N-terminal Protein Processing: A Comparative Proteogenomic Analysis*

Stefano Bonissone‡**, Nitin Gupta‡‡‡, Margaret Romine§, Ralph A. Bradshaw¶, and Pavel A. Pevzner**

N-terminal methionine excision (NME) and N-terminal acetylation (NTA) are two of the most common protein post-translational modifications. NME is a universally conserved activity and a highly specific mechanism across all life forms. NTA is very common in eukaryotes but occurs rarely in prokaryotes. By analyzing data sets from yeast, mammals and bacteria (including 112 million spectra from 57 bacterial species), the largest comparative proteogenomics study to date, it is shown that previous assumptions/perceptions about the specificity and purposes of NME are not entirely correct. Although NME, through the universal enzymatic specificity of the methionine aminopeptidases, results in the removal of the initiator Met in proteins when the second residue is Gly, Ala, Ser, Cys, Thr, Pro, or Val, the comparative genomic analyses suggest that this specificity may vary modestly in some organisms. In addition, the functional role of NME may be primarily to expose Ala and Ser rather than all seven of these residues. Although any of this group provide “stabilizing” N termini in the N-end rule, and de facto leave the remaining 13 amino acid types that are classed as “destabilizing” (in higher eukaryotes) protected by the initiator Met, the conservation of NME-substrate proteins through evolution suggests that the other five are not crucially important for proteins with these residues in the second position. They are apparently merely inconsequential players (their function is not affected by NME) that become exposed because their side chains are smaller or comparable to those of Ala and Ser. The importance of exposing mainly two amino acids at the N terminus, i.e. Ala and Ser, is unclear but may be related to NTA or other post-translational modifications. In this regard, these analyses also reveal that NTA is more prevalent in some prokaryotes than previously appreciated. Molecular & Cellular Proteomics 12: 10.1074/mcp.M112.019075, 14–28, 2013.

Although methionine is used to initiate protein synthesis for essentially all proteins, it is subsequently removed in a large percentage of cases, either by cleavage of an N-terminal “signal ” peptide (as part of cellular translocation mechanisms or precursor activations) or by the action of specific methionine aminopeptidases (MetAPs). Approximately two-thirds of the proteins in any proteome are potential substrates for the latter N-terminal methionine excision (NME),1 and MetAPs appear in all organisms from bacteria to eukaryotes (1). The second, or P2, amino acid in protein substrates is crucially important for NME because MetAP specificity mainly depends on the nature of this residue, a selectivity that is conserved across all species (1–5). These enzymes generally excise the N-terminal Met when the second residue is Gly, Ala, Ser, Thr, Cys, Pro, or Val (3, 6, 7) which are the amino acids smallest in size (based on radius of gyration of the side chain (8)). NME is a necessary process for proper cell functioning; it is included in the minimal genome set of eubacteria (9). Eukaryotes contain two MetAPs derived from a version in bacteria (MetAP1), and another found in archea (MetAP2) (11). Just as the deletion of MetAP eubacteria is lethal, the deletion of both MetAPs in yeast is also lethal (10).

In 1988, Arfin and Bradshaw (2) observed that the specificity of NME coincided with that of the N-end rule (NER) (12, 13), a ubiquitin-dependent protein degradation process that is based on the recognition of N-terminal residues. The stabilizing residues for the NER include Gly, Ala, Ser, Cys, Thr, Pro, and Val and, with the exception of Met, the destabilizing residues are all found to be in the class of P2-residues that are not substrates for the MetAPs. This suggested that NME acts to release Met from proteins whose stability is unaffected by the NER creating at the same time a second class of proteins, who have the potential for regulated turnover downstream of the cotranslational processing. However, despite ex-
tensive studies, this type of programmed protein turnover (requiring downstream removal of Met) has not been demonstrated to occur. An implication of this correlation is that exposing of the stabilizing residues may also contribute to increasing their lifetime.

The stabilizing residues exposed by the action of the MetAPs can be further modified. The most extensive of these reactions is N-terminal acetylation (NTA), which can occur on as much as 70–80% of the mass of the soluble protein in eukaryotes. Although the specificity of the N-acetyltransferase (NAT) responsible is not as rigid as the MetAPs, the principal substrates in the stabilizing class are usually the four smallest residues (Gly, Ala, Ser, and Thr) (6, 14). A second class of NATs can also modify the retained Met when the adjacent residues are Asp, Glu or Asn (15). The functional importance of this modification (in either case) is not known although it has been suggested that it may exert a protective effect against spurious aminopeptidase cleavages. Recently, Hwang et al. (16) have extended the NER to include $N$-acetylated termini as also destabilizing thus providing another possible function for this modification. In contrast, to date, very few instances of $N$-acetylation have been observed in bacteria. Other modifications can also occur in both eukaryotes and prokaryotes although they are generally much more limited in scope.

The specificity of the MetAPs suggest an apparent connection between NME and protein degradation. However, this connection has never been examined using high-throughput mass spectrometric data or a comparative genomics approach; thus it remains unclear whether exposing these stabilizing residues contributes to increasing protein half-life and thus represents a primary purpose of NME. (The connection between NME and NER in bacteria, which has an NER with a somewhat different profile (17), is even more obscure.) Recent studies provide some examples where disruption of NME via a single-residue substitution in the P2 position causes protein degradation (18–20); however, some of these experimental results are in conflict with the NER (13). Giglione et al. (20) have shown that NME triggers degradation of D2 protein in Caenorhabditis reinhardtii in the PSII complex after replacing the second (stabilizing) Thr residue by another amino acid to prevent NME. This replacement results in early degradation of D2 and instability of the PSII complex. From this, Giglione et al. (20) postulated that NME determines protein life-span via currently unknown machinery. However, because Bachmair et al. (12) classified Met as a stabilizing residue, it is not entirely clear why substituting one stabilizing residue (Met) by another one (Gly, Ala, Ser, Cys, Thr, Pro, or Val) should affect protein stability and the substitution may have other deleterious effects that are manifested in different ways.

The logic for analyzing NME and NER is shown in Fig. 1. NME exposes seven residues to be new N termini of proteins. Because this is presumably for some functional reason, the conventional assumption is that all seven residues must have functional importance as N termini. By the comparative genomics postulate (as defined in the text), evolutionary conservation of all seven at P2 should be observed. If all of these residues are not conserved, one of the two assumptions must be incorrect; either not all seven residues are important or the comparative genomics postulate holds, and only two of the seven residues are of functional importance as N termini, then the other five residues are inconsequential players and only these two residues should be evolutionarily conserved.

![Diagram of NME and NER](image)

**Fig. 1. Two alternative cases for NME function.** A, NME exposes seven residues to be new N termini of proteins. Because this is presumably for some functional reason, the conventional assumption is that all seven residues must have functional importance as N termini. By the comparative genomics postulate (as defined in the text), evolutionary conservation of all seven at P2 should be observed. If all of these residues are not conserved, one of the two assumptions must be incorrect; either not all seven residues are important or the comparative genomics postulate holds, and only two of the seven residues are of functional importance as N termini, then the other five residues are inconsequential players and only these two residues should be evolutionarily conserved.

B, Given that the comparative genomics postulate holds, and only two of the seven residues are of functional importance as N termini, then the other five residues are inconsequential players and only these two residues should be evolutionarily conserved.
N-terminal Protein Processing

parative genomics postulate (function suggests conservation) leads to the conclusion that the seven residues should be evolutionarily conserved at position P2 of proteins. However, because only two out of the seven residues are conserved, we argue that one of the two assumptions in Fig. 1A must be incorrect and put forth the alternative logic depicted in Fig. 1B, which matches our analysis across dozens of species. According to this logic, NME accomplishes the goal of exposing Ala and Ser by exposing all residues with side chains smaller or comparable in size to Ala and Ser (G, T, V, P, and C). These residues are thus inconsequential players that are not functionally important (and are not evolutionarily conserved) at P2.

In this report, we examine the connection between the specificity of NME and stabilizing residues of NER. In doing so, data sets from bacteria (including 112 million mass spectrometric spectra from 57 species), yeast, and mammals, were analyzed for N-terminal peptides both with respect to the excision (or not) of initiator Met residues and the distribution of P2-residues. The results reveal a strong preference of Ala and Ser as P2-residues. However, this process does not appear to be linked to the NER other than being generally compatible with it. These studies also demonstrate a much greater than expected number of Nα-acetylation events in some bacteria.

MATERIALS AND METHODS

Spectral Data Sets—A tandem mass spectrometry (MS/MS) data set from Saccharomyces cerevisiae was used in the analysis of NME and other post-translational modifications (PTMs). Six million spectra were analyzed, searched against an S. cerevisiae database, yielding 410 protein identifications with N-terminal evidence.

In addition, 112 million spectra from 57 different bacterial organisms were searched to provide a large data set containing 16511 classifiable proteins. InsPecT (21) was used for searching the spectra followed by the MS-generating function approach (MSGF), as detailed in Kim et al. (22) to filter for significant identifications at a false discovery rate (FDR) of 1%. Further description of this data set is presented in supplemental Text S1.

N-terminal PTM Search—Of the available 57 data sets from bacterial organisms, 47 were searched for all N-terminal modifications (including unexpected ones) across 68 million spectra. To identify PTMs occurring on the N terminus of proteins, MSGF (22) was used. Given a protein p in proteome P, the first tryptic cut of p was taken as the peptide to search. Therefore, the N terminus of the peptide was the N terminus of p, whereas the C terminus was either K or R. In addition to searching for this peptide, the initial Met was removed and the Met-less version was also searched. So should P denote the number of proteins in the proteome P, there are \( M = 2^p \) peptides to search.

Each of the \( M \) peptides was paired with each available spectrum for the organism in question. Therefore, with \( N \) different spectra, one obtained \( MN \) potential peptide-spectra matches (PSM) to test. For each PSM, there will likely be a difference in parent mass of the spectra and of the peptide. This difference, or offset, was concatenated to the peptide annotation as an N-terminal modification to be searched. As a result our search represents a blind (or universal) search for PTM (23) that allows us to find unexpected PTMs.

For most bacteria, \( M \) is between 1000 and 8000, whereas \( N \) varies greatly. However, there are frequently 40 million to 400 million PSMs with accompanying spectral probabilities. Only those PSMs with spectral probabilities less than \( 1.0 \times 10^{-10} \) were retained as statistically significant. This greatly reduced the number of PSMs, and potential PTMs, to consider.

A permutation test was used to determine over-representation of Nα-acetylated Ser-proteins from the 248 proteins found by MS/MS data in yeast containing evidence of N-terminal acetylation. This test was performed by considering all proteins in S. cerevisiae with Ala, Ser, or Thr at position 2. Of all proteins satisfying that constraint, 248 proteins were randomly selected. Counts for the number of proteins with Ala, Ser, and Thr at P2 were determined for each iteration. This process of randomized sampling was iterated 10,000 times to obtain a null distribution. Thresholds for 5 and 1% were then determined using the conservative Bonferroni correction for the tests of the three residues.

Sequelogs—To test if NME was critical for a function of a specific protein, its sequelogs from related species must be analyzed (in this report the term sequelog from (24) is used instead of the commonly used term ortholog). The protein sequence data used for comparing sequelogs were compiled into three sets: bacteria, yeast, and mammal. The bacterial data set comprises 19 Shewanella species (1860 proteins had sequelogs in all species). The yeast data set comprised seven Saccharomyces species (1502 proteins had sequelogs in all species) obtained from the Saccharomyces Genome Database (http://downloads.yeastgenome.org/). The mammalian data set was composed of six species: human, chimpanzee, macaque, cow, opossum, and rat, obtained from the MSOAR project (http://msoar.ucr.edu/). There were 7934 proteins that had sequelogs in all species. Sequelog mappings for each data set were obtained from each respective project. A protein was included in a particular data set if there existed a sequelog in all species considered in the data set in question. No further annotations or processing were performed to determine or modify the provided sequelog mapping.

RESULTS

NME Specificity—Large-scale mass spectrometric studies (25) have confirmed that NME occurs with nearly 100% efficiency from proteins containing Gly, Ala, Ser, Thr, Cys, Pro, and Val as the P2-residues, which provides the simplest rule for predicting NME. More detailed studies (3, 7) have devised rules for NME prediction that use the amino acids in positions 2 and 3. The simple rule (that classifies a protein as NME-positive if P2 is in X where X = (G, A, P, S, T, V, C)) results in an 8% error rate, most of which represents false positives (96 of 107 errors are false positives). The accuracy of this simple prediction rule on the Shewanella MS/MS data set is acceptable, given that some of the errors may reflect false positive peptide identifications rather than a shortcoming of the rule.

As part of these analyses, an MS/MS data set consisting of ~112 million spectra from 57 different bacteria generated at Pacific Northwest National Laboratory were examined (see (26) for a detailed description of this data set). 6811 N-terminal peptides representing Met retention and 9700 representing Met removal (16511 peptides in total) were identified. This is the largest data set of N-terminal peptides analyzed to date. Fig. 2A and 2B represent the logos for peptides with retained and cleaved Met, respectively. supplemental Table S1, S2, and S3,
Interestingly, these analyses also revealed previously unknown variability in NME specificity across different species. An example of a species with a less specific NME (with eight rather than seven residues in the second position that trigger NME) is shown in Fig. 3, which compares NME specificities in *Brachybacterium faecium*, *Escherichia coli*, and *Deinococcus radiodurans*. Asn is exposed in 20 of 22 cases, in contrast to the standard NME specificity that conventionally includes this residue in the group of non-MetAP substrates. However, it is consistent with the observations of Walker and Bradshaw (19) that yeast MetAP1 could readily cleave a synthetic peptide with a Met-Asn N terminus. *Deinococcus radiodurans* and various *Geobacter* species show a different variance where, contrary to the usual MetAP substrate profile, NME retains rather than removes Met before Thr. This is unlikely to be an artifact because the same effect is observed in all *Geobacter* species in this data set, which includes *Geobacter metallireducens*, *Geobacter sulfurreducens*, and *Geobacter uranireducens*. Another smaller but significant variation in NME specificity relates to Val P2-residues where Met is mainly retained rather than excised in ~30% of the species considered (cleavage specificity is less than 50%). Therefore, although the simple MetAP specificity is a good overall predictor for most species, there is some flexibility observed in some species. This poses an opportunity for NME predictor software, such as the TERMINATOR web server (27), to improve and reflect these variations in NME specificity (currently all eubacteria are assumed to have the same NME specificity).

N-terminal Modifications—A typical MS/MS experiment identifies only a fraction of all tryptic peptides (and hence only a fraction of those derived from the N terminus) and the accuracy of MS/MS-based peptide identification tools further.
deteriorates while detecting modified (e.g. acetylated) peptides (see (28) for complications in identification of N-acetylated peptides). Therefore, the acetylation status of most proteins cannot be inferred from a typical MS/MS experiment. Mass spectrometry data for S. cerevisiae were searched for PTMs (Fig. 4). Nα-acetylation is a common PTM in yeast proteins, with Ser, Ala, Gly, and Thr being the acetylated residues when the initiating Met is cleaved (29). Indeed, of the

Fig. 4. Results from PTM searches in Saccharomyces cerevisiae. A, Breakdown of NME, non-NME, and NME+NTA (NME and Nα-acetylation) events. Approximately 65% of N-terminal events in yeast were found to be Nα-acetylations. B, Residue breakdown of the P2 position of proteins that underwent NME+NTA. Ser is strongly favored in yeast, much as Ser-conservation is also pronounced in yeast. C, 221 identifications of only NME+NTA proteins show that Ser dominates the composition. D, The intersection between NME+NTA and NME-only show a wide distribution among NME residues.
270 Nα-acetylated sites found in *S. cerevisiae* by mass spectrometry, 197 are on Ser, 31 on Ala, and 20 on Thr. Permutation tests revealed that only Ser is significantly over-represented in these PTM observations, the only residue with a p-value less than 0.01 (see Methods). In addition, Fig. 4D shows that Nα-acetylation prefers Ser-proteins because Thr- and Ala-proteins increase in frequency when considering the intersection of Nα-acetylation and NME-only events. Although proteins with N-terminal Ser, Ala, Thr, and Gly are all Nα-acetylated in yeast, N-terminal Ser-proteins appear to be heavily favored.

Serine is favored in *S. cerevisiae*, however, it is not universally favored for Nα-acetylation events. Helbig et al. (28) provide mass spectrometry evidence for Nα-acetylation in human cells, showing a strong preference for Ala. In the same study, the Nα-acetylation profile for *Drosophila melanogaster* was also shown to differ, with Met, Ala, and Ser nearly equally represented. The bacterial Nα-acetylation data in Table I and Fig. 5 also show a slightly different profile, still with Ala and Ser as prominent residues.

**Widespread N-terminal Acetylation in Bacteria**—Acetylation is currently viewed as a rare modification in bacteria. In previous studies with *E. coli* five different Nα-acetylated proteins were identified: ribosomal subunits S18 and S5 (30), L12 (31), along with SecB (32), and elongation factor Tu (33). Utilizing the data set of bacterial organisms, blind PTM searches detected many Nα-acetylated proteins in multiple organisms, although certainly not all.
Fig. 6 shows the potential PTMs as the number of occurrences of a particular mass shift. Mass shifts of -131 correspond to a loss of Met (i.e., an NME event) whereas a shift of 0 signifies the retention of the initial Met residue; these are expected events. The surprising number of -89 mass shifts, corresponding to a loss of Met (-131) along with the addition of an acetyl group (+42), was considered indicative of Nα-acetylation. A protein was considered to be Nα-acetylated if it was observed with this offset within ±1Da.

Fig. 5A diagrams the findings of NME-only, non-NME, and NME proteins that also undergo Nα-acetylation (NME+NTA), using this un-targeted PTM search method. There is considerable overlap between NME-only and NME+NTA sets, suggesting that many proteins undergoing NTA do not require this modification, or require both (as with ribosomal protein L7/L12 in E. coli). Fig. 5 summarizes the overall P2 residue distribution of NME events and NME+NTA events from Table II and Table I, respectively. It was observed that Ser is well
represented in both groups, showing that it is an important component of NME. Table I shows the breakdown of amino acids at position P2 of each protein with observed NME and NTA. Ala, Ser, and Thr showed a strong tendency for losing the initial Met and for being acetylated. Table II shows the same amino acid breakdown for P2 of identified NME events using this same MSGF approach.

Although Thr is highly represented in these N/H9251-acetylated identifications, seen in Fig. 5C, no such elevated Thr-conservation at position P2 was observed in the Shewanella data set. This could be because of a more complicated role of Thr in N/H9251-acetylated proteins, or a reduced role of NTA in the Shewanella organisms compared with other bacteria (data not shown). Unfortunately, sequelogs are difficult to determine for distantly related bacterial organisms, preventing further analysis of conservation in bacteria.

NME-conserved Proteins—To observe NME importance using comparative genomic approaches, we made certain assumptions about the sequence data.

Postulate 1 (Comparative genomics postulate): If a protein undergoes NME in one species and it is essential for its function, then related proteins in other species are likely to undergo NME as well. A protein is called (strongly) NME-conserved if its related proteins (sequelogs) across all studied species also undergo NME. Alternatively, a protein can be called mostly NME-conserved if most of its sequelogs (statistically significant increase as compared with null hypothesis) also undergo NME. This fuzzier definition would also capture functional importance regarding NME, but would add additional complexity to our analysis. We opt for a strict definition of (strongly) NME-conserved through the remainder of the text.

Postulate 2: Finding X-residues in all sequelogs of a protein is statistically surprising unless NME is truly important for the function of the protein. Although similar to the first axiom, it is distinct from it because the first one specifies a general relation among sequelogs, and this one relates X-residues to NME and describes the type of signal we should observe from proteins that require NME.

Postulate 3: The baseline compares X-residues at P2 with X-residues at P3, P4, and P5. Because X-residues at P2 are related to NME, and presumably only NME, those at P3, P4, and P5 are not related to NME but are instead related to each individual protein’s function. These three positions can be used as a baseline for the background of expected X-residue statistics.

Our statistic for X-residues is now defined as follows: For a protein P (having sequelogs in all species) and a position i, define \( X(P,i) \) as the number of species in which the amino acid at position i is an X-residue. \( X(P,i) \) can vary from 0 to the number of species, n. The number of proteins with \( X(P,i) = t \) defines count(t,i) (count(t,i) varies from 0 to the number of proteins, k). Because we care about proteins with X-residues at P2 across all n species, as per postulates 1 and 2, the statistics we compute are count(n,2), count(n,3), count(n,4) and count(n,5). To obtain an estimate on each statistic’s variance, bootstrap analysis is applied to each data set using 10,000 iterations.

The mean counts for each data set are shown in Fig. 7A, for the Shewanella data set there are 375 such proteins for position 2, in contrast to only 162 for position 3. This is a significant increase in conservation of X-residues at the position directly affecting NME. Similar trends are observed in yeast and mammalian species, also seen in Fig. 7A. Yeast shows a similar pattern with \( \sim 300 \) more proteins with perfectly conserved X-residues at position 2 as compared with other positions. The mammalian species also show a large difference (\( \sim 500 \) proteins) between position 2 and other

Fig. 6. Offset counts for the range of \(-150 \) to \(+150 \) in A. variabilis. NME events are represented as peaks located at \(-131 \), NME and N\( \alpha \)-acetylation at \(-89 \). Retention of Met (no NME event) is represented by a peak at 0.
positions. Furthermore, the differences in X-residue counts between positions 2 compared with 3, 4, and 5 cannot be attributed to variance in the statistic used.

The higher conservation of X-residues at the second position, seen in all data sets in Fig. 7A, indicates that NME may be functionally critical for some proteins and provides a candidate list of such proteins. Proteins whose sequelogs in all species have an X-residue in the second position are designated as NME-conserved proteins.

Conservation of Ala- and Ser- as the P2-Residue in Proteins—After analyzing for group conservation of all seven X-residues at initial positions of the proteins (checking whether the second residue in each sequelog is an X-residue), analyses of individual amino acids (instead of the set X) at the second position and comparison to the conservation at the third position (where conservation is not expected) as a control were performed. Fig. 8 and Table III show that Ala is conserved among all Shewanella species in 79 proteins at position 2 but only in 14 proteins at position 3. Ser is conserved in 75 proteins at position 2 but only in 16 proteins at position 3. Conservation levels of other amino acids are rather similar between positions 2 and 3, suggesting that proteins with Ala and Ser in the P2 position may be the important targets for NME in Shewanella. Because Ala and/or Ser are exceptionally well-conserved, it is reasonable to assume that NME affects the function of these proteins (if one accepts the comparative genomics postulate). This Ala and/or Ser importance is further underlined in supplemental Text S1, specifically Tables S4, S5, and S6, where T-tests showed significant differences in P2 from other N-terminal positions.

Ala and Ser also exhibit much higher conservation at the second position as compared with other N-terminal positions.
indicating that they are specifically required here (rather than in a few initial positions) and suggests that this elevated conservation is relevant to NME. *Shewanella*, yeast, and mammalian species all show much higher levels of conservation for Ala and/or Ser in the second position compared with other residues. The one exception to this pattern is seen for Lys in *Shewanella*, Fig. 8A, where Lys is elevated at P2 and P3 compared with P4 and P5. This violates our third postulate, and is therefore considered to not be an NME relevant pattern.

This trend of Ala/Ser importance appears again in Fig. 7B, which displays X-residue conservation across bacterial, yeast, and mammalian data sets. For the data in this figure, Ala and Ser are removed from the X set, i.e. $X = \{G,P,T,C,V\}$ and there is little change in the level of conservation from P2 to P3, P4 or P5 for sequlogos within each data set. Most of the change from P2 to P3, P4 or P5, of which there is little, can be accounted for by variability of the statistic. Comparing Fig. 7A to Fig. 7B shows that Ala and Ser are responsible for the elevated levels of conservation at position 2 for these sequlogos.

**DISCUSSION**

Although it has been known for some time that protein synthesis is essentially universally initiated with Met (34), it has only become appreciated over the intervening years that this residue is subsequently removed from a considerable portion of the proteins actively synthesized in any given organism. Much of this NME for cytoplasmic and nuclear proteins occurs as a cotranslational event catalyzed by specific MetAPs (5). There are two classes of these enzymes but both have quite similar substrate profiles (11). In a eukaryote such

---

**Fig. 7.** Mean NME-conserved protein counts are shown for P2 through P5 for bacterial, yeast, and mammalian data sets. An NME-conserved protein is one containing the residue at the position in question (two through five) in $X$ across all sequlogos of that protein. All values are obtained from bootstrap estimated mean and standard deviation of each statistic. A The set of residues considered is defined as $X = \{G,A,S,P,V,T,C\}$. It is evident that P2 contains a larger number of NME-conserved than P3-P5. B NME-conserved protein counts, without Ala and Ser, are shown for P2 through P5 for the three data sets. Here the X set does not contain Ala and Ser, and is defined as $X = \{G,P,V,T,C\}$. Not including Ala and Ser in $X$ produces no elevated counts at P2 compared with P3, P4, and P5. This suggests that the boost in NME-conserved counts seen in A can be attributed to the Ala and Ser residues.
Fig. 8. Conservation of residues are shown for P2 through P5 for data sets from A, Shewanella B, Saccharomyces C, mammalian. Ala and Ser are by far the most conserved in P2 across all three data sets. The number of proteins with Ala in the second position in Shewanella, Saccharomyces, and the mammalian data sets are 79, 24, and 541, respectively. The number of proteins with Ser in P2 in Shewanella, Saccharomyces, and mammalian data sets are 75, 130, and 268, respectively.
as yeast, the number of proteins predicted to lose or retain the initiator Met (on an individual gene basis) is nearly equal; however, based on protein mass, a great preponderance of soluble proteins undergo this modification and are, moreover, then N-acetylated.

The functional purposes of NME are poorly understood and there are several possible reasons that NME is such an essential process. In the first place, it has been argued that NME allows the rapid recycling of Met to keep the levels of the intracellular pools at an adequate level for the many roles that this amino acid plays in addition to its involvement in protein synthesis (both as an initiator and as a constituent residue). Met is generally relatively scarce and the recovery of most (on a mass basis) of the Met expended in initiating protein synthesis is certainly likely to be beneficial to the cell. However, this ascribed role for NME begs the question of why Met is recovered from only a subset and not all proteins. The answer, although likely complex, may be that having two classes of protein (those that lose their Met and those that do not) is desirable for other reasons. Indeed, it is possible that primordial MetAPs did not make the distinction between these two substrate types and Met was recovered from all synthesized proteins. However, if this was the case, evolution altered these enzymes to their present day specificity very early on and these changes were adopted in all life forms that are presently known. There are few biological processes that show this level of conservation in all living organisms. Walker and Bradshaw (19) demonstrated that NME specificity in vitro can be changed (by site-directed mutagenesis of yeast MetAP1) to expand the number of P2-residues that allowed Met removal, so even within the present MetAPs there is no structural reason to prevent the development of an enzyme (through mutation and natural selection) that has a broader and possibly comprehensive substrate profile. The fact that this did not happen suggests additional roles for NME and the need for the two classes of protein that it generates.

By the same token, being able to expand the specificity does not ensure that it would be possible to further "close down" the repertoire of substrates. It may well be that the substrate specificity that the present day MetAPs display is basically at the lower limit (in terms of the size of the P2-residue) that nature can engineer with this scaffold. This may explain the dominance of Ala/Ser substrates (controlled by evolutionary pressure on the second position) while still allowing the five or so additional substrates that are also cleaved.

A second proposed function of NME is linked to protein stability and degradation. Thus, the profile of NME produced by MetAP action fits, at least in eukaryotes, to that dictated by the NER. This suggests, given the very early development of MetAP specificity, that the recognition component of the NER that binds certain N-terminal residues to allow the polyubiquitinylation and subsequent proteasomal degradation evolved to be compatible with it, rather than the other way around. As such, the removal of Met (from the seven amino acids with the smallest side chains) is not so much a stabilizing event as it is a permissive one, i.e. these N termini are simply ignored by the NER machinery. Because there is a singular lack of evidence that the proteins in the non-NME class are ever subsequently degraded via the NER (after exposure downstream of the destabilizing residues), it suggests that the role of NER is to capture and destroy misfolded and/or damaged proteins, spuriously formed fragments and any other peptidic detritus that may arise in the course of other cellular activities.

### Table III
**Conservation of single amino acids for the Shewanella sequence data set (P2 and P3)**

| Amino acid | P2 | P3 |
|------------|----|----|
| G          | 11 | 11 |
| A          | 79 | 14 |
| S          | 75 | 16 |
| P          | 16 | 17 |
| V          | 6  | 25 |
| T          | 21 | 20 |
| C          | 3  | 2  |
| L          | 39 | 52 |
| I          | 28 | 45 |
| N          | 21 | 16 |
| D          | 13 | 19 |
| Q          | 20 | 20 |
| K          | 58 | 55 |
| E          | 19 | 21 |
| M          | 1  | 1  |
| H          | 6  | 8  |
| F          | 14 | 27 |
| R          | 23 | 28 |
| Y          | 7  | 17 |
| W          | 3  | 4  |

### Table IV
**Conservation of single amino acids for the Saccharomyces sequence data set (P2 and P3)**

| Amino acid | P2 | P3 |
|------------|----|----|
| G          | 18 | 13 |
| A          | 24 | 15 |
| S          | 130| 13 |
| P          | 15 | 9  |
| V          | 14 | 8  |
| T          | 10 | 10 |
| C          | 1  | 0  |
| L          | 25 | 18 |
| I          | 1  | 12 |
| N          | 11 | 9  |
| D          | 13 | 6  |
| Q          | 3  | 10 |
| K          | 14 | 19 |
| E          | 3  | 4  |
| M          | 0  | 0  |
| H          | 2  | 4  |
| F          | 11 | 14 |
| R          | 7  | 33 |
| Y          | 1  | 10 |
| W          | 1  | 2  |
Asp, Glu, or Asn (15). Hwang addition, the initiator Met is acetylated when the P2-residue is residues exposed by NME but is certainly not the only one. In acetylation is a prominent modification of four of the seven residues of this data set (P2 and P3).

### Table V

| Amino acid | P2  | P3  |
|------------|-----|-----|
| G          | 168 | 119 |
| A          | 541 | 186 |
| S          | 268 | 190 |
| P          | 116 | 81  |
| V          | 60  | 64  |
| T          | 90  | 104 |
| C          | 15  | 29  |
| L          | 94  | 171 |
| I          | 24  | 25  |
| N          | 53  | 59  |
| D          | 97  | 93  |
| Q          | 28  | 55  |
| K          | 67  | 97  |
| E          | 177 | 138 |
| M          | 25  | 30  |
| H          | 18  | 17  |
| F          | 39  | 49  |
| R          | 45  | 121 |
| Y          | 19  | 24  |
| W          | 19  | 29  |

rather than as part of programmed protein degradation. Recently, Hwang et al. (16) have revised the original NER by showing that Nα-acetylation can produce N termini that are also recognized by the germine ubiquitin ligase and thus arguing now that NER is really a quality control mechanism for protein stoichiometry in the cell. They also conclude that it mainly functions as a scavenger for improperly folded proteins.

A third raison d’être for NME is the preparation of the N terminus for further post-translational modification. For purposes of this discussion, no strong distinction is drawn between co- and post-translational events. In eukaryotes, Nα-acetylation is a prominent modification of four of the seven residues exposed by NME but is certainly not the only one. In addition, the initiator Met is acetylated when the P2-residue is Asp, Glu, or Asn (15). Hwang et al. (16) have recently reported that five residues (Ala, Val, Ser, Thr, and Cys) are the primary sites of NTA after NME and argue that they should now be considered destabilizing residues in the NER. However, in an extensive study of Nα-acetylation by mass spectrometry, it was shown that Ala and Ser account for 89% of all the proteins modified by acetylation that are NME targets in Homo sapiens, implying that Nα-acetylation of Val, Cys, and even Thr is also rare in eukaryotes (28). The NATs responsible for NTA (the NatA, NatB, and NatC complexes) have been studied in a variety of eukaryotic organisms (35). NatA modifies Ala, Ser, Gly, and Thr so it requires NME to prepare its substrates. Knockdown of Nna10 (previously Ard1, standardization introduced (35)) and Naa15 (previously Nat1), which comprise NatA, in Caenorhabditis elegans resulted in embryonic lethality (36). The human NatA complex has also been shown to be important in apoptosis as shown in Naa10 (previously hARD1) and Naa15 (previously NATH) knockdowns in human HeLa cells (37). The necessity of NatA in yeast is less clear-cut. Polevoda et al. (15) produced deletion mutants of Naa10/Naa15 (previously ard1 and nat1, of the NatA complex), Naa20 (previously Nat3, of the NatB complex), and Naa30 (previously Mak3, of the NatC complex), all of which are viable; however, all display negative phenotypes.

The extensive occurrence of this reaction in eukaryotes suggests that it is an important modification. However, there is little evidence for any direct role of this group in any function. It may be that it is involved in, as yet unidentified, protein-protein interactions perhaps through a specific binding motif such as found for other PTMs. Alternatively it may contribute generally to the cellular environment by blocking a significant number of α-amino groups to prevent nonenzymatic glycation with reducing sugar metabolites (which would make these energy rich compounds unavailable for catabolic breakdown). It seems not to be significantly involved, as was once proposed, in stabilizing proteins against spurious aminopeptidase attack.

In contrast, modifications such as N-myristoylation clearly have functional consequences. This modification has been shown to be important for a small subset of proteins, and it is estimated that 0.5% of eukaryotic proteins undergo this modification (38). The functional role of NME is unlikely to be strongly connected to N-myristoylation as only a small percentage of proteins undergo this addition and even fewer require it for proper protein function. Also, N-myristoylation is not as universal as NME because it does not occur in bacteria. The analyses herein (Table V) reveal some over-conservation of Gly in mammals (168 Gly-conserved proteins for the second position as compared with 119 Gly-conserved proteins for the third position). However, this 40% increase provides a somewhat weaker support for the functional role of Gly as compared with 190% increase for Ala. Therefore, although Gly-conserved proteins undoubtedly exist, their modest amount of over-conservation does not allow one to conclude that exposing Gly is an important role of NME.

The inclusion of the commonly acetylated N termini in the NER (16) significantly changes the original model. The N termini now considered to be degrons compose 18 of 20 residues, including 5 of 7 X-residues: A, V, S, T, and C, leaving out only X-residues G and P because of the fact that they are rarely Nα-acetylated (in yeast). We argue that if G and P are not considered degrons, then V, T, and C should also be removed from the degron list because of their rare acetylation state, (seen in Fig. 4). This leaves Ala and Ser as degrons that are heavily Nα-acetylated in yeast, suggesting that the predominant removal of Met and subsequent acetylation of Ala and Ser combine to provide a principal role for NME and the connection to the NER degradation pathway.

One explanation for the conservation of Ala and Ser might be the influence of the underlying nucleotide sequences of the
Met and second codon. The Kozak consensus RccATGG is a regulatory motif playing an important role in the initiation of translation in vertebrate and some other eukaryotes (39). Because the Kozak consensus (where ATG represents the first codon in the protein) suggests that the second codon starts with G, it might be expected that there would be higher levels of Ala at P2 in vertebrates because of the frequency of G at the first position of Ala codons. However, by using much larger data sets than available to Kozak at the time the rule was formulated, Xia (40) showed that, contrary to the Kozak rule, there is no link between G at the fourth position and translation initiation efficiency. Xia further suggested that it is the presence of specific amino acids at P2 that led Kozak to (erroneously) conclude that G at position 4 affects the translation efficiency. The observations reported herein that Ala and/or Ser are universally over-represented in the second position in nearly all species cannot be explained by the Kozak consensus rules because it is not universal, e.g., it is not applicable to bacteria and does not extend beyond the first codon in fruit flies. Yeast consensus codons UC[U/C] as reported in Hamilton et al. (41) also appear not to be in support of translation initiation efficiency (shown in supplemental Text S2, specifically Fig. S1, and S2).

It is appropriate to emphasize that the goal of this study was to detect whatever statistically significant differences might occur between the second and third positions that could offer insight into NME function. Indeed, the fact that Ser/Ala have elevated levels of conservation at P2 in species as diverse as bacteria and humans illustrates that there exists an enormous evolutionary pressure for retaining them in the second position. Because this analysis reveals no over-conservation for 18 out of 20 amino acids, it can be argued that exposure of Ala and/or Ser by NME defines its key function.

Although this study reveals only two over-conserved residues in the second position in the species studied, it cannot be ruled out that additional residues (e.g. Thr) may be over-conserved in other species. However, the seven seemingly equivalent residues (with respect to NME specificity) clearly have vastly different evolutionary patterns suggesting that some of them (e.g. Ala and Ser) play an important functional role whereas others may represent inconsequential players. Thus, NME appears to be a mechanism evolved to resolve the “conflict” between the uniformity of the translation initiation mechanism and the variety of functions affecting the N termini of proteins.

** To whom correspondence should be addressed: Computer Science and Engineering, University of California San Diego, La Jolla, CA. Tel.: 858-822-4365; E-mail: ppevzner@ucsd.edu.

† Current affiliation: National Institutes of Health, Bethesda, Maryland. E-mail: sboniss@ucsd.edu.

**REFERENCES**

1. Giglione, C., Boularat, A., and Meinnel, T. (2004) Protein N-terminal methionine excision. Cell. Mol. Life Sci. 61, 1455–1474
2. Arfin, S., and Bradshaw, R. (1988) Cotranslational processing and protein turnover in eukaryotic cells. Biochemistry 27, 7979–7984
3. Frottin, F., Martinez, A., Peynot, P., Mitra, S., Holz, R. C., Giglione, C., and Meinnel, T. (2006) The Proteomics of N-terminal Methionine Cleavage. Mol. Cell. Proteomics 5, 2336–2349
4. Meinnel, T., Serero, A., and Giglione, C. (2008) Impact of the N-terminal amino acid on targeted protein degradation. Biol. Chem. 389, 839–851
5. Bradshaw, R. A., Brickey, W. W., and Walker, K. W. (1998) N-Terminal processing: the methionine aminopeptidase and N[alpha]-acetyl transferase families. Trends Biochem. Sci. 23, 263–267
6. Huang, S., Elliott, R. C., Lui, P. S., Koduri, R. K., Weickmann, J. L., Lee, J. H., Blair, L. C., Ghosh-Dastidar, P., Bradshaw, R. A., and Bryan, K. M. (1987) Specificity of cotranslational amino-terminal processing of proteins in yeast. Biochemistry 26, 8242–8246
7. Meinnel, T., and Giglione, C. (2008) Tools for analyzing and predicting N-terminal protein modifications. Proteomics 8, 626–649
8. Levitt, M. (1976) A simplified representation of protein conformations for rapid simulation of protein folding. J. Mol. Biol. 104, 59–107
9. Hutchison, C. A., Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., Smith, H. O., and Venter, J. C. (1999) Global Transposon Mutagenesis and a Minimal Mycoplasma Genome. Science 286, 2165–2169
10. Li, X., and Chang, Y. H. (1999) Amino-terminal protein processing in Saccharomyces cerevisiae is an essential function that requires two distinct methionine aminopeptidases. Proc. Natl. Acad. Sci. U.S.A. 92, 12357–12361
11. Arfin, S. M., Kendall, R. L., Hall, L., Weaver, L. H., Stewart, A. E., Matthews, B. W., and Bradshaw, R. A. (1995) Eukaryotic methionyl aminopeptidases: two classes of cobalt-dependent enzymes. Proc. Natl. Acad. Sci. U.S.A. 92, 7714–7718
12. Bachmair, A., Finley, D., and Varshavsky, A. (1986) In vivo half-life of a protein is a function of its amino-terminal residue. Science 234, 179–186
13. Varshavsky, A. (1996) The N-end rule: functions, mysteries, uses. Proc. Natl. Acad. Sci. U.S.A. 93, 12142–12149
14. Persson, B., Flinta, C., Heijne, G. von, and Jörnvall, H. (1985). Structures of N-terminally acetylated proteins. Eur. J. Biochem. 152, 523–527
15. Polevoda, B., Norbeck, J., Takakura, H., Blomberg, A., and Sherman, F. (1999) Identification and specificities of N-terminal acetylated transferases from Saccharomyces cerevisiae. EMBO J. 18, 6155–6168
16. Hwang, C. S., Shenmory, A., and Varshavsky, A. (2010) N-Terminal acetylation of cellular proteins creates specific degradation signals. Science 327, 973–977
17. Tobias, J. W., Shradrer, T. E., Rocap, G., and Varshavsky, A. (1991) The N-end rule in bacteria. Science 254, 1374–1377
18. Polevoda, F. S. B. (2002) The diversity of acetylated proteins. Genome Biol. 3, 59–107
19. Walker, K. W., and Bradshaw, R. A. (1999) Yeast Methionine Aminopeptidase I. J. Biol. Chem. 274, 13403–13409
20. Giglione, C., Vallon, O., and Meinnel, T. (2003) Control of protein life-span by N-terminal methionine excision. EMBO J. 22, 13–23
21. Tanner, S., Shu, H., Frank, A., Wang, L. C., Zandi, E., Mumba, M., Pevzner, P. A., and Bafna, V. (2005) InsPeCt: identification of posttranslationally modified peptides from tandem mass spectra. Anal. Chem. 77, 4626–4639
22. Kim, S., Gupta, N., and Pevzner, P. A. (2008) Spectral Probabilities and Generating Functions of Tandem Mass Spectra: a Strike Against Decoy Databases. J. Proteome Res. 7, 3354–3363
23. Bauer, D., Tanner, S., Zandi, E., Bafna, V., and Pevzner, P. A. (2005) Identification of Post-translational Modifications via Blind Search of Mass Spectra. Nat. Biotechnol. 23, 1562–2567
24. Varshavsky, A. (2004) ‘Spalog’ and ‘sequelog’: neutral terms for spatial and sequence similarity. Current Biol. 14, R181–R183
25. Gupta, N., Tanner, S., Jaitly, N., Adkins, J. N., Lipton, M., Edwards, R.,
N-terminal Protein Processing

Romine, M., Osterman, A., Bafna, V., Smith, R. D., and Pezrner, P. A. (2007) Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation. Genome Res. 17, 1362–1377

26. Frank, A. M., Monroe, M. E., Shah, A. R., Moore, R. J., Anderson, G. R., Smith, R. D., and Pezrner, P. A. (2010). Spectral Archives: A Novel Approach To Analyzing Tandem Mass Spectra. Submitted

27. Martinez, A., Traverso, J. A., Valot, B., Ferro, M., Espagne, C., Ephritikhine, G., Zivy, M., Giglione, C., and Meinnel, T. (2008) Extent of N-terminal modifications in cytosolic proteins from eukaryotes. Proteomics 8, 2809–2831

28. Helbig, A. O., Gauci, S., Rajmakers, R., Breukelen, B. van, Slijper, M., Mohammed, S., and Heck, A. J. (2010) Profiling of N-Acetylated Protein Termini Provides In-depth Insights into the N-terminal Nature of the Proteome. Mol. Cell. Proteomics 9, 928–939

29. Walsh, C. (2005). Posttranslational Modification of Proteins. Expanding Nature’s Inventory. Englewood, CO: Roberts and Co. Publishers

30. Yoshikawa, A., Isono, S., Sheback, A., and Isono, K. Cloning and nucleotide sequencing of the genes rimI and rimJ which encode enzymes acetylat- ing ribosomal proteins S18 and S5 of Escherichia coli K12. Mol. General Gen. 209, 481–488

31. Tanaka, S., Matsushita, Y., Yoshikawa, A., and Isono, K. (1989). Cloning and molecular characterization of the gene rimI which encodes an enzyme acetylating ribosomal protein L12 of Escherichia coli K12. Mol. General Gen. 289–293

32. Smith, V. F., Schwartz, B. L., Randall, L. L., and Smith, R. D. (1996) Electrospray mass spectrometric investigation of the chaperone SecB. Protein Sci. 5, 488–494

33. Arai, K., Clark, B. F., Duffy, L., Jones, M. D., Kaziro, Y., Laursen, R. A., L’Italien, J., Miller, D. L., Nagarkatti, S., Nakamura, S., Nielsen, K. M., Petersen, T. E., Takahashi, K., and Wade, M. (1980) Primary structure of elongation factor Tu from Escherichia coli. Proc. Natl. Acad. Sci. U.S.A. 77, 1326–1330

34. Jackson, R., and Hunter, T. (1970) Role of Methionine in the Initiation of Haemoglobin Synthesis. Nature 227, 672–676

35. Polevoda, B., Arnesen, T., and Sherman, F. (2009) A synopsis of eukaryotic Nalpha-terminal acetyltransferases: nomenclature, subunits and substrates. BMC Proceedings 3, S2

36. Sönntichsen, B., Koski, L. B., Walsh, A., Marschall, P., Neumann, B., Brehm, M., Alleaume, A. M., Artelt, J., Bettencourt, P., Cassin, E., Hewitson, M., Holz, C., Khan, M., Lazik, S., Martin, C., Nitzsche, B., Ruer, M., Starnford, J., Winzi, M., Heinikel, R., Roder, M., Finell, J., Häntsch, H., Jones, S. J., Jones, M., Piano, F., Gunsalus, K. C., Oegema, K., Gönzczy, P., Couolson, A., Hyman, A. A., and Echeverri, C. J. (2005) Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. Nature 434, 462–469

37. Arnesen, T., Gromyko, D., Pendino, F., Ryningen, A., Varhaug, J. E., and Lillegaard, J. R. (2006) Induction of apoptosis in human cells by RNAi-mediated knockdown of hARD1 and NATH, components of the protein N-alpha-acetyltransferase complex. Oncogene 25, 4350–4360

38. Maurer-Stroh, S., Eisenhaber, B., and Eisenhaber, F. (2002) N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. J. Mol. Biol. 317, 541–557

39. Kozak, M. (1989) The scanning model for translation: An update. J. Cell Biol. 108, 229–241

40. Xia, X. (2007) The +4G Site in Kozak Consensus Is Not Related to the Efficiency of Translation Initiation. PLoS ONE 2, e188

41. Hamilton, R., Watanabe, C., and Boer, H. de (1987). Compilation and comparison of the sequence context around the AUG startcodons in Saccharomyces cerevisiae mRNAs. Nucleic Acids Res. 15, 3581–3593