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

**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

**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-kB, 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[5-7]. 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 sequences in the NLSdb are used to identify the known non-nuclear proteins.

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

| Gene   | Gene Description                        | Phos. Site | Disease                      | Score | Assoc. Type         |
|--------|-----------------------------------------|------------|------------------------------|-------|---------------------|
| TP53   | tumor protein p53                        | S315; T312 | Adenocarcinoma               | 0.5   | Biomarker           |
| DKC1   | dyskeratosis congenita 1, dyskerin       | S451       | Hoyeraal-Hreidarsson syndrome | 0.60  | Genetic Variation   |
| XPC    | xeroderma pigmentosum, complementation C | S61        | Xeroderma Pigmentosum        | 0.70  | Genetic Variation   |
| VDR    | vitamin D(1,25-di-hydroxyvitaminD3)      | S51        | Vitamin D-Dependent          | 0.70  | Genetic Variation   |
| PTHLH  | parathyroid hormone-like hormone         | T108; T121 | Brachydactyly, Type E2       | 0.60  | Genetic Variation   |
| MYH3   | myosin, heavy chain 3, skeletal muscle,  | S949; T961 | Freeman-Sheldon syndrome     | 0.60  | Genetic Variation   |
| MYH6   | myosin, heavy chain 6, cardiac muscle,   | S950; T962 | Cardiomyopathy, Familial     | 0.70  | Genetic Variation   |
| MYH7   | myosin, heavy chain 7, cardiac muscle,   | S948; T960 | Myopathy, Myosin Storage     | 0.60  | Genetic Variation   |
|        | beta                                     |            |                              |       |                     |
| MYH8   | myosin, heavy chain 8, skeletal muscle,  | S951; T963 | Hecht syndrome               | 0.60  | Genetic Variation   |
Note: 1DisGeNET score was developed and calculated based on the number of sources that report the association, the type of curation of each of these sources, the animal models whose localization has been studied, and the number of supporting publications from text-mining based sources as described in the literature \[9-11\].

Results and Discussion

A 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 determination of genomic and proteomic approaches, gene expression at both mRNA and protein levels can be quantitatively analyzed in a global manner.

Changes in localization are often triggered post-transcriptional modifications such as phosphorylation. Using the Python language, we developed a bioinformatics module to survey phosphorylation sites and disease associations. Using the Uniprot database, we found phosphorylation sites near the nuclear localization signals of the 138 disease associated gene hits as defined in text. We then calculated the DisGeNET score for each association, ranging from 0 to 1 that indicated the strength of the association.

Figure 1: Flowchart of building the bioinformatics tool. A) Dataset preparation. The program starts with a list of NLS-containing proteins in NLSdb, finds 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-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.

| Gene  | Description                           | DisGeNET Score |
|-------|---------------------------------------|----------------|
| POR   | P450 (cytochrome) oxidoreductase       | 0.60           |
| TJP2  | tight junction protein 2              | 0.60           |
| KCNJ1 | potassium inwardly-rectifying channel, subfamily J, member 1 | 0.70           |
| type 2 | 0.70 | Genetic Variation |
| DSP   | desmoplakin                           | 0.60           |
| ADD1  | adducin 1 (alpha)                     | 0.52           |
| VCP   | valosin containing protein            | 0.71           |
| ATRX  | alpha thalassemia/mental retardation syndrome X-linked | 0.71           |
| TCOF1 | Treacher Collins-Franceschetti syndrome 1 | 0.72           |
| TNNT2 | tropomin T type 2 (cardiac)           | 0.70           |
| PLEC  | pectin                                | 0.70           |

Jianjun Hu., et al.  
J Bioinfo Proteomics Rev | Volume 3: Issue 1
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).

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 with 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 for ubiquitin-proteasome-mediated degradation.

Some associations identified in our analysis appear consistent with existing research, like that between p53 phosphorylation and localization. TP53 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 TP53 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[22,24]. 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 BIOD-BI-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., Queralta-Rosinach, N., Bravo, A., et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. (2015) Database (Oxford).
Computational identification of Genes associated with Human Diseases

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.