POSSIBLE ROLE OF SIGNALLING PEPTIDES IN MAMMALIAN PROTEIN EXPRESSION, *IN SILICO*

AFSHAN KALEEM[1*], ANAM KHAN[2], MEHWISH IQTEDAR,[1] ROHEENA ABDULLAH[1], IRFANA IQBAL[2] AND SHAGUFTA NAZ[1]

[1]Department of Biotechnology, Lahore College for Women University, Lahore, Pakistan

[2]Department of Zoology, Lahore College for Women University, Jail Road, Lahore, Pakistan *

Corresponding author e-mail: afshankaleem1904@gmail.com

خلاصہ

سگنل-پیپٹائڈز پروٹین میں موجود ہیں،اور سیل میں مختلف اجزاء کو نشانہ بنانے والے پروٹینز ہیں. پروٹین جو اینڈوپلاسمیک ریٹیکولک(ER) کو نشانہ بنایا جاتا ہے یا تو ER کے ڈھونڈے میں برقرار رکھا جاتا ہے، یا پھر دوسرے سیلولر آرگنائز میں منتقل کیا جاتا ہے. سیل کے اندر پروٹین کے مخصوص ھدف بندی مخصوص ترتیب مقاصد کی طرف سے بیان کی گئی ہے. انسانی پروٹینز میں سگنل پیپٹائڈز کی موجودگی سیلیکو میں پروٹین کی مجموعی تقسیم کا تعین کرنے کے لئے سلیکو میں تحقیق کی گئی تھی. 1000 سے زیادہ انسانی پروٹین کو انوکٹ بی بی ڈیٹا بیس سے حاصل کیا گیا تھا. ان پروٹین سیلولر لوکلائزیشن کا تعین کرنے کے لئے مختلف بائیوٹیکنیک اوزار استعمال کیے گئے تھے. پروٹین کی زیادہ سے زیادہ تعداد 85 ~) ER٪ (میں واقع ہوئی اور گوجی کے آلات یا مونو کوڈیا میں تقریبا 15٪ پایا گیا تھا، جو مشورہ دیتے ہیں کہ ER سے نکلنے والے پروٹینز، جیولوجی یا منوکوہنڈیا کو سب سے زیادہ ہدایت دی جاتی ہے. لیسوسوم، سیوٹپلاسم یا جھلی میں بھی کم سے کم پروٹین پایا گیا تھا. تنظیموں میں سیلولر compartmentalization سیل کی ایک اہم میکانیزم ہے. پروٹین کے انٹرایکولر ٹوکری کے غلط ٹرانسمیشن راستیاتی حالات کی وجہ سے ہو سکتی ہے، کیونکہ پروٹین اپنی مناسب سائٹ پر نہیں پہنچتا، جس میں اہم سگنلنگ کیسیڈس کی غیر فعال یا غلطی پیدا ہوتی ہے.

Abstract

Signal-peptides are found in proteins, and target proteins to different organelles in the cell. Proteins that are targeted to the endoplasmic reticulum (ER) are either retained in the lumen of the ER, or further transported to other cellular organelles. Specific targeting of proteins to inside the cell is defined by specific sequence motifs. The presence of signalpeptides in human proteins was investigated *in silico* to determine overall distribution of proteins in the cell. 1000 human proteins were retrieved from the UniprotKB database. Different bioinformatic tools were utilized to determine these proteins cellular localization. Highest number of proteins was found to be located in the ER (~85%) and around 15% were found in the golgi apparatus or the mitochondria, suggesting that proteins leaving the ER, are most likely to be directed to the golgi or mitochondria. Smaller number of proteins was also found in lysosomes, cytoplasm or membrane. Cellular compartmentalization into organelles is an important mechanism of the cell. Incorrect transport of proteins to intracellular compartment can lead to pathological conditions, as the protein does not reach its proper site, causing either inactivation or misregulation of important signalling cascades.

Introduction

The *N*-terminal newly synthesized secretory and membrane proteins are usually signal sequences ranging from 16 to 50 amino acid residues. The signal sequences direct proteins to different organelles like insertion into the endoplasmic reticulum (ER) membrane in eukaryotes and consequently get cleaved off by signal peptidases found in the ER. These signal peptides are either degraded or may function on their own. Different signal peptides direct the protein towards different organelles, and mostly signal peptides are composed of a recognition motif, which is interpreted by the targeting machinery (Kunze and Berger 2015; Mukhopadhyay *et al.*, 2004).

The *N*-terminal signal sequence dictates the targeting of membrane and nascent secretory proteins to the ER. The signal sequences (normal 15-30 amino acids) consist of three different regions: A hydrophobic core region (h-region) which is flanked by an n- and c-region. The h-region, the most vital part for targeting and membrane insertion, comprise of mostly hydrophobic residues (7-15 amino acids). The n-region consist of positively charged hydrophobic residues such as Arg and Lys, and the c-region contains mostly neutral and polar residues with helix breaking Pro, Gly and small uncharged residues around the cleavage sites of signal peptides. The signal sequences get cleaved off either co-translationally or post-translationally following protein translocation (Mukhopadhyay *et al.*, 2004; Nicchitta, 2002). The great variability of signal sequences length and amino acid composition suggests that ER targeting and translocation etc. is modulated by the signal sequence. The nuclear localization signal (NLS) is a signal peptide that directs the protein to the nucleus. It mostly consists of five basic, positively charged amino acids and can be located anywhere in the peptide chain (Kosugi *et al.*,

2009). The mitochondrial signal peptides comprise of a sequence with alternating hydrophobic and positively charged amino acids at the *N*-terminus (mitochondrial targeting signal (MTS)). Matrix proteins have a signal sequence of 20-30 residues, which fold into an amphiphilic helix with positive charge. Proteins which are imported into other subcompartments of the mitochondria contains an additional signal sequence that re-routs the protein. These additional targeting signals are normally removed when the protein has reached its destination (Kunze *et al.*, 2015; Mukhopadhyay *et al.*, 2004; Schatz, 1996).

N-acetylation which is a co-translatinal process can inhibit ER translocation, and represents an early determining step in the cellular localization (Forte *et al.*, 2011). In this work the second amino acid (P2) in mammalian proteins is investigated to determine its role in proteins intracellular targeting.

The goal of this present study is to investigate the presence of signal peptides in mammalian proteins and their subcellular location by using readily available bioinformatic prediction tools.

Materials and Methods

Proteins FASTA sequences were retrieved from the UniprotKB database (Wu *et al.*, 2006). Two different prediction methods were utilized: PrediSi (Prediction of Signal peptides) and Predotar.

PrediSi is a prediction tool for signal peptide sequences and gives the position of cleavage in eukaryotic and bacterial amino acid sequences (Hiller *et al.*, 2004). It is based on neural networks that are trained on different sets of eukaryotic and prokaryotic sequences, and trained on using sequences extracted from the SwissProt databases.

Predotar is also a neural network based prediction program, which is capable of recognizing signal peptides targeted for ER and mitochondrial. Predotar was designed for systematic secreening of large batches of proteins for the identification of putative targeting sequences. Its best quality is its accuracy as it has very low rate of false positives as compared to other programs (Small *et al.*, 2004).

More than 80% of mammalian proteins undergo acetylation in the *N*-terminal, and is amongst the most common eukaryotic protein modifications. NetAcet is developed on a neural network based method with a 74 % sensitivity on mammalian data (Kiemer *et al.*, 2005). This tool was utilized to determine *N*-terminal acetylation of proteins and the role of P2 amino acid in cellular compartilization.

Results

In this work the sequence of 1000 randomly chosen mammalian proteins sequence data and their subcellular location was retrieved from UniProtKB. PrediSi was utilized for prediction of signal peptides, and the Predotar database was used to predict the presence and their subcellular locations.

Subcellular locations of mammalian protein from UniProtKb/SwissProt: Mostly proteins are secreted and are located in the ER or lysosomes and few are located in the mitochondria, cytoplasmic membrane, cell surface cell membrane and other as not found as shown in Table 1. The 1000 proteins which were collected from UniproKB, 801 were secreted proteins, 50 protein‘s sub cellular location was ER, 35 contain lysosome as their subcellular location, 19 as membranous, 3 as cell surface, 3 as cell membrane, 4 as cytoplasm, 4 resides in golgi apparatus and 81 are not found. The important thing about the secretory proteins is that these proteins do pass through the ER and often these proteins contain signal peptides that target them to the ER. The proteins that contain the subcellular location of ER include all the location in endoplasmic reticulum like ER lumen, sarcoplasmic reticulum, and rough ER. The proteins that contain the subcellular location of Golgi apparatus include all the areas in golgi like its lumen. Some proteins that localized in ER were also seen to be localized in golgi apparatus and lysosomes. Few proteins reside as membranous, cell membranous, cytoplasm and cell surface proteins.

Prediction of signal peptides by PrediSi: The prediction of presence or absence of signal peptides in mammalian proteins is determined by calculating its score and cleavage site by the Predisi tool. The signal peptide prediction, score and site of cleavage is shown in Table 1. PrediSi server provides a score on a scale between 0 and 1. A score greater than 0.5 indicates that the polypeptide or protein contains a signal peptide. PrediSi also give putative signal peptidase cleavage position in the examined protein sequence as shown in the Table 1. The predicted signal peptides were present in 933 sequences out of 1000 and absent in 67 sequences. Number of the proteins targeted to specific organelles is: 887 to ER, 16 to mitochondria, 17 to possibly mitochondria, 11 to none and 2 were discarded due to the absence of *N*-methionine terminal.

Prediction of organelle targeting by Predotar: Predotar provides a probability estimate, which tells if a sequence has an ER or mitochondrial targeting sequence. A large number of proteins were directed to the ER, some to mitochondria and others elsewhere. The probabilities of each protein are predicted by Predotar as

shown in Table 1. Out of 1000 mamalian proteins, the frequency of proteins targeting to ER is 852, to mitochondria is 13 and 73 proteins predicted as none. While the proteins that targeted to possibly ER or mitochondria are about 43 and 17 respectively and 2 sequences are not predicted and discarded as they contain no terminal methionine.

Determination of second amino acid in mammalian protein: *N*-terminal acetylation, which is an early determining step in the cellular sorting of polypeptides, is an important and conserved process. The bioinformatics tool NetAcet was utilized to determine *N*-terminal acetylation of proteins and to determine the P2 amino acid. When examining P2 amino acid in 1000 mammalian proteins, it became evident that the occurence of Ala was highest, followed by Lys, Arg and Gly (Fig. 1). 981 proteins showed potential for *N*-terminal acetylation except 19 proteins which lacked Met at position 1 in their sequence. The sequence of amino acids from highest to lowest present at second positions in ER targeting peptides was found as following:

Ala> Lys> Arg> Gly> Leu>Ser>Pro

Proteins targeted to mitochondria showed highest number of Ala at P2 followed by Val and Gly. When proteins were targeted to the golgi apparatues it was found that Gly, Ser and His were most likely to be present at P2.

Table 1. Subcellular locations of mammalian Protein form UniProtKb/SwissProt.

Subcellular Locations	Uniprot	Predisi	Predotar
Secreted	801	-	-
Endoplasmic Reticulum (Possibly ER)	50	887	852 (43)
Golgi Apparatus (Post Golgi)	4	-	-
Mitochondria (Possibly Mitochodria)		16 (17)	13 (17)
Lysozomes	35	-	-
Cell surface/ Cellular membrane	6	-	-
Cytoplasm	4	-	-
Membrane	19	-	-
Not found	81	11	73
Signal Peptide	-	67	-
Discarded	-	2	2
Total	1000	1000	1000

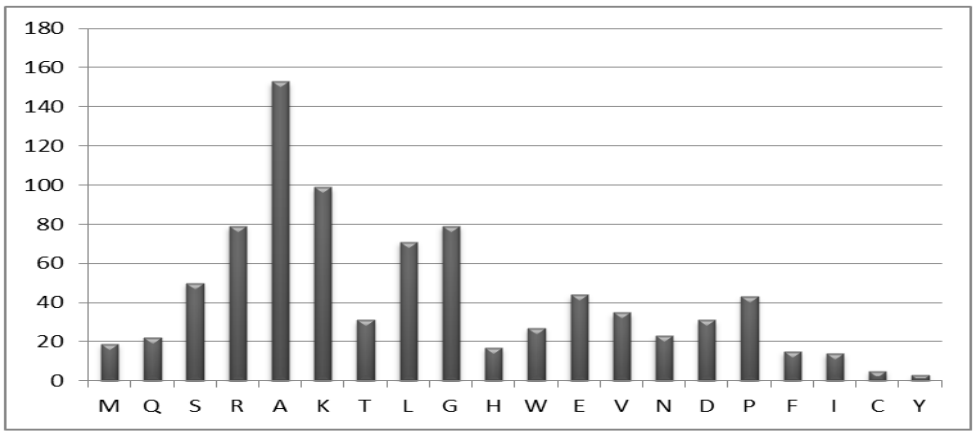

Fig. 1. Distribution of P2 amino acids in 981 mammalian proteins targeted to the ER.

Discussion

Protein subcellular localization is vital for understanding the function of proteins, but also the organization of the cell as a whole. Proteins subcellular localization is predicted by using different bioinformatic tools, thereby providing information for a large numbers of proteins in a short period of time.

The signal peptide has a crucial part in protein translocation and directed targeting in cells of eukaryotic and prokaryotic origin (Redrejo-Rodríguez *et al.*, 2012). Proteins encoded with a short sequence peptides acting like postal addresses are targeted for secretion or transferred to other organelles. But signal peptides are not found in all proteins, which suggest that other protein targeting mechanisms exist. Mostly secreted proteins contain an

ER signal sequence, generally at the *N*-terminus consisting of 5-10 hydrophobic amino acids. All these proteins are further delivered to the Golgi apparatus from the ER. In the *C*-terminal proteins containing an additional sequence, KDEL (lys-asp-glu-leu) end are retained or are returned back to the ER. This *C*-terminal KDEL motif is conserved in mammals, and works as an ER retention signals (Stornaiuolo *et al.*, 2003). The mitochondrial signal peptide directs proteins to the matrix, and contains a sequence consisting of positive charged and hydrophobic amino acid residues at the *N*-terminus, also called MTS (Schatz, 1996).

All the mammalian proteins and their subcellular location were retrieved form the UniProtKb/SwissProt database (Table 1). The number of secreted proteins were 80%, followed by proteins that direcly go to the ER or its lumen (about 50 proteins). In the ER secretory proteins are synthesized and assembled, and then further processed in the Golgi apparatus. First they arrive at the trans-Golgi where they are sorted and packed into carriers of post-Golgi, and then fuse with the cell surface after being transported into the cytoplasms. Secreted proteins do pass through the ER, so maximum number of proteins either resides in the ER or pass through it.

For targeting protein to its subsequent location the presence of signal peptide is compulsory that cleaves from protein upon reaching its targeted location. Once the protein has reached its destination, its signal peptide is cleaved, and it becomes a matured protein (Kunze and Berger 2015; Mukhopadhyay *et al.*, 2004). By using the PrediSi software signal peptides were determined and it was found to be present in 933 proteins, whereas 67 proteins lacked signal peptide. Mostly proteins containing signal peptides targeted to specific locations in the cell were localized in the ER. Predotar was used to determine the targeting locations of *N*-terminal signals. For every sequence, the tool provides a probability estimate about whether a protein sequence contains an ER or mitochondrial targeting sequence, and sometimes no targeting sequence is present. It was found that the frequency of proteins targeting to ER was 852, to mitochondria 16 and 35 proteins predicted as none. The proteins targeted to possibly ER or mitochondria are about 43 and 17 respectively and 2 sequences are not predicted and discarded as they contain no terminal Met. This indicates that the ER has first preference regarding localization of proteins and are either utilized in the ER or further transported to other organelles in the cell. This phenomenon was also observed in the proteins targeted to the golgi apparatus, as they also have the ER targeting sequence included in their targeting sequence.

Conclusion

The insight about the function of a protein can be determined by knowing where and when it is located in the cell i.e. the proteins location. Most proteins contain *N*-terminal sequences that, in one way or another, guide the cell where to move unmatured proteins. The information regarding proteins subcellular localization in medical sciences can be helpful in target directed drug discovery. For example, drug molecules are easily accessible to secreted and plasma membrane proteins due to their localization. Aberrant subcellular localization of proteins may lead to several pathological conditions, such as Alzheimer's disease and cancer (Davis and Williams 2012; Hughes *et al.*, 2000).

References

Davis, R.E. and Williams, M. (2012). "Mitochondrial Function and Dysfunction: An Update," *J Pharmacol Exp Ther* 342: 598-607.

Forte, G.M.A., Pool, M.R. and Stirling, C.J. (2011). "N-Terminal Acetylation Inhibits Protein Targeting to the Endoplasmic Reticulum," *PLoS Biol* 9: e1001073.

Hiller, K., Grote, A., Scheer, M., Münch, R. and Jahn, D. (2004). "PrediSi: prediction of signal peptides and their cleavage positions," *Nucleic Acids Res* 32 (Web Server issue): W375-379.

Hughes, A.E., Ralston, S.H., Marken, J., Bell, C., Macpherson, H., Wallace, R.G., Van Hul, W., Whyte, M.P., Nakatsuka, K., Hovy, L. and Anderson, D.M. (2000). "Mutations in TNFRSF11A, affecting the signal peptide of RANK, cause familial expansile osteolysis," *Nat Gen* 24: 45-48.

Kiemer, L., Bendtsen, J.D. and Blom, N. (2005). "NetAcet: prediction of N-terminal acetylation sites," *Bioinform* 21: 1269-1270.

Kosugi, S., Hasebe, M., Matsumura, N., Takashima, H., Miyamoto-Sato, E., Tomita, M. and Yanagawa, H. (2009). "Six classes of nuclear localization signals specific to different binding grooves of importin alpha," *J Biol Chem* 284: 478-485.

Kunze, M. and Berger, J. (2015). "The similarity between N-terminal targeting signals for protein import into different organelles and its evolutionary relevance," *Front Physiol*, 259. Mukhopadhyay, A., Ni, L. and Weiner H. (2004). "A co-translational model to explain the in vivo import of proteins into HeLa cell mitochondria," *Biochem J* 382: 385-392.

Mukhopadhyay, A., Ni, L. and weiner, H. (2004). A co-translational model to explain the in vivo import of proteins into HeLa cell mitochondria," *Biochem J* 382: 385-392.

Nicchitta C.V. (2002). "A platform for compartmentalized protein synthesis: Protein translation and translocation in the ER," *Curr Opin Cell Biol* 14: 412-416.

Redrejo-Rodríguez, M., Muñoz-Espín, D., Holguera, I., Mencía, M. and Salas, M. (2012). "Functional eukaryotic nuclear localization signals are widespread in terminal proteins of bacteriophages," *Proc Natl Acad Sci USA* 109: 18482-18487.

Schatz, G. (1996). "The Protein Import System of Mitochondria," *J Biol Chem* 271: 31763-31766.

Small, I., Peeters, N., Legeai., F. and Lurin, C. (2004). "Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences," *Proteom* 4: 1581-1590.

Stornaiuolo, M., Lotti, L.V., Borgese, N., Torrisi, M.-R., Mottola, G., Martire, G. and Bonatti, S. (2003). "KDEL and KKXX Retrieval Signals Appended to the Same Reporter Protein Determine Different Trafficking between Endoplasmic Reticulum Intermediate Compartment, and Golgi Complex," *Mol Biol Cell* 14: 889-902.

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'donovan, C., Redaschi, N. and Suzek, B. (2006). "The Universal Protein Resource (UniProt): an expanding universe of protein information," *Nucleic Acids Res* 34 (Database issue): D187-191.