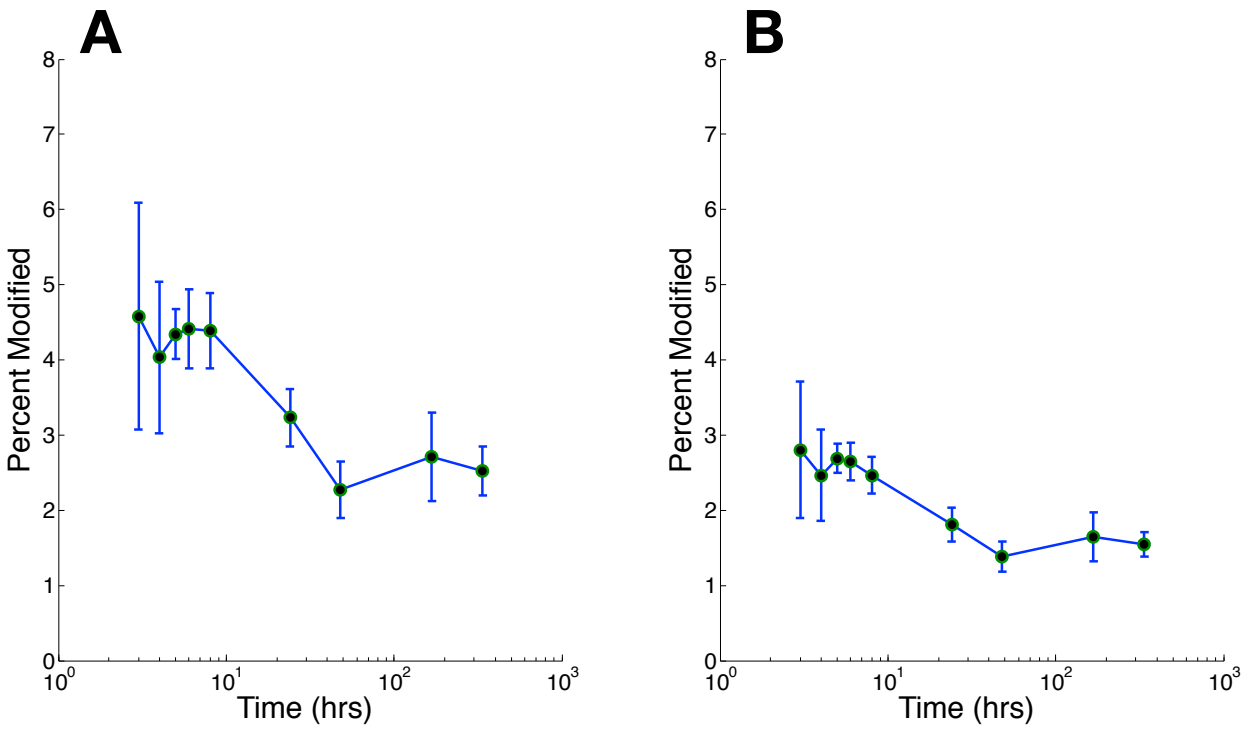

Figure 5. Oxidations go down over 2 weeks. (A) Total number of oxidations seem to go down from exponential to stationary phases. (B) The same trend follows for methionine oxidations too.

Supplementary Figures

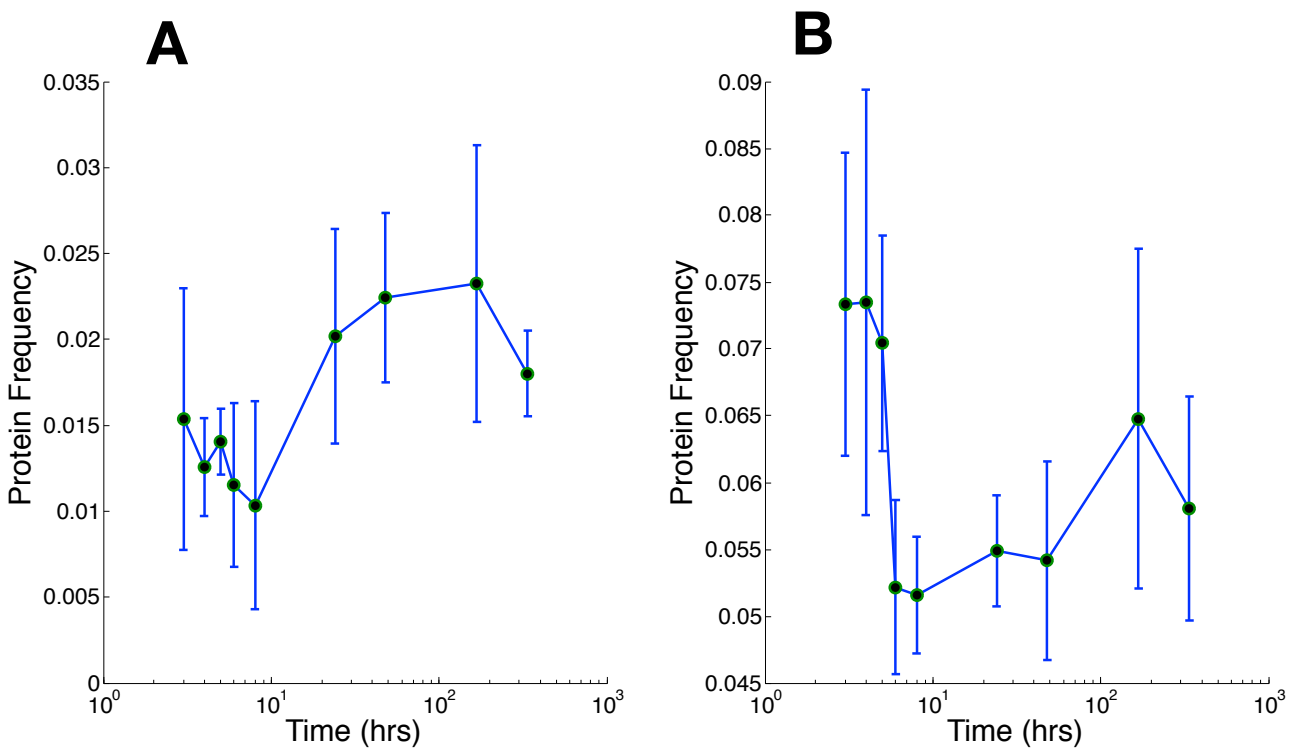

Figure 6. Relative protein abundances of methionine sulfoxide reductases MsrA and MsrB. (A) Protein abundances of MsrA seems to increase going from exponential to long-stationary phase. One of the sites oxidized on MsrA is FQAA[M+16]LAADDDR. (B) MsrB also seems to increase going into the long-stationary phase, protecting *E. coli* from oxidative damage.

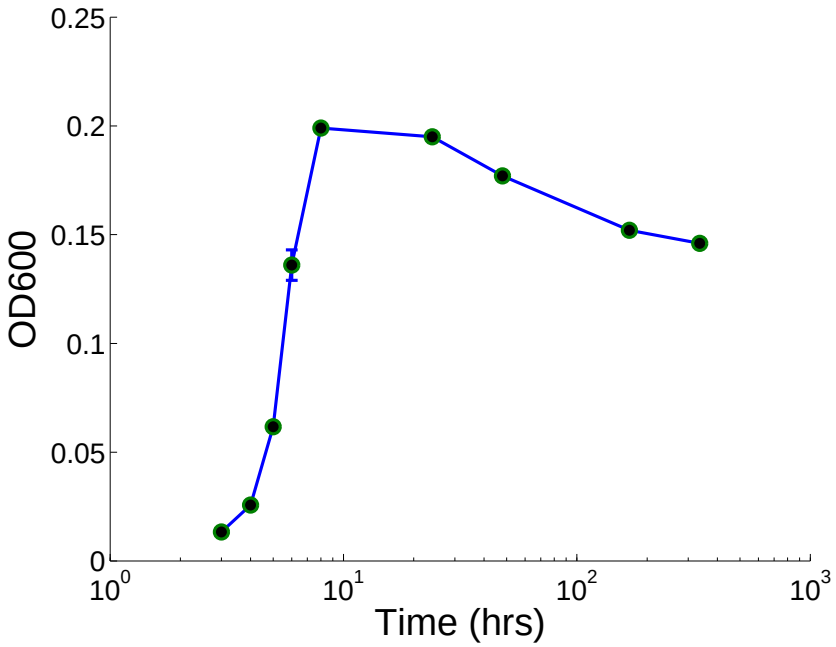

Figure S1: OD600 curve . Growth curve (OD600) of REL606 under glucose starvation conditions.

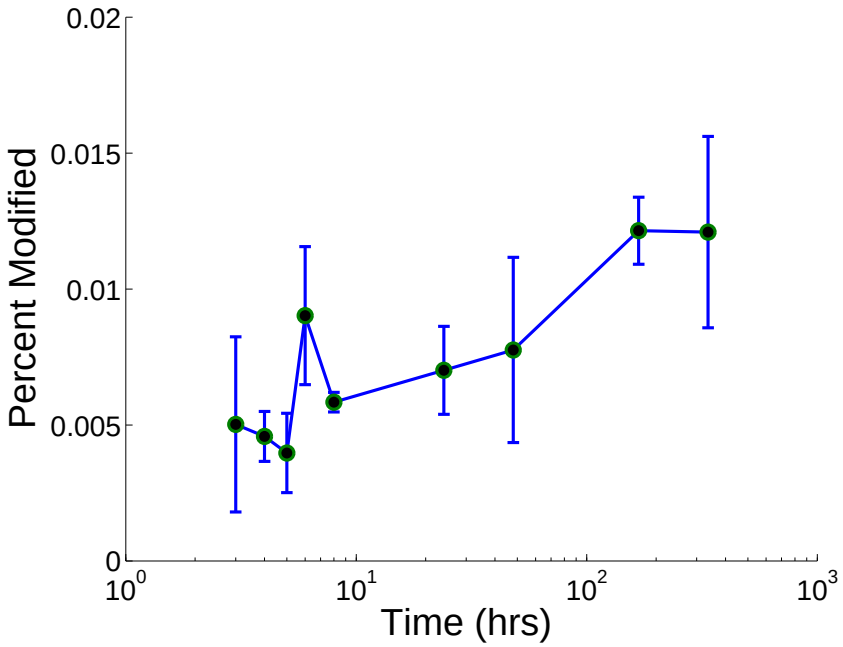

Figure S2: E. coli Carboxylations . Carboxylations seem to go up.

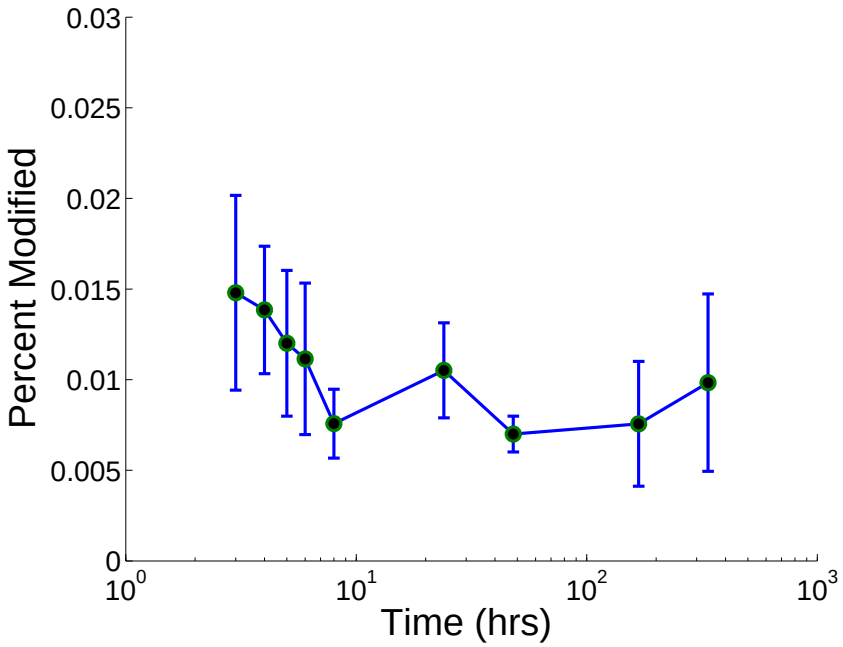

Figure S3: E. coli Nitrosylations. Nitrosylations seem not to change much during the entire growth period.

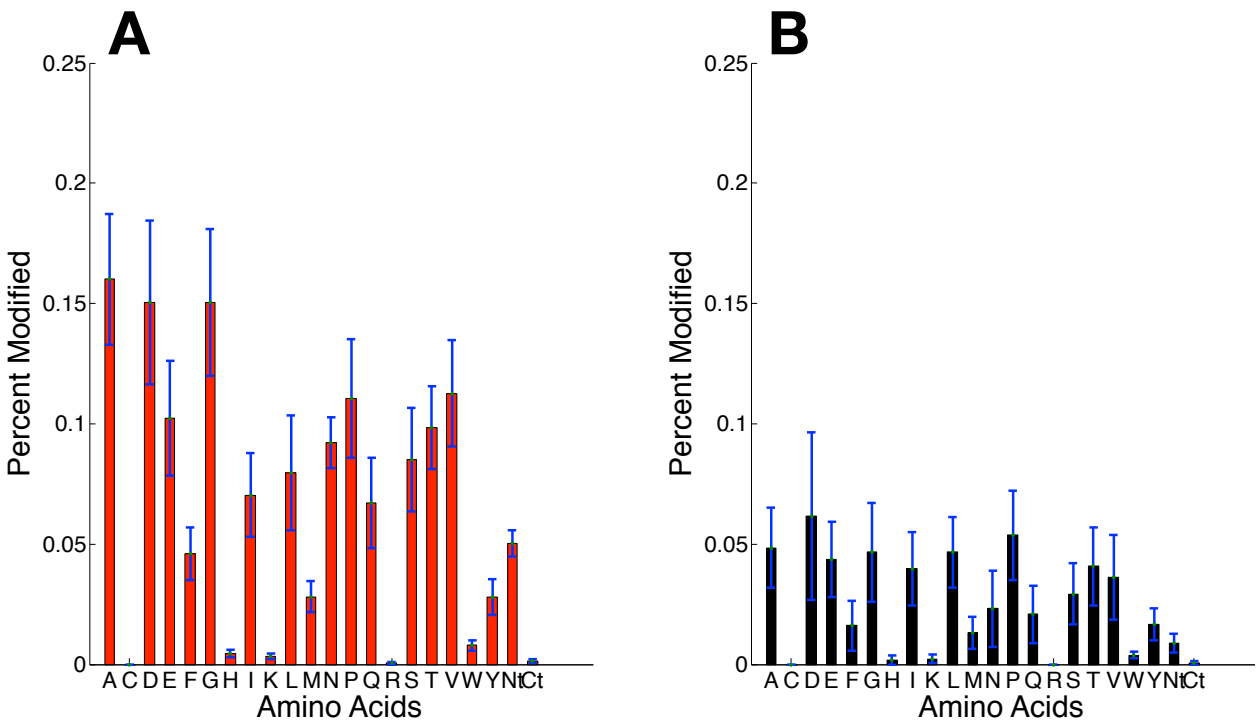

Figure S4: Na and K adducts. (A) Na and (B) K adducts seem to happen on all amino acids except histidine, lysine and arginine. H, K and R are amino acids that are basic and carry some (+) charge at physiological pH.

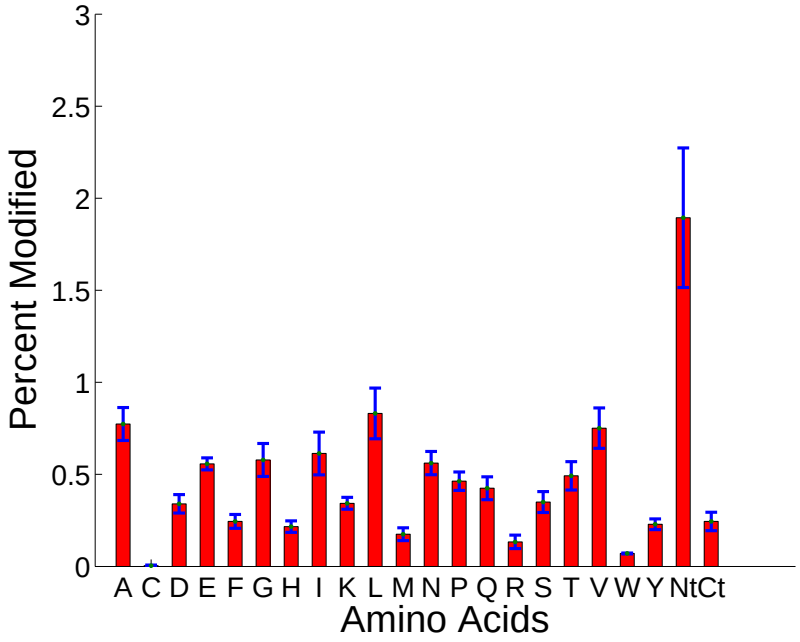

Figure S5: +1Da shift occurs randomly. We cannot infer that this +1 Dalton mass-shift is deamidation as it occurs randomly on all amino acids, inferring it is carbon-13 (C13) peak picking as previously shown in many studies.

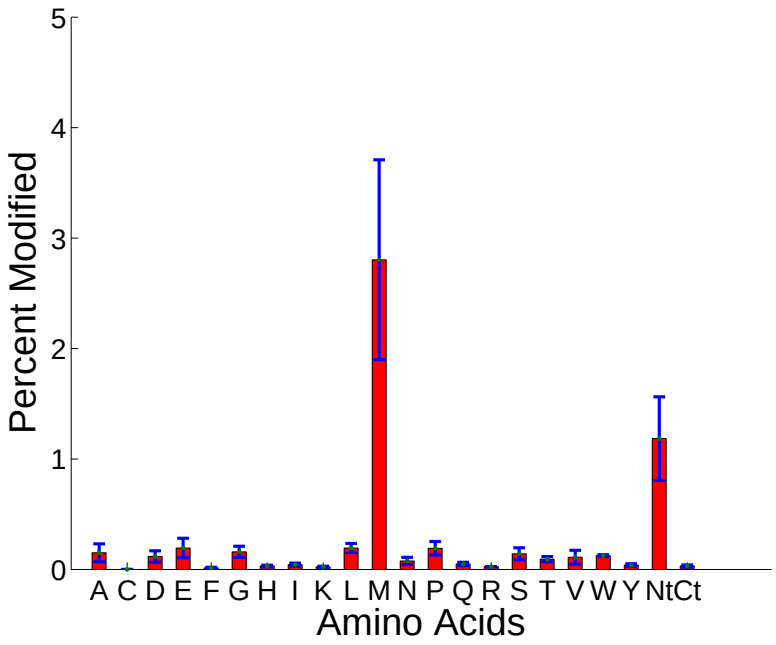

Figure S6: Oxidation is dominant on methionine. Even though many amino acids could be oxidized, in this data set, oxidation seems to occur primarily on methionine, as expected.

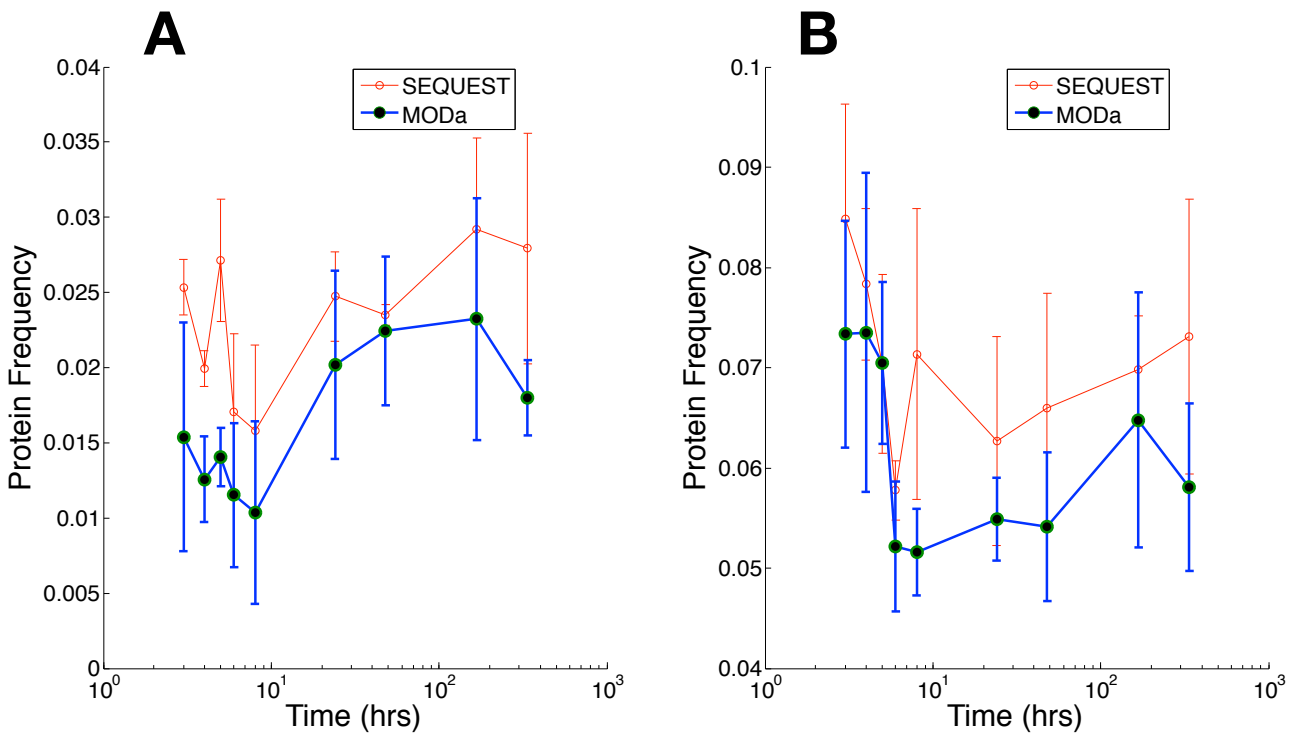

Figure S7: Sequest and MODa results agree. We compared MODa protein abundance results with Sequest to double-check the protein levels of the sulfoxide reductases that protect *E. coli* from oxidative stress by fixing MetSO to Met. As seen, MODa and Sequest results seem to agree well for (A) MsrA and (B) MsrB.

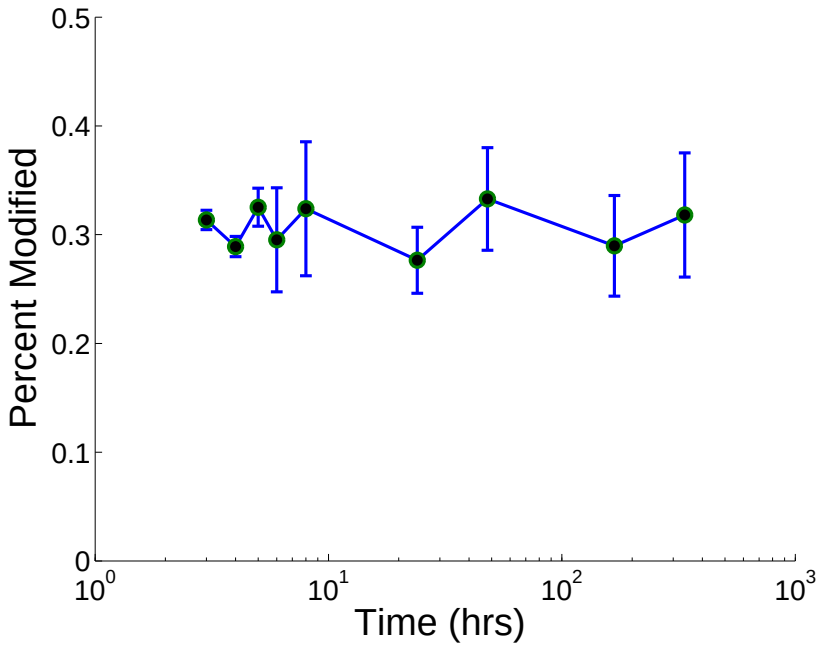

Figure S8: Glutamine to pyroglutamate conversion. Glutamine to pyroglutamate happens to stabilize the protein. This conversion seems to be consistent across both the exponential and stationary phases.