Bioinformatics, Proteomics and Immaging Analysis

OMMEGA Publishers
Answering Intelligence
www.ommegaonline.org

Research Article Open Access

Computational Identification of Phosphorylation Sites around Nuclear Localization Signal Sequence Reveals New Insight into Genes Associated With Human Diseases

Eric Chen, Jianjun Hu*

Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA

***Corresponding author:** Dr. Jianjun Hu, Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, 29208, USA; Tel: +1(803)7777304; Fax:+1(803)7773747; Email: jianjunh@cse.sc.edu

Abstract

Alterations in protein Subcellular localization often contribute to the development of human diseases. Post-transcriptional modifications, such as phosphorylation on the Nuclear Localization Signal (NLS), may change the protein's localization. However, little is known about the frequency and local effects of phosphorylation near NLS sites. In this study, a computer program was developed to search various databases in order to find proteins with NLS Phosphorylations, and any diseases that are associated with those genes. 308 NLS sequences were found in the NLSdb database, resulting in the identification of 133,448 NLS-containing proteins in the Uniprot database. We cross-referenced these proteins with phosphorylation data available from the PhosphoSitePlus database and found that about 21% of these NLS-containing proteins have evidence of phosphorylation sites. After plugging this into the gene disease association database, 138 of disease-associated genes (1% of NLS-containing proteins)were identified to have phosphorylation sites on their NLS sequences. Further evaluation of the NLS phosphorylation status of these genes in clinical samples may lead to development of new biomarkers for human diseases, and shed new light into the pathogenesis of these gene-associated diseases.

Keywords: Nuclear localization signal sequence; Phosphorylation; Biomarker

Received Date: November 11, 2016
Accepted Date: December 02, 2016
Published Date: December 17, 2016

Citation: Jianjun Hu., et al. Computational Identification of Phosphorylation Sites around Nuclear Localization Signal Sequence Reveals New Insight into Genes Associated With Human Diseases (2017) Bioinfo Proteom Img Anal 3(1): 177- 181.

DOI: 10.15436/2381-0793.16.1217

Introduction

All eukaryotic cells have a complex endomembrane system and contain elaborate organelles that provide distinct compartments for different metabolic activities. Protein translation is confined to only one of these compartments, the cytosol, but proteins are needed for nearly all cellular functions. Thus, the translocation of proteins is a fundamental requirement for proteins to exert their functions in different organelles. In fact, approximately half of the proteins generated by a cell have to be transported across at least one cellular membrane to reach their functional destinations[1]. Subcellular localization is essential to protein function, as it determines the access of proteins to interacting partners and the post-translational modification machinery that allows the integration of proteins into functional biological networks.

Considering the importance of the Subcellular localization of proteins, it is not surprising that the disruption of nuclear-cytoplasmic transport is responsible for cause of many diseases, and is a potent mechanism for resistance to drug treatments. Aberrant protein localization can be caused by mutation, altered expression of cargo proteins or transport receptors, or deregulation of components of the trafficking machinery[2]. For example, a number of the major oncogenes and tumor suppressors such as p53, BCRA1, APC, and retinoblastoma (Rb), β-catenin, NF-κB, survivin and cyclin D1 have been reported to have aberrant Subcellular localization in various types of cancers[3,4]. The mis-localization of these proteins can alter their function so that their ability to suppress tumor cells is diminished, or their ability to induce cancer development, metastasis, or drug resistance is increased.

Nuclear transport is highly regulated by the nuclear pore complex (NPC) to ensure that proteins can enter when their functions are required and exit into the cytoplasm when they are not needed. While proteins less than 40 kDa in size are free to traverse the NPC, larger proteins require active transport directed by nuclear localization or nuclear export sequences (NLSs). For nuclear entry, proteins must negotiate with a NPC that is comprised of over 30 different protein components called nucleoporins (Nups)[4].

Post-transcriptional modification (PTM)-based modulation of the NLS binding affinity to import receptors is one of the most understood mechanisms that regulates the nuclear import of proteins[4-7]. Our previous study has developed an effective algorithm to predict nuclear import activity, in which molecular interaction energy components (MIECs) were used to characterize NLS-import receptor interaction, and a support vector regression machine(SVR) was used to learn the relationship between the characterized NLS-import receptor interaction and the corresponding nuclear import activity[8]. Based on our model, we developed a systematic framework to precisely predict how potential PTM, such as phosphorylation, regulates nuclear import of human and yeast nuclear proteins. In this study, we developed a computation-based screening method to survey NLS sites for potential modifications like phosphorylations that may lead to protein mislocalization in human disease cells. We hypothesize that the mislocalizations of such proteins, not just expression levels, can serve as novel diagnostic markers or therapeutic targets for human diseases.

Methods

With Python, a bioinformatics tool has been developed to search for phosphorylation sites on localization signals of disease-associated genes. The program starts by reading through all the rows of database of nuclear localization signals (NLSdb, https://rostlab.org/services/nlsdb/). The database contains 114 experimentally determined NLSs that were obtained through an extensive literature search, extended to 308 experimental and potential NLSs using “in silico mutagenesis”. This final set matched over 43% of all known nuclear proteins and matched no known non-nuclear proteins.

NLS sequences in the NLSdb are used to identify the predicted nuclear proteins with links to Uniprot (http://www.uniprot.org/). Uniprot provides the FASTA sequence for each protein, as well as the protein’s accession ID, which is used to search the phosphorylation site database. PhosphoSitePlus (http://www.phosphosite.org) is an open, comprehensive, manually curetted and interactive resource for studying experimentally observed post-translational modifications, primarily of human and mouse proteins. It encompasses 1,300,000 non-redundant modification sites, primarily phosphorylation, ubiquitinylation, and acetylation sites. If a phosphorylation site was found on the NLS, the phosphorylation site and sequence and a gene symbol were recorded.

This gene symbol was then used to search the gene-disease association dataset (DisGeNET, http://www.disgenet.org). DisGeNET is a discovery platform integrating information on gene-disease associations (GDAs) from several public data sources and the literature[9-11]. The current version (DisGeNET v3.0) contains 429,111 associations between 17,181 genes and 14,619 diseases, disorders, and clinical or abnormal human phenotypes. A gene-disease score was provided to rank the strength of the associations based on the level supporting evidence, taking into account the number and type of sources (level of curation, organisms), including the number of publications supporting the association. Searching through the DisGeNET database yielded a list of associated diseases, as well as the gene-disease score. These results were ll written to a table that was further processed by hand to remove duplicates, and filter by score(Table 1).

Table 1: Selected top gene list with a 0.5 gene-disease score cutoff filter.

Gene	Gene Description	NLS	Phos. Site	Disease	Score[1]	Assoc. Type
TP53	tumor protein p53	304kRALPNNtsssPQPkkkP322	S315; T312	Adenocarcinoma	0.51	Biomarker
DKC1	dyskeratosiscongenita 1, dyskerin	445KRKREsEsEsDEtPPAAPQLIKKEKKK472	S451	HoyeraalHreidarsson syndrome	0.60	Genetic Variation
XPC	xerodermapigmentosum, complementation group C	54KRKRGCsHPGGSADGPAKKKVAK77	S61	XerodermaPigmentosum	0.70	Genetic Variation
VDR	vitaminD(1,25-dihydroxyvitaminD3) receptor	48RRsMKRK55	S51	Vitamin D-Dependent Rickets, Type 2A	0.70	Genetic Variation
PTHLH	parathyroid hormone-like hormone	102YLTQEtNKVEtYKEQPLKtPGKKKKGKP130	T108; T121	Brachydactyly, Type E2	0.60	Genetic Variation
MYH3	myosin, heavy chain 3, skeletal muscle, embryonic	939KKRKLEDECsELkkDIDDLELtLAKVEKEK969	S949; T961	Freeman-Sheldon syndrome	0.60	Genetic Variation
MYH6	myosin, heavy chain 6, cardiac muscle, alpha	940KKRKLEDECsELkkDIDDLELtLAKVEKEK970	S950; T962	Cardiomyopathy,Familial Hypertrophic, 14	0.70	Genetic Variation
MYH7	myosin, heavy chain 7, cardiac muscle, beta	938KKRKLEDECsELkRDIDDLELtLAKVEKEK968	S948; T960	Myopathy, Myosin Storage	0.60	Genetic Variation
MYH8	myosin, heavy chain 8, skeletal muscle, perinatal	941KKRKLEDECsELkkDIDDLELtLAKVEKEK971	S951; T963	Hecht syndrome	0.60	Genetic Variation

POR	P450 (cytochrome) oxidoreductase	44RKKKEEVPEFTkIQtLTssVRESS-FVEKMK74	T59	Antley-Bixler Syndrome With Genital Anomalies and Disordered Steroidogenesis	0.60	Genetic Variation
TJP2	tight junction protein 2	232RDRDRDRsRGR243	S240	Hypercholanemia, Familial	0.60	Genetic Variation
KCNJ1	potassium inwardly-rectifying channel, subfamily J, member 1	184PKKRAKt191	T191	Bartter syndrome, antenatal ,		
type 2	0.70	Genetic Variation				
DSP	desmoplakin	2448KKKQVQTSQKNtLRKRR2465	T2460	Skin Fragility-Woolly Hair Syndrome	0.60	Genetic Variation
ADD1	adducin 1 (alpha)	716KKKKKFRtPsFLKKsKKK734	T724	Hypertension	0.52	Biomarker
VCP	valosin containing protein	58LkGKKRR65	T56	Paget Disease, Frontotemporal Dementia	0.71	Genetic Variation
ATRX	alpha thalassemia/ mental retardation syndrome X-linked	1043KKSKKIRDKTSKKKDELsDy-AEKSTGKGDsCDssEDKKSK1083;	S1061; S1073	ATR-X syndrome	0.71	Genetic Variation
TCOF1	Treacher Collins-Franceschetti syndrome 1	1438KKEKKKSDKRKKDKEK-KEKKKKAKKASTKDSEsPsQKKK-KKKKK1482;	S1471; S1350	Mandibulofacial Dysostosis	0.72	Genetic Variation

TNNT2	troponin T type 2 (cardiac)	182RKKKALsNMMHFGGyIQKQA-QtERKsGKRQtEREKKKK220	S189;
			Y197;
			T213
				Cardiomyopathy, Dilated, 1D	0.70	Genetic Variation

PLEC	pectin	2044RQKGLVEDtLRQRR2058	T2053	Epidermolysis bullosa simplex, Ogna type	0.70	Genetic Variation

Note: [1]DisGeNET score was developed and calculated based on thenumber of sources that report the association, the type ofcuration of each of these sources, the animal models wherethe association has been studied, and the number of supportingpublications from text-mining based sources as described in the literature[9-11].

Results and Discussion

Aberrant protein localizations are responsible for many diseases. Considerable effort has been devoted to developing reliable methods to predict the effect of mutations on the Subcellular localization of disease related proteins[12-14], and a substantial amount of experimental data has been collected on their mislocalizations. For example, loss of the nuclear localization signal (NLS) due to a missense mutation within the NLS of the sex-determining region of the Y protein (SRY) has been shown to be associated with XY sex reversal in Swyer syndrome[15]. With the development of genomic and proteomic approaches, gene expression at both mRNA and protein levels can be quantitatively analyzed in a global manner[16,17], but these "omic" methods are not sufficient to detect alterations in the protein localizations. Currently there is no high throughput technology available to screen for mislocalized proteins in a global way.

Changes in localization are often triggered post-transcriptional modifications. While it is currently estimated that 40 to 50% of eukaryotic proteins are phosphorylated, little is known about the frequency and local effects of phosphorylation near nuclear binding sites. NLS signals tend to be conserved between proteins, and are thus sensitive to alterations[13]. Our previous study has shown that phosphorylation on NLS residues has a dramatic impact on the binding ability of nuclear proteins to nuclear import receptors, and therefore affects nuclear localization efficiency[8].

In this study, we investigated how frequently phosphorylation sites are near the NLS, how they may be modified to affect the protein localization, and the human diseases they are potentially associated with. This was done with the NLS motif database NLSdb, phosphorylation site database Phosphodict, and disease-gene association database disGeNET. The NLSdb website contained 308 potential NLS motifs, which matched over 43% of all known nuclear proteins and no currently known non-nuclear proteins. DisGeNET gave a score for each association, ranging from 0 to 1 that indicated the strength of the association.

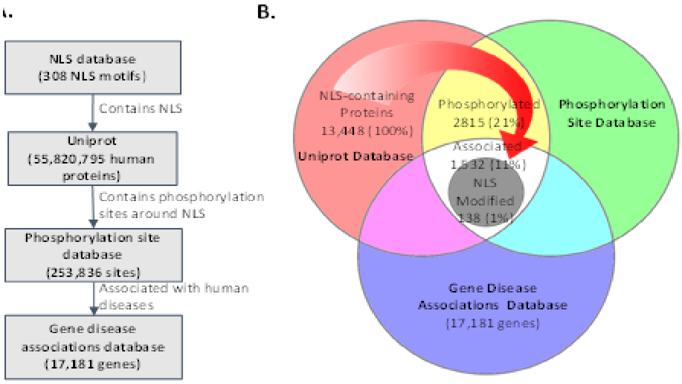
A.
NLS database
(308 NLS motifs)
Contains NLS
Uniprot
(55,820,795 human proteins)
Contains phosphorylation sites around NLS
Phosphorylation site database
(253,836 sites)
Associated with human diseases
Gene disease associations database
(17,181 genes)
B.
NLS-containing Proteins
13,448 (100%)
Uniprot Database
Phosphorylated
2815 (21%)
Phosphorylation Site Database
Associated
1,532 (11%)
NLS Modified
138 (1%)
Gene Disease Associations Database
(17,181 genes)

Figure 1: Flowchart of building the bioinformatics tool. A) Dataset preparation. The program starts with a list of NLS-containing proteins in NLSdb, find additional information on Uniprot, then narrows the list by searching for phosphorylation sites and disease associations. B) Graphical representation of program search algorithm for identification of the 138 disease associated gene hits as defined in text.

Our study aims to identify the NLS-containing proteins whose localization may be regulated by post-transcriptional modifications such as phosphorylation. Using the Python language, we developed a bioinformatics module to survey phosphorylation sites on nuclear localization signals and potential disease associations. We first identified 13,448 NLS-containing proteins(Figure.1A). Next, we search through the phosphorylation site database for the NLS-containing proteins. Out of 13448 NLS-containing proteins 2,815 (21%) were also found to have a phosphorylation site listed in the phosphorylation database (Figure.1B). Next, we cross-referenced the disease associated entries for the proteins with potential phosphorylation sites at NLS residues.

Through this screening process, we identified 270 phosphorylation sites on the NLS sequences of 138 proteins which potentially contribute to disease development. These genes were found to be associated with many different types of diseases including cancer, cardiovascular disease, obesity etc. Interestingly, 32 out of the 138 (27.8%) proteins were known to be associated with cancers (Figure.2).

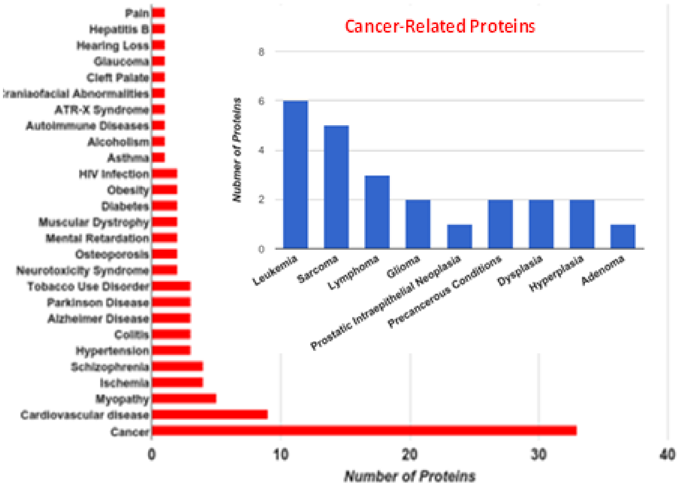

Figure 2: Distribution of potential disease-related proteins with phosphorylation sites in their NLS sequences among human diseases, including different types and stages of cancer.

These cancer-related genes are found in different types of cancer and different stages of cancer (Figure. 2). The fact that there are so many cancer related genes on our list reflects the large amount of publications and research work on cancer in the literature. However, when we restricted associations to those with a score of 0.5 or higher, only 5% of the genes on the shortened list were cancer-associated genes, and the rest of genes were strongly associated other genetic diseases. The reason for this discrepancy may be due to the fact that protein mislocalization was not a factor which was used to calculate the score in the DisGeNET database. This means that a large number of genes can be potential diagnostic and/or prognostic markers if the localization of the gene products is taken into account For example, the disease-association score for metastasis associated 1 (MTA1) gene is only 0.01 in prediction of breast neoplasm, because even in normal cells, MTA1 levels vary a great deal from tissue to tissue[18-20], and little association is found when looking at expression level alone. However, several studies have shown that MTA1 is located in the nucleus, cytoplasm, and the nuclear envelope[21], and further investigations are needed to identify the exact Sub-cellular localizations of MTA1 proteins. We reviewed the sub-cellular localization patterns of the MTA family members and gavea comprehensive overview of their respective molecular activities in multiple contexts.

Some associations identified in our analysis appear consistent with existing research, like that between p53 phosphorylation and localization. TP^{53} is a tumor suppressor gene, i.e., its activity stops the formation of tumors[22,23]. Under normal conditions, the p53 protein, which is encoded by the TP^{53} gene, is a labile and inactive protein. Cytoplasmic p53 interacts with MDM2, which serves as an E3 ubiquitin ligase and targets p53 for ubiquitin-proteasome-mediated degradation[23,24] When cells are exposed to DNA damage and other stress, p53 accumulates in the nucleus and becomes active[22]. Previous studies report that phosphorylation of p53 at Ser315 inactivates p53 by enhancing its proteolytic degradation in the cytoplasm[25,26]. Although the mechanism by which Ser315 regulates cytoplasmic retention of p53 is largely unknown, these observations are consistent with our hypothesis that phosphorylation around the NLS may reduce its binding affinity to nuclear import receptors and therefore inhibit its nuclear translocation.

Overall, we developed a program to systematically survey for proteins whose cellular localizations may be altered due to phosphorylation of their nuclear localization signal sequences during the development of human diseases. Further experimental validation of their expression levels and localizations in clinical samples may lead to the identification of a new set of diagnostic and/or prognostic biomarkers for human diseases.

Acknowledgments: This work was supported in part by the National Science Foundation Career Award [grant number BIODBI-0845381] to Hu J.

Conflict of interest: The authors declare no conflict of interest.

Author's contribution: EC, development of software program, interpretation of data, writing paper; JH, Study conception and design, writing paper.

References

1. Chacinska, A., Koehler, C.M., Milenkovic, D., et al. Importing mitochondrial proteins: machineries and mechanisms. (2009) Cell 138(4): 628-644.
2. Hung, M.C., Link, W. Protein localization in disease and therapy. (2011) J Cell Sci 124(Pt 20): 3381-3392.
3. Hill, R., Cautain, B., de Pedro, N., et al. Targeting nucleocytoplasmic transport in cancer therapy. (2014) Oncotarget 5(1): 11-28.
4. Kau, T.R., Way, J.C., Silver, P.A. Nuclear transport and cancer: from mechanism to intervention. (2004) Nat Rev Cancer 4(2): 106-117.
5. Poon, I.K., Jans, D.A. Regulation of nuclear transport: central role in development and transformation? (2005) Traffic 6(3): 173-186.
6. Gorner, W., Schuller, C., Ruis, H. Being at the right place at the right time: the role of nuclear transport in dynamic transcriptional regulation in yeast. (1999) Biol Chem 380(2): 147-150.
7. Okada, N., Sato, M. Spatiotemporal Regulation of Nuclear Transport Machinery and Microtubule Organization. (2015) Cells 4(3): 406-426.
8. Lin, J.R., Liu, Z., Hu, J. Computational identification of post-translational modification-based nuclear import regulations by characterizing nuclear localization signal-import receptor interaction. (2014) Proteins 82(10): 2783-2796.
9. Gutierrez-Sacristan, A., Grosdidier, S., Valverde, O., et al. PsyGeNET: a knowledge platform on psychiatric disorders and their genes. (2015) Bioinformatics 31(18): 3075-3077.
10. Pinero, J., Queralt-Rosinach, N., Bravo, A., et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. (2015) Database (Oxford).

11. Queralt-Rosinach, N., Pinero, J., Bravo, A., et al. DisGeNET-RDF: harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases. (2016) Bioinformatics 32(14): 2236-2238.
12. Emanuelsson, O., Brunak, S., von Heijne, G., et al. Locating proteins in the cell using TargetP, SignalP and related tools. (2007) Nat Protoc 2(4): 953-971.
13. Laurila, K., Vihinen, M. Prediction of disease-related mutations affecting protein localization. (2009) BMC genomics 10: 122.
14. Nair, R., Rost, B. Protein subcellular localization prediction using artificial intelligence technology. (2008) Methods Mol Biol 484: 435-463.
15. McLane, L.M., Corbett, A.H. Nuclear localization signals and human disease. (2009) IUBMB Life 61(7):697-706.
16. Reinhold, W.C., Varma, S., Rajapakse, V.N., et al. Using drug response data to identify molecular effectors, and molecular “omic” data to identify candidate drugs in cancer. (2015) Hum Genet 134(1):3-11.
17. Villoslada, P., Baranzini, S. Data integration and systems biology approaches for biomarker discovery: challenges and opportunities for multiple sclerosis. (2012) J Neuroimmunol 248(1-2):58-65.
18. Levenson, A.S., Kumar, A., Zhang, X. MTA family of proteins in prostate cancer: biology, significance, and therapeutic opportunities. (2014) Cancer metastasis reviews 33(4):929-942.
19. Nicolsonm, G.L., Nawa, A., Toh, Y., et al. Tumor metastasis-associated human MTA1 gene and its MTA1 protein product: role in epithelial cancer cell invasion, proliferation and nuclear regulation. (2003) Clinical & experimental metastasis 20(1): 19-24.
20. Toh, Y., Nicolson, G.L. The role of the MTA family and their encoded proteins in human cancers: molecular functions and clinical implications. (2009) Clinical & experimental metastasis 26(3): 215-227.
21. Liu, J., Wang, H., Huang, C., et al. Subcellular localization of MTA proteins in normal and cancer cells. (2014) Cancer metastasis reviews 33(4): 843-856.
22. Levine, A.J. p53 the cellular gatekeeper for growth and division. (1997) Cell 88(3): 323-331.
23. Vousden, K.H., Prives, C. p53 and prognosis: new insights and further complexity. (2005) Cell 120(1): 7-10.
24. Prives, C. Signaling to p53: breaking the MDM2-p53 circuit. (1998) Cell 95(1): 5-8.
25. Katayama, H., Sasai, K., Kawai, H., et al. Phosphorylation by aurora kinase A induces Mdm2-mediated destabilization and inhibition of p53. (2004) Nat Genet 36(1): 55-62.
26. Wang, L., Wang, M., Wang, S., et al. Actin polymerization negatively regulates p53 function by impairing its nuclear import in response to DNA damage. (2013) PLoS ONE 8(4): e60179.

Ommega Online Publisher
Bioinformatics, Proteomics and Immaging Analysis
Short Title :Bioinfo Proteom Img Anal

ISSN: 2381-0793
E-mail : bioinfo@ommegaonline.org
website: www.ommegaonline.org