Amino acid preferences at neddylation sites

Ahmet Sinan Yavuz, Namık Berk Sözer and Osman Uğur Sezerman*

***Abstract*— Neddylation is a dynamic post-translational modification in which NEDD8 proteins are covalently attached to the target site lysine residue. Neddylation may affect a target protein's localization, binding partners and structure. Targets of this modification have commonly found in nucleus and the most well characterized target family is cullins, which is modulating ubiquitination and proteosomal degradation system in a cell. Disruptions in neddylation pathway implicated in various diseases such as Alzheimer's, Parkinson's and cancer. Therefore, understanding neddylation site recognition bears a huge importance in understanding the complete functional mechanism of this post-translational modification and revealing the mechanisms of associated diseases towards a cure. However, there is no study in literature investigating whether a common neddylation site motif exists or not. In this work, we have identified various amino acid preferences and hydrophobicity patterns seen in neddylation sites, differing from not neddylated lysine residues.**

INTRODUCTION

NEDD8 is an ubiquitin-like modifier, which is encoded by *NEDD8* gene in humans and *Rub1 gene* in *S. cerevisiae*. NEDD8 was initially identified as one of the ten neural precursor cell-expressed, developmentally down regulated genes (NEDD) and defect on NEDD8 pathway shown to be lethal in many organisms [1]. NEDD8 shares ~60% sequence identity with ubiquitin, and it is the most similar known ubiquitin-like protein (Ubl) [1].

Neddylation is the covalent attachment of NEDD8 proteins to the target sites. Similar to SUMO and other Ubl proteins, NEDD8 is synthesized in an immature form [2]. Cleavage of extra amino acids catalyzed by UCH-L3 enzyme located beyond Gly76 reveals the mature isopeptide linkage site, which will form a bond with target site's lysine residue [2]. Following the maturation, NEDD8 can be activated to bind to an E1 enzyme (UBA3-APPBP1 heterodimer) consuming 1 ATP in the process (Figure 1). Afterwards, E1 bound NEDD8 is loaded on an E2 enzyme (UBC12) and from E2, with or without help of an E3 enzyme, it is transferred to the target site's lysine residue [2]. Attached NEDD8 proteins then can be removed by NEDD8 isopeptidases, making neddylation a dynamic and reversible process.

As many other post-translational modifications, neddylation directly affects 3D surface of a target protein, which may alter binding partners of the substrate and/or stimulate a conformational change in the structure [2].

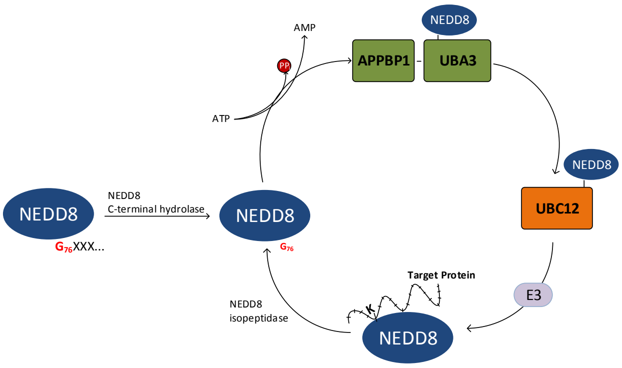

Figure 1. Neddylation pathway. Basic schematic representation of neddylation pathway shows maturation by cleavage of extra amino acid residues after G76, NEDD8 activation by APPBP1-UBA3 heterodimer using an ATP, transfer of NEDD8 to UBC12 (E2), and conjugation of NEDD8 into target site's lysine residue with or without a help of an E3.

Additionally, neddylation can encourage the recruitment of NEDD8-binding proteins, causing new protein complexes to occur [2]. All of these direct effects may also influence further changes such as in subcellular localization [2]. In conclusion, NEDD8 attachment to a substrate may significantly alter target protein's lifespan, role, subcellular localization and structure.

Neddylated proteins can be found mostly in the nucleus and the most well-characterized targets are the cullin proteins [3], [4]. Cullin proteins are scaffold proteins of SCF-ubiquitin ligase complex, which controls the ubiquitination and proteosomal degradation system [4]. Neddylation of cullins increases the ubiquitination and proteosomal degradation of substrates. As main targets of cullins are cell-cycle regulation, transcriptional regulation and signal transduction proteins, neddylation plays a significant role in the maintenance of cell machinery [4]. Hence, disruptions in neddylation pathway has observed in many diseases, such as Alzheimer's [5], Parkinson's [6], and cancer [7].

Although neddylation plays significant roles in cells, target recognition and specificity is still unclear. There is no previously reported neddylation site motifs or published neddylation site prediction tools. Identification of neddylation target sites experimentally is also expensive and laboursome. Therefore, there is a need of identifying possible sequence properties of neddylation sites to aid in prediction of such sites.

In this short work, we aim to identify common seen amino acid preferences or hydrophobicity patterns seen in the neddylation sites.

METHODS

Dataset

We have searched PubMed with keywords "nedd8", "neddylation", "nedylation", "rub1", "rub2", "rub3", and "rubylation", and manually collected 63 sites in 29

*Corresponding Author (e-mail: ugur@sabanciuniv.edu).
Ahmet Sinan Yavuz and Osman Uğur Sezerman are with Biological Sciences and Bioengineering Program, Faculty of Engineering and Natural Sciences, Sabancı University, Orhanli, Tuzla, Istanbul, Turkey (emails: asinanyavuz@sabanciuniv.edu and ugur@sabanciuniv.edu).
Namık Berk Sözer is with Department of Genetics and Bioengineering, Faculty of Engineering and Architecture, Yeditepe University, Istanbul, Turkey, (e-mail: namikberk.sozer@std.yeditepe.edu.tr).

proteins from ~600 articles, published until July 1st, 2014. Among these sites, 6 were discarded due to neddylation was shown only in vitro, and 3 were discarded as neddylation was not reported in a single amino acid resolution. After this elimination, primary sequences of 28 proteins were retrieved from UniProt [8].

Redundancy elimination was performed with CD-HIT [9]. This program clusters sequence datasets and selects a representative sequence of each cluster having at least a given percent identity. We clustered sequences with 0.4 threshold, so that no two sequences sharing a sequence identity >40% left in the dataset. After such an elimination procedure, dataset was left with 22 proteins and 48 sites.

We prepared dataset for analysis by defining sequence windows as lysine residues flanked by 10 residues upstream and 10 residues downstream, forming a 21 amino acid long sequence segments. All sequence windows that contain experimentally identified neddylation sites were considered as the positive set and rest of the sequence windows was assumed as not neddylated and formed the negative set.

Amino Acid Grouping and Hydrophobicity Scale

In order to assess common biochemical properties in the sequence windows we have used both 20-letter amino acid alphabet and a 11-letter grouping of amino acids based on physicochemical properties, named as Sezerman grouping [10], [11] (Table 1).

SEZERMAN AMINO ACID GROUPING

Groups	Amino Acids
A	IVLM
Q	RKH
C	DE
D	QN
E	ST
F	A
G	G
H	W
I	C
W	YF
K	P

Additionally, we have used Kyte & Doolittle [12] hydrophobicity scale to assess the hydropathy difference between neddylated and not neddylated sequence windows.

Statistical Testing

In order to assess statistical difference in hydrophobicity values between known neddylated sequence windows and not neddylated sequence windows, two-tailed Mann-Whitney U tests were performed.

Additionally, amino acid profiles of neddylated and not neddylated sequence windows were compared with chi-square test of independence. Two strategies were employed to identify differences clearly. First one was implemented by creating 20x2 and 11x2 contingency tables for each position in the window, for normal amino acid and Sezerman grouping amino acid distributions, respectively. This approach aims to identify general differences in amino acid distributions. Second strategy was to identify whether particular amino acids are over or underrepresented in particular positions of the sequence windows. For this strategy, we have created twenty 2x2 contingency tables for normal amino acid representation and eleven 2x2 contingency tables for Sezerman grouping.

Benjamini- Hochberg [13] procedure has been applied for controlling false discovery rate at α = 0.05. All p-values have been adjusted according to this procedure.

All statistical tests were performed using R (version 3.1.0, The R Foundation for Statistical Computing, Vienna, Austria; http://www.r-project.org) and an in-house program written in Python 2.7.5 [14], with the SciPy [15] library (version 0.11.0).

RESULTS

Sequence Logos

In order to identify amino acid preferences visually, we have created sequence logos using WebLogo 3 [16] to represent probability of an amino acid to be present in a certain location in the sequence window (Figure 2).

Sequence logos identified a strong positively charged/polar amino acid preference difference in -3rd position of the window (Figure 2a,b). Sezerman grouping results also supported this finding by more than 40% probability assigned to positively charged amino acids group (Q), and more than 20% probability assigned to polar amino acid groups (D, E) (Figure 2c,d). Similar difference can be observed in +8th position, in which both Figure2a-b comparison and Figure 2c-d comparison reveals a different preference of amino acids. Lastly, Figure 2a and 2b reveals additional differences between groups, such as overrepresentation of A at position -7, and overrepresentation of positively charged amino acids at position +3, however, commenting on these differences may require additional evidence, such as statistical testing.

Statistical Testing

We have performed statistical testing to identify differences in two aspects: overall amino acid composition, single amino acid over/underrepresentation in each position of sequence windows. Overall amino acid compositions showed no statistically significant difference occurs between positions (all p-values > 0.05).

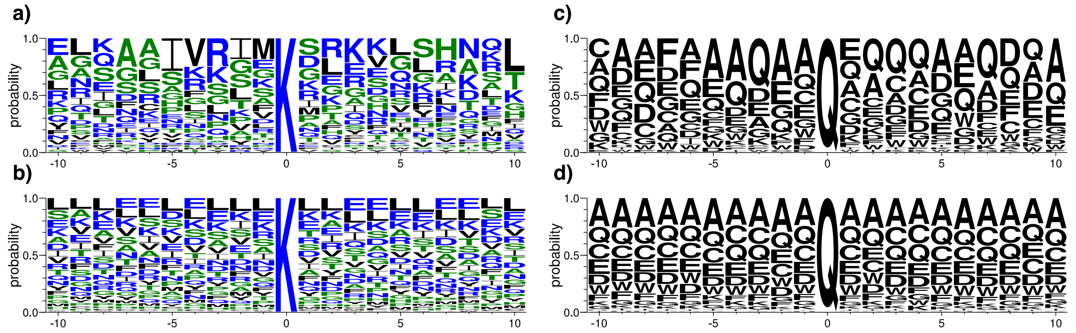

Figure 2. Sequence logos of neddylated and not neddylated sequence windows, centered around lysine residues. (a) neddylated sites, (b) not neddylated sites, (c) neddylated sites in Sezerman grouping, (d) not neddylated sites in Sezerman grouping. All sequence logos were created using WebLogo 3 [16].

On the other hand, single amino acid preference tests revealed several under/overrepresented amino acids. As it has been identified from sequence logos, positively charged amino acids (Sezerman grouping: Q) were significantly overrepresented in position -3 (44% of positive windows, while 15% of the negative windows), χ^2 (1, N = 960) = 28.44, $p < 0.001$. In this particular site, only arginine presence was statistically significant too, with arginine present in 25% of the positive sites and 4% of the negative sites, χ^2 (1, N = 960) = 40.09, $p < 0.001$. Charged and polar amino acids were found statistically significant in other two locations: +7 and +8. Histidine was overrepresented in position +7. 19% of the positive sites and only 2% of the negative sites have histidine in this position, χ^2 (1, N = 960) = 40.75, $p < 0.001$. Asparagine presence was found to be significant in position +8, where it is present in the 17% of the positive sites and 4% of the negative sites, χ^2 (1, N = 960) = 16.23, p = 0.004.

Apolar amino acids showed overrepresentation on various positions as well. Alanine was statistically significantly overrepresented with 25% of the positive sites, and 6% of the negative sites in position -7, χ^2 (1, N = 960) = 23.44, $p < 0.001$. Similarly, a methionine overrepresentation (19% of the positive sites, 2% of the negative sites) in -1st position was declared statistically significant, χ^2 (1, N = 960) = 38.95, $p < 0.001$. Valine was overrepresented in neddylated windows at position -4 as well. It was present at 25% of the positive sites and 7% of the negative sites, χ^2 (1, N = 960) = 21.76, $p < 0.001$. Lastly, isoleucine was overrepresented at position -5. It was present 29% of the positive sites, while only 6% of the negative sites, χ^2 (1, N = 960) = 34.19, $p < 0.001$. This overrepresentation can also easily seen from the sequence windows in Figure 2a-b.

It should be worth noting that frequency based statistical testing results reported in this section should be taken into consideration carefully, as some of the amino acids may be declared significant due to only dataset-specific frequency differences, and they may not imply anything on underlying biological principle.

Hydrophobicity

Hydrophobicity can differentiate target sites from non-target sites as it is an important effect in protein-protein binding [17]. Efficacy of hydrophobicity has shown in sumoylation site prediction previously [18]–[20]. Therefore, same principle may lead neddylation site recognition. In order to identify hydrophobicity differences, we have plotted boxplots of Kyte-Doolittle [12] hydrophobicities for each location and performed Mann-Whitney U tests to determine statistical significance (Figure 3). In Kyte-Doolittle [12] hydrophobicity scale, while hydrophobic residues have positive scores, hydrophilic residues have negative scores.

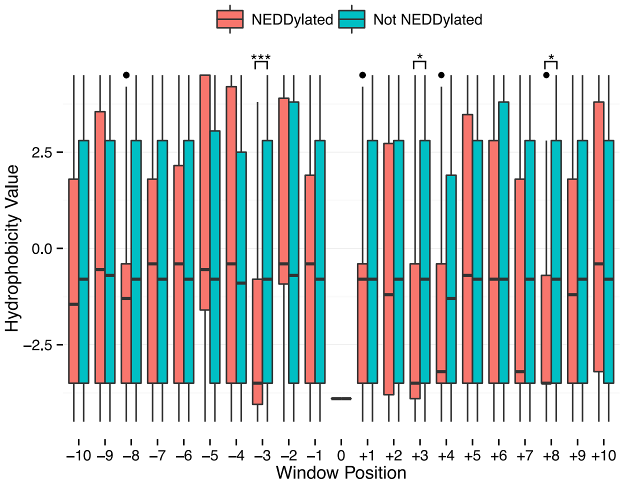

Figure 3. Boxplots of Kyte-Doolittle [12] hydrophobicity distribution among 21-amino acids long sequence windows. Statistical significance was assessed with Mann-Whitney U tests. All p-values were corrected using Benjamini-Hochberg [13] procedure. *: $p < 0.05$, ***: $p < 0.001$.

Supporting the sequence logos and statistical testing results, hydrophobicity plots has showed a statistically significant charged residue preference in -3rd position (Figure 3, $p<0.001$). Similarly, boxplots have showed a polar tendency in positions +3 and +8 (Figure 3, $p<0.05$). Although, they were not statistically significant, positions +4 and +7 also shows a polar tendency, as well.

DISCUSSION

On overall, we have showed that neddylation sites does not have significantly overrepresented "consensus" motif, as it was the case for sumoylation. However, this may only because of the limitations of the dataset. On the other hand, we have identified various significant amino acid preferences, especially charged amino acids in -3rd position. This fact may imply a significance of this position in neddylation site recognition by UBC12.

Identification of amino acid preferences may be the first step in decrypting neddylation site recognition, and accomplishing successful *in silico* identification of neddylation sites may open up various new application fields, such as studying ubiquitination and proteome degredation abnormalities and associated diseases. However, small size of experimentally identified neddylation sites seriously limits the *in silico* efforts to identify neddylation sites. Hence, with the ever-increasing amount of experimentally identified neddylation targets, we expect neddylation site identification methodologies will grow significantly.

In addition to obtaining primary sequences of new experimentally validated neddylation sites, we may need additional structural insights of site recognition, as neddylation site recognition may not only determined by primary sequence and hydrophobicity, but also conformational state, flexibility and subcellular localization. As most of the post-translational modifications are dynamic processes affected significantly by subcellular localization and conformational state, obtaining such information would be enlightening in understanding the actual mechanism of site recognition. However, it seems unlikely to have access to this information on neddylation sites, soon. Therefore, systematical identification of neddylation proteome still presents a great challenge.

CONCLUSIONS

Neddylation, a vital post-translational protein modification, plays significant roles in cellular machinery as it mainly functions as a regulator of ubiquitin-protein ligases and proteome degradation system. In this paper, we showed several amino acid preferences in neddylation targets via sequence logos, statistical testing and hydrophobicity scales. Future work lays in developing a neddylation site predictor for the use of research community that uses other possible properties that may affect neddylation site recognition as well.

ACKNOWLEDGMENTS

Ahmet Sinan Yavuz is supported by 2211 – National Scholarship Program for PhD Students of The Scientific and Technological Research Council of Turkey (TUBITAK).

REFERENCES

[1] A. G. van der Veen and H. L. Ploegh, "Ubiquitin-like proteins.," *Annu. Rev. Biochem.*, vol. 81, pp. 323–57, Jan. 2012.

[2] G. Rabut and M. Peter, "Function and regulation of protein neddylation. 'Protein modifications: beyond the usual suspects' review series.," *EMBO Rep.*, vol. 9, no. 10, pp. 969–76, Oct. 2008.

[3] J. Herrmann, L. O. Lerman, and A. Lerman, "Ubiquitin and ubiquitin-like proteins in protein regulation.," *Circ. Res.*, vol. 100, no. 9, pp. 1276–91, May 2007.

[4] D. P. Xirodimas, "Novel substrates and functions for the ubiquitin-like molecule NEDD8.," *Biochem. Soc. Trans.*, vol. 36, no. Pt 5, pp. 802–6, Oct. 2008.

[5] Y. Chen, R. L. Neve, and H. Liu, "Neddylation dysfunction in Alzheimer's disease.," *J. Cell. Mol. Med.*, vol. 16, no. 11, pp. 2583–91, Nov. 2012.

[6] Y. S. Choo, G. Vogler, D. Wang, S. Kalvakuri, A. Iliuk, W. A. Tao, R. Bodmer, and Z. Zhang, "Regulation of parkin and PINK1 by neddylation.," *Hum. Mol. Genet.*, vol. 21, no. 11, pp. 2514–23, Jun. 2012.

[7] W.-T. Yao, J.-F. Wu, G.-Y. Yu, R. Wang, K. Wang, L.-H. Li, P. Chen, Y.-N. Jiang, H. Cheng, H. W. Lee, J. Yu, H. Qi, X.-J. Yu, P. Wang, Y.-W. Chu, M. Yang, Z.-C. Hua, H.-Q. Ying, R. M. Hoffman, L. S. Jeong, and L.-J. Jia, "Suppression of tumor angiogenesis by targeting the protein neddylation pathway.," *Cell Death Dis.*, vol. 5, p. e1059, Jan. 2014.

[8] The Uniprot Consortium, "Activities at the Universal Protein Resource (UniProt).," *Nucleic Acids Res.*, vol. 42, no. 1, pp. D191–8, Jan. 2014.

[9] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.," *Bioinformatics*, vol. 22, no. 13, pp. 1658–9, Jul. 2006.

[10] M. C. Cobanoglu, Y. Saygin, and U. Sezerman, "Classification of GPCRs Using Family Specific Motifs.," *IEEE Trans. Comput. Biol. Bioinforma.*, pp. 1–15, Sep. 2010.

[11] A. S. Yavuz, B. Ozer, and U. Sezerman, "Pattern recognition for subfamily level classification of GPCRs using motif distillation and distinguishing power evaluation," *Lect. Notes Comput. Sci.*, vol. 7632, pp. 267–276, 2012.

[12] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein.," *J. Mol. Biol.*, vol. 157, no. 1, pp. 105–32, May 1982.

[13] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *J. R. Stat. Soc. Ser. B*, vol. 57, no. 1, pp. 289–300, 1995.

[14] The Python Consortium, "Python." [Online]. Available: http://www.python.org.

[15] The SciPy Consortium, "SciPy." [Online]. Available: http://www.scipy.org.

[16] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, "WebLogo: a sequence logo generator.," *Genome Res.*, vol. 14, no. 6, pp. 1188–90, Jun. 2004.

[17] N. T. Southall, K. A. Dill, and A. D. J. Haymet, "A view on the hydrophobic effect," *J. Phys. Chem B*, vol. 106, pp. 521–533, 2002.

[18] A. S. Yavuz and U. Sezerman, "SUMOtr: SUMOylation site prediction based on 3D structure and hydrophobicity," in *2010 5th International Symposium on Health Informatics and Bioinformatics*, 2010, pp. 93–97.

[19] Y. Z. Chen, Z. Chen, Y. A. Gong, and G. Ying, "SUMOhydro: A novel method for the prediction of SUMOylation sites based on hydrophobic properties," *PLoS One*, vol. 7, no. 6, p. e39195, 2012.

[20] A. S. Yavuz and O. U. Sezerman, "Predicting sumoylation sites using support vector machines based on various sequence features, conformational flexibility and disorder," *BMC Bioinformatics*, 2014, in press.